GENERATE

Breakthroughs Information Technology

Vol 01 (1) 2025 p. 31-47

**Corresponding author:**
Wang Ying
Email : 2023103058@nwnu.edu.cn

*Received 28 June 2025;*
*Accepted 15 July 2025;*
*Published 21 July 2025.*

**Conflict of interest statement:**
Author(s) reported no conflict of interest

DOI: http://doi.org/10.70764/gdpu-bit.2025.1(1)-03

# A SYSTEMATIC LITERATURE REVIEW ON THE ROLE OF EXPLAINABLE AI IN ENHANCING ALGORITHMIC FAIRNESS ACROSS APPLICATION DOMAINS

**Wang Ying**[1]

[1] *Northwest Normal University, China*

## ABSTRACT

**Objective**: Explainable Artificial Intelligence (XAI) is a field dedicated to improving the understanding and clarification of Machine Learning (ML) algorithms and their results. This research aims to determine commonly used XAI techniques, examine their classification, and assess their impact on the decision-making process.

**Research Design & Methods**: This research utilizes the PRISMA framework to conduct a systematic literature review of 310 Scopus-indexed articles on Explainable Artificial Intelligence (XAI) from 2018 to 2024, using targeted keyword searches to ensure rigorous selection and transparency in the research process.

**Findings:** The findings suggest that SHAP and LIME are the most frequently used explainable artificial intelligence (XAI) methods in the financial sector, due to their adaptability, clarity, and compatibility with various predictive models. However, there is still no standardized taxonomy, and only a few studies have focused on fairness or algorithmic fairness as a primary goal.

**Implications & Recommendations:** This analysis highlights the need for a more comprehensive framework that combines explanation with fairness metrics. Future research should investigate the integration of Explainable Artificial Intelligence (XAI) within regulatory structures, such as the General Data Protection Regulation (GDPR), to ensure that it meets the needs of technical and ethical assessment.

**Contribution & Value Added:** This research formulates a concise and applicable classification of XAI methods in the financial sector and highlights its relevance to equity issues to encourage the development of transparent, ethical, and auditable AI systems.

**Keywords:** Explainable AI, Fairness, Review, Algorithmic.

JEL codes: C88, O33
**Article type:** research paper

## INTRODUCTION

Explainable Artificial Intelligence (XAI) is a field dedicated to improving the understanding and clarification of Machine Learning (ML) algorithms and their results. With the increasing adoption of complex models such as deep neural networks and ensemble methods in various sectors, especially the financial sector, there is a pressing need to explain how and why algorithmic decisions are made (Attaran and Deb, 2018). This is particularly important as the decisions generated by these algorithms can significantly impact individuals, particularly in credit scoring systems, fraud detection, and loan processing. The most complex models, which learn from data through supervised learning methods such as neural networks or randomization, often lack transparency or provide limited insights. These models, while providing high performance, can be difficult to understand. Therefore, it is imperative to explain their outputs, so that users can understand the underlying rationale and thus facilitate informed decision-making (Martins et al., 2023).

While there is no single, universally accepted definition of explainability, the various studies focusing on XAI serve to outline the main goals in this field, each with different objectives and levels of detail. XAI techniques aim to explain the behaviour of ML models and their outputs. According to M. Turek, XAI aims to help users understand, trust, and effectively manage the new generation of Artificial Intelligence (AI) systems (Martins et al., 2023). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) characterizes intelligibility as the quality of an AI system that a non-expert can understand. An AI system is considered understandable if its functions and operations can be conveyed in non-technical terms to individuals without specialized knowledge. In addition to these definitions, there is significant interest among experts in understanding the internal workings of a model, referred to as ML model interpretability. Christoph Molnar states that interpretable ML includes methods and models that make the behavior and predictions of machine learning systems understandable to humans (Molnar, 2021).

One critical issue that arises along with the use of ML-based systems is fairness in algorithmic decision-making (Atmaja, 2025). High-accuracy models are not necessarily fair, as hidden biases in the training data or algorithm structure may lead to discrimination against certain groups (Qureshi et al., 2024). Therefore, the need for an XAI approach that not only explains model predictions but can also uncover potential biases and injustices is very important (Jain, 2024). Previous studies show that although various XAI techniques have been developed, only a few explicitly consider fairness aspects, especially in finance, which relies heavily on algorithmic decisions (Agu et al., 2024). This gap reflects the need to integrate the XAI taxonomy with the principle of fairness, as the financial sector is one of the areas with high risk if algorithm-based decisions are not accompanied by transparency and accountability (Oguntibeju, 2024). In this context, XAI has an important role as an internal audit tool that helps stakeholders understand the decision-making structure while assessing and mitigating potential bias (Deokar et al., 2024).

Fairness in algorithms is not only an ethical issue, but also closely related to regulation and social interests, especially to fulfill the principles of fairness and transparency stipulated in the General Data Protection Regulation (GDPR) (Official Journal of the European Union, 2016). The regulation emphasizes that personal data must be processed lawfully, fairly, and transparently (Article 5). It gives individuals the right to obtain information about automated decision-making (Article 14 paragraph 2.g), the right to be forgotten (Article 17), and the right to object to data processing (Article 21). These provisions make XAI not just an optional feature but a normative necessity to ensure data protection and fairness in AI-based systems. Implementing XAI in the financial sector is not only technically important but also urgent in fulfilling legal obligations and building public trust in an increasingly complex digital system. The purpose of this paper is twofold: first, to consolidate existing knowledge on XAI techniques and methods, particularly concerning tabular data; second, to present specific XAI methods and techniques used in the financial sector, informed by the findings of a systematic literature review.

The structure of this paper is as follows: Section 2 provides a brief overview of XAI fundamentals, Section 3 outlines the methodology used for the systematic search of articles; Section 4 presents a quantitative analysis of the search results; Section 5 conducts a qualitative analysis of the reviewed literature and identifies the XAI methods currently used in the financial industry; and finally, Section 6 offers conclusions along with a critical evaluation of the contributions of this work and the limitations encountered in this research.

## LITERATURE REVIEW

Explainable Artificial Intelligence (XAI) is a crucial approach in developing modern artificial intelligence systems, which seeks to increase transparency and clarity in the algorithmic decision-making process. By enabling users to understand the reasoning and mechanisms behind model predictions, XAI not only strengthens trust in the system but also plays a strategic role in detecting and mitigating potential hidden biases in algorithms (Oyeniran et al., 2022). This bias can have serious repercussions in many cases, especially when modelled decisions unfairly target social groups based on race, gender, or economic background (Aninze, 2024). Therefore, XAI opens up opportunities for ethical and technical audits, allowing user verification, such as doctors, judges, or financial analysts, to ensure that model decisions are relevant and non-discriminatory (Shuford, 2024). XAI is not only a diagnostic tool but also a literature tool for model refinement; explanatory results that are not in line with equity principles can trigger adjustments to the data, model structure, or objective function. This is particularly important in health, finance, education, and law sectors, where fairness and accountability are key principles.

Meanwhile, XAI also plays an important role in supporting algorithmic fairness. As a technical basis, Zafar et al. (2017) pioneered the development of algorithms that explicitly integrate fairness through concepts such as equal opportunity, demographic parity, and disparate mistreatment. This approach balances model accuracy and predictive fairness, marking a step forward in formal efforts to reduce algorithmic discrimination. XAI in this context is not just an additional feature, but an ethical foundation for inclusive and responsible AI design. The widespread implementation of XAI not only minimizes the risk of systemic bias but also paves the way for the development of AI systems that are transparent, auditable, and aligned with social and regulatory values as mandated in global regulations.

As the application of XAI grows in various domains, it is important to understand the taxonomic framework underlying XAI methods to assess how explanations are generated by machine learning systems (Saarela and Podgorelec, 2024). While many XAI studies focus on its applications in finance, healthcare, or law, understanding the underlying mechanisms of XAI methods is crucial for the findings to be interpreted broadly by various parties. For a more in-depth exploration, recent systematic literature and comprehensive books provide a thorough analysis of the advantages and disadvantages of each approach (Saarela and Podgorelec, 2024). Generally, XAI methods are classified based on the explanation mechanism used, including example-based approaches, counterfactuals, hidden semantics, rules, and features or saliency (Saarela and Podgorelec, 2024). The feature importance approach is the most dominant in the context of classification models due to its ability to score and rank features, thus explaining the contribution of each feature to the model's predictions (Saarela and Jauhiainen, 2021; Wojtas and Chen, 2020). In image-based models, features are represented as super pixels, so visual methods such as saliency maps and pixel attribution become important tools to account for local influences on predictive output (Arrieta et al., 2020; Burkart and Huber, 2021).

## METHODS

The PRISMA framework, which stands for Preferred Reporting Items for Systematic Reviews and Meta-Analyses, is essential in preparing review papers. PRISMA enhances systematic reviews by promoting transparency in meta-analysis, ensuring proper reporting of objectives, and facilitating drawing reliable and relevant conclusions from research results. Therefore, this review utilized the PRISMA methodology to conduct an extensive literature review on XAI, as illustrated in Figure 1.

The research materials used in this study were drawn from leading scientific journals covering January 2018 to December 2024, with searches conducted through the Scopus.com platform. The SCOPUS citation database was chosen for the survey search and literature review because it has stricter standards than Google Scholar or other search engines that do not provide a validation component. The keywords used are divided into two search elements: first, definitions of the field of study; second, filters for surveys or literature reviews. TITLE (〝Explainable Artificial Intelligence〞 OR 〝Explainable AI〞 OR 〝XAI〞 AND TITLE (〝Systematic Review〞 OR 〝Review〞 OR 〝Survey〞). The search was conducted without specifying the application domain, which was considered irrelevant as the aim was to obtain an overview of the definition of XAI and a clear specification of the method categories. An initial 310 papers were selected based on their relevance to the specified topics and keywords.

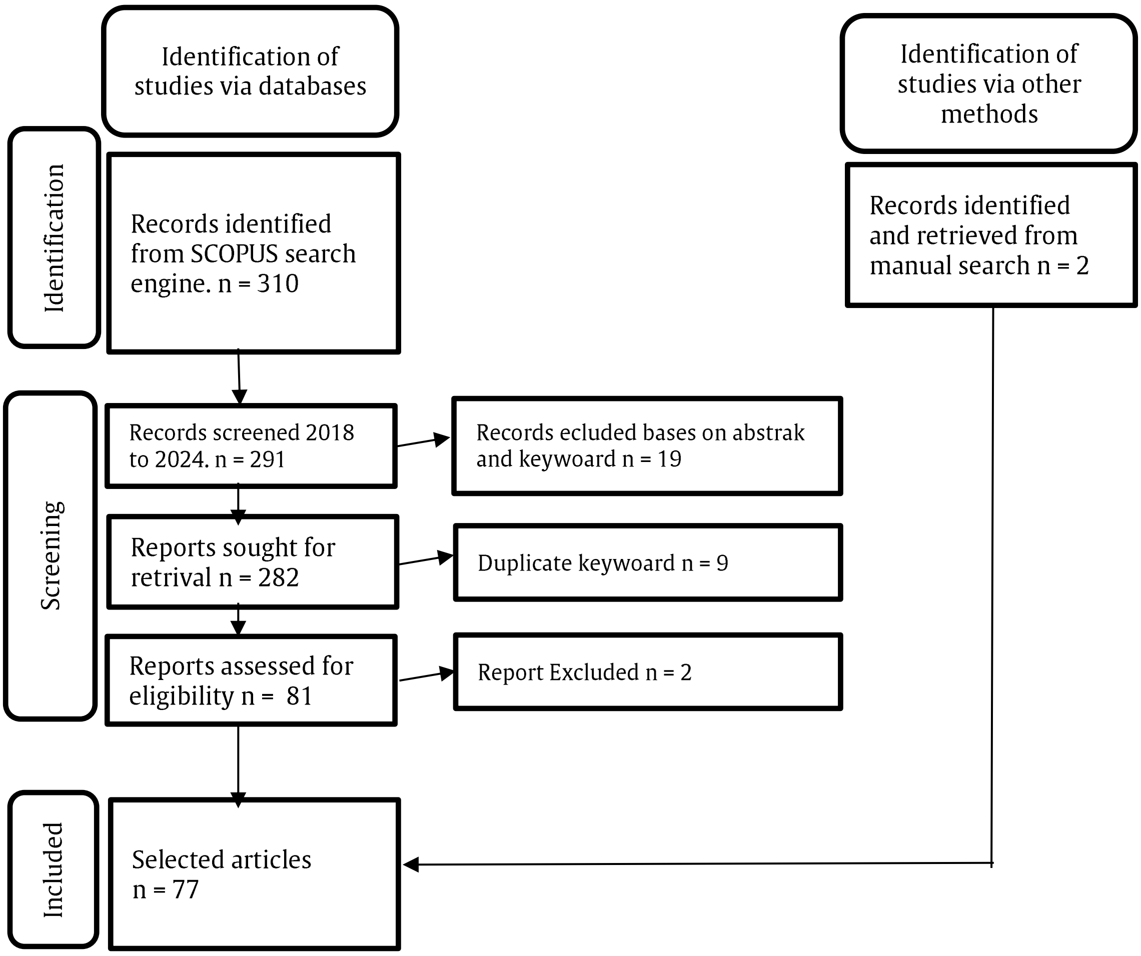
Identification of studies via databases
Identification of studies via other methods
Identification
Records identified from SCOPUS search engine. n = 310
Records identified and retrieved from manual search n = 2
Screening
Records screened 2018 to 2024. n = 291
Records ecluded bases on abstrak and keyword n = 19
Reports sought for retrival n = 282
Duplicate keyword n = 9
Reports assessed for eligibility n = 81
Report Excluded n = 2
Included
Selected articles n = 77

Figure 1. PRISMA framework for survey research

## RESULT

The results of a bibliometric analysis using VOSviewer of recent scientific publications show that the topics of Explainable AI (XAI), interpretability, and trust in artificial intelligence systems are an important focus in academic studies. In recent years, there has been a significant increase in interest in transparency and fairness in algorithmic systems. This is reflected in the dominance of keywords such as "explainable artificial intelligence," which appears 112 times with an average publication year of 2023 and a high citation rate. Similarly, the terms "XAI" and "interpretability" appeared 66 and 74 times, respectively, with relevance and citation rates showing great attention to the ability of AI systems to explain the reasoning behind their decisions.

The thematic mapping performed by VOSviewer grouped these keywords into several main clusters. The first cluster focuses on the interpretability aspect centered on the system's and the user's relationship. Keywords such as "transparency", "interpretable AI", and "trust" appear frequently in this cluster and illustrate the importance of AI systems to be explainable in the context of human understanding. This led to creating a co-occurrence network of keywords illustrated in Figure 2. User trust in the system is strongly influenced by the extent to which the system can transparently explain the logic or process used in making decisions.

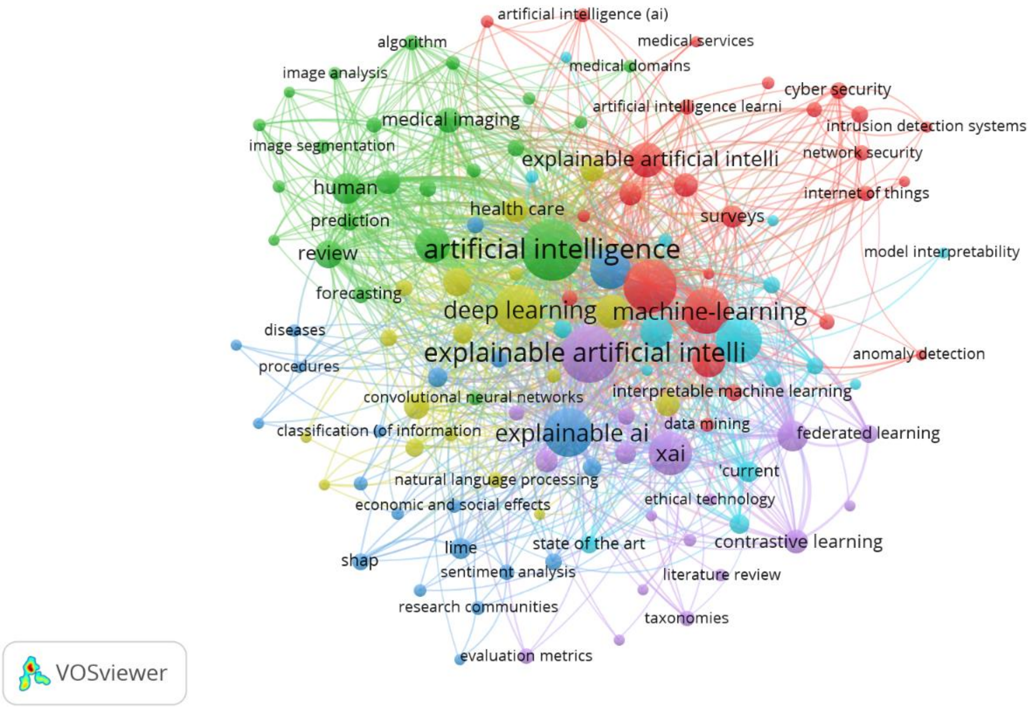

Figure 2. Keyword co-occurrence map in explainable artificial intelligence (XAI) research

The fifth cluster focuses on evaluating XAI systems and frameworks. It includes keywords such as "systematic literature review", "responsible AI", and "taxonomies". This cluster has important research value as it leads to the development of evaluative and methodological approaches to ensure that AI systems developed are not only accurate but also ethical and accountable. The keyword "responsible AI" indicates increased attention to fairness, reliability, and ethical aspects in the design and application of AI technologies. While the term "fairness" does not explicitly dominate the keyword map, the concept is implicitly reflected through terms such as "trustworthiness", 'transparency', and "responsible AI". This indicates that the issue of fairness has become an important concern in discussions around XAI, especially when AI is applied in the context of decision-making that impacts certain individuals or groups of people. For example, in a financial context, AI is used to assess credit risk, approve loans, or detect fraud- all require assurance that the decisions taken are unbiased and can be explained rationally.

Interestingly, research trends in the past five years also show an increase in systematic review-based approaches such as systematic literature review (SLR), which underscores researchers' efforts to establish a strong theoretical and practical foundation for XAI development. In addition, the emergence of the keyword "finance" (with an average publication year of 2024) reinforces the assumption that the application of XAI in the financial sector is starting to receive greater attention. This shows great potential for further research combining XAI with fairness principles in the finance domain, especially to ensure that automated decision-making systems can be audited and held accountable.

## Explainable AI Publication Trends

Graph 1 publication trends show a significant increase in related studies, starting with only two publications in 2018 and ending with 121 publications in 2024. After slow growth from 2018 to 2020, a spike is seen in 2021 with 26 publications, stabilizing at 39 through 2022 and 2025. However, a drastic spike occurred in 2023 with 74 publications and peaks in 2024, signalling increased academic and practical attention to Explainable AI (XAI) and fairness in algorithmic systems. This increase reflects the global urgency to deliver AI systems that are not only technically advanced but also transparent, explainable, and fair, especially in the context of applications in critical areas such as finance and personal data regulation.

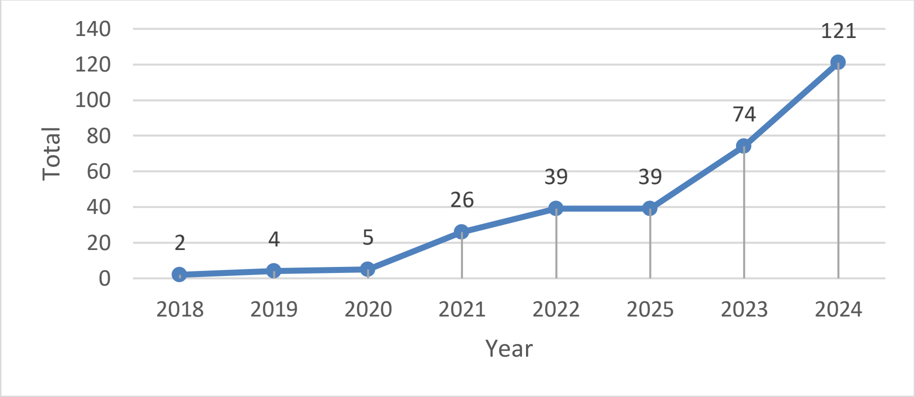

Figure 3. XAI Publication Trends year over year

Based on an analysis of the distribution of SCOPUS documents by discipline, most publications on Explainable AI (XAI) come from Computer Science, reflecting XAI's primary focus on algorithm development and intelligent systems. Significant contributions are also seen from engineering, mathematics, decision sciences, and social sciences, demonstrating the relevance of XAI in AI-based decision-making and behavioural studies. The field of Medicine occupies an important position due to the need for interpretability in automated diagnosis. Meanwhile, contributions from Economics, Finance, and Management are still limited, indicating that the utilization of XAI in these domains is still in its infancy or relies on AI methods that are already interpretative. Although technical domains still dominate XAI, the opportunities for its application in social, medical, and financial fields are immense, especially as the need for transparent, ethical, and accountable systems increases.

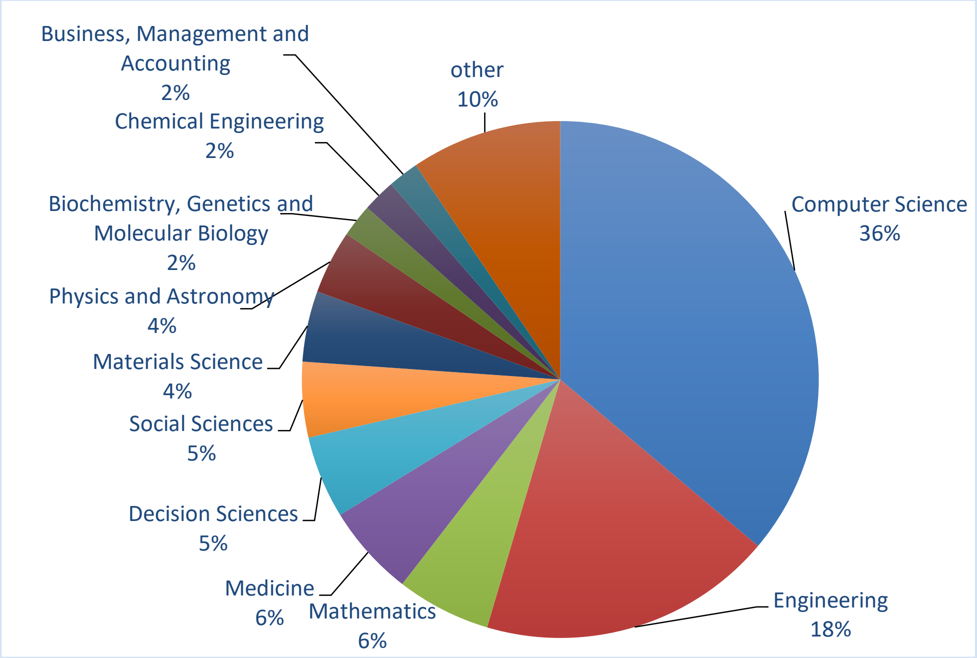

Figure 4. The primary focus area of the journal is the article

## Review of Literature and Examination of Current Surveys

A systematic review of several key relevant publications was conducted to gain a more comprehensive understanding of the development and focus of Explainable Artificial Intelligence (XAI) research, particularly in the context of fairness and its applications in various domains. Table 1 summarizes the literature review results by presenting the main references, the focus of each study, the main contributions made, and the important insights that can be drawn from each study. This summary not only describes the theoretical and methodological landscape of XAI but also identifies research gaps that can be used as a basis for developing further studies.

Table 1. Explainable Artificial Intelligence Literature Review

| No. | Author/ Reference | Study Focus | Contributions | Insights |
|---|---|---|---|---|
| 1 | (Adadi and Berrada, 2018) | Reason for need XAI | Emphasizes the importance of XAI for legal compliance, ML system understanding, and model improvement. Provides a taxonomy for the characterization of XAI methods. | XAI helps with identification of system weaknesses and knowledge extraction. Applications include finance, demonstrating the relevance of fairness in this domain. |
| 2 | (Angelov et al., 2021) | XAI history and taxonomy | Offers a historical perspective and classification of XAI methods. Describes various methods with a new taxonomy. | XAI is discussed in the context of criminal justice and fraud detection, demonstrating the importance of fairness in social and financial systems. |
| 3 | (Islam et al., 2022) | SLR XAI (137 papers) | Systematic review of 137 publications, identification of application domains, and new taxonomies. | Only 3 out of 137 studies touched on the finance domain, suggesting that XAI in finance is under-researched and has great potential. |
| 4 | (Stepin et al., 2021) | Contrastive & counterfactual explanations | Taxonomy for two explanatory approaches: contrastive and counterfactual. | This explanation is important for fairness because it gives users specific reasons for the AI system's decisions. |
| 5 | (Linardatos et al., 2021) | XAI to AI & DL | Reviewed XAI techniques and developed a taxonomy based on white-box and black-box approaches. | Provides a comprehensive overview of relevant XAI methods for fairness, especially in hard-to-describe models such as Deep Learning. |
| 6 | (Minh et al., 2022) | XAI theory review | Classification papers by type, explanation, and discussion of the advantages and disadvantages of methods. | Important for fairness as it helps to choose the right XAI method based on the context of use. |
| 7 | (Lin et al., 2021) | XAI to Deep Learning | A hierarchical taxonomy that emphasizes the DL interpretation method. | Discussing the trade-off between interpretability and performance is a key issue in building fair AI systems. |

| No. | Author/ Reference | Study Focus | Contributions | Insights |
|---|---|---|---|---|
| 8 | (Darias et al., 2021) | XAI method library | Analyze XAI's method library and how each generates explanations. | Even without an explicit taxonomy, focusing on explanatory mechanisms is useful for fairness practices. |
| 9 | (Bertrand et al., 2022) | Cognitive bias in XAI | SLR on cognitive bias in XAI-based decision-making systems. | Fairness is threatened by human biases absorbed into the XAI system - it is important to mitigate these biases in the system design. |
| 10 | (Hanif et al., 2021) | Ethics in XAI | Review of XAI method in ethics and taxonomy based on ethical values. | Evaluating XAI methods based on ethical principles is very important in the context of AI fairness. |
| 11 | (Lopes et al., 2022) | XAI method evaluation | Taxonomy for evaluating XAI methods rather than the methods themselves. | A comprehensive evaluation is essential to measure whether the XAI method promotes fairness. |
| 12 | (Vilone and Longo, 2020) | Definition, explanation, and formal challenge | Emphasized the lack of a standard definition for 'explanation' in ML and the formal challenges in XAI. | Demonstrates the importance of a consistent approach to the definition of fairness and interpretability in XAI. |
| 13 | (Molnar, 2021) | XAI method category | Three main categories of XAI methods. | The different approaches point to the need for uniform classification for fairness and accountability. |

## DISCUSSION

### Integration of Fairness in Explainable Artificial Intelligence (XAI) Taxonomy and Applications

Based on literature analysis, this research indicates that the field of Explainable Artificial Intelligence (XAI) is experiencing rapid development, but still faces fundamental challenges related to taxonomy consistency and integration of justice values. Although various classification approaches of XAI methods have been proposed by researchers, such as Adadi and Berrada, 2018; Angelov et al., 2021; Islam et al., 2022; Molnar, 2021, there is no widely accepted taxonomic standard yet. Each study offers a varied classification structure based on the stage of application (a posteriori vs. a priori), the scope of the model (model-specific vs. model-agnostic), and the type of explanation (local vs. global), reflecting the complexity and diversity of XAI applications in various domains. The importance of XAI in meeting regulatory demands, such as the General Data Protection Regulation (GDPR), as well as fulfilling ethical and practical needs for explaining the results of automated decisions, is discussed. However, research by Islam et al., (2022) shows that only a small proportion of XAI research explores its application in finance. This sector relies heavily on algorithmic decisions and is at high risk if not accompanied by transparency and accountability. This suggests a significant gap between the development of methods and their practical implementation in critical sectors.

In addition to the challenges in taxonomy structure, the issue of fairness is also a dimension that has not been fully accommodated in existing XAI frameworks. Some research has implicitly touched on fairness through terms such as "responsible AI", 'trust', and "transparency". However, very few have methodologically integrated fairness metrics in XAI evaluation models. Fairness is one of the key elements to ensure that AI systems are not only explainable but also non-discriminatory and uphold the principles of social justice. Important contributions also come from studies that promote counterfactual explanation methods and user-based evaluation approaches (Lopes et al., 2022). This method provides a new direction in developing explanations that are more intuitive and acceptable to various stakeholders, from system developers to end users. These findings further reinforce the urgency of developing a unified framework that not only classifies XAI methods technically but also integrates ethical, social, and regulatory dimensions. Therefore, future research directions should focus on: (1) developing an inclusive and cross-disciplinary taxonomy of XAI, (2) developing fairness metrics that can be applied in evaluating model explanations, and (3) exploring broader applications of XAI in the financial sector and other sensitive areas. Thus, XAI is not only a technical tool for model interpretation but also a normative framework for building a fair, transparent, and accountable artificial intelligence system.

### A comprehensive classification of XAI techniques

As the literature in the field of XAI advances, the importance of systematic efforts to classify XAI methods based on certain characteristics is increasing. This helps not only in understanding the various approaches that have been developed but also in assessing the extent to which such methods can contribute towards achieving transparency and fairness in machine learning (ML)-based systems. One taxonomy approach proposed in the literature categorizes XAI methods based on three main dimensions: Stage, Model, and Scope. The first category, Stage, refers to the time or phase in which XAI methods are applied to ML models. In this context, XAI methods are divided into two categories: ante-hoc and post-hoc. Antecedent methods are interpretation techniques intrinsically integrated into the model from the start, such as in decision trees or linear regression. On the other hand, post-hoc methods are applied after the model has generated predictions, without knowing the internal mechanism of the model (Adadi and Berrada, 2018; Islam et al., 2022; Speith, 2022). Examples include LIME, a surrogate model that attempts to mimic the functionality of a black-box model by performing local sampling and analysis of the data (Ribeiro et al., 2016). This category is very important as it is directly related to how transparent the model is from its initial design, which can affect users' perception of fairness.

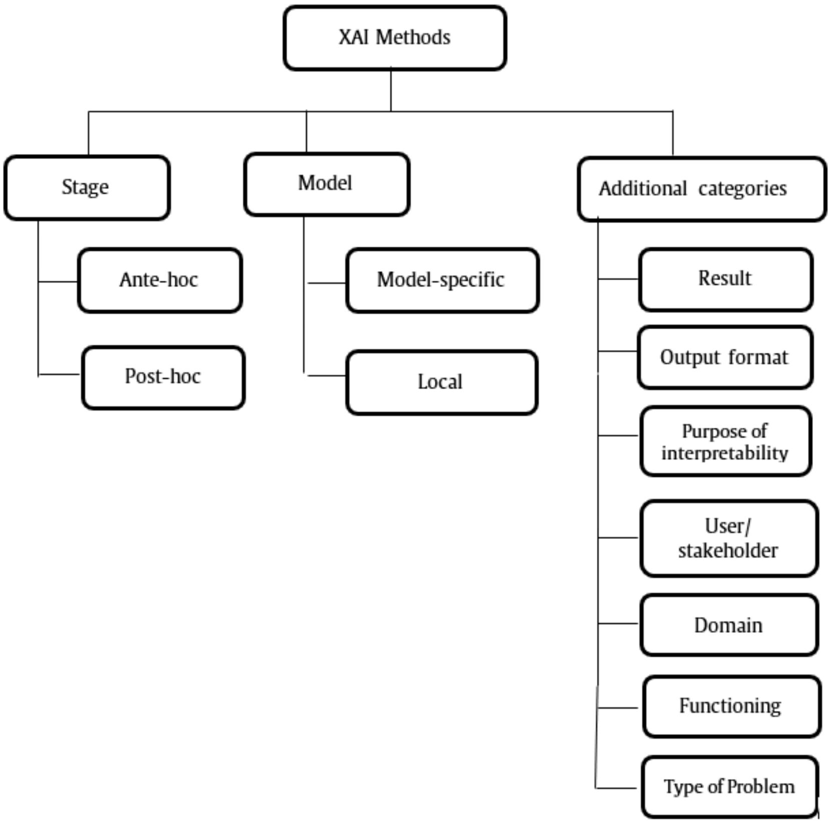

Figure 5. XAI methods specific

The second category, Model, describes whether an XAI method is model-specific or model-agnostic. Model-specific methods can only be used on certain models because they depend on the internal architecture, such as decision trees or neural networks in the form of layer visualizations (Arrieta et al., 2020; Lin et al., 2021; Sahakyan et al., 2021). Conversely, model-agnostic methods such as LIME or SHAP can be widely applied to different models as they do not require access to the internal structure of the model. In the context of fairness, the generalizability of agnostic methods is often an advantage as they can be applied across domains and models to compare fairer outcomes between different systems (Jiang et al., 2024).

The third category, Scope, refers to the scope of explanation provided by XAI methods, i.e., global versus local interpretability. Global methods explain the behavior of the entire model, providing an overview of how features are used to make predictions in aggregate. In contrast, local methods explain one specific prediction or a specific subset of data. An example of a local method is LIME, which shows the importance of a feature for a single instance of data, whereas a global method such as SHAP can provide a generalized contribution of features across predictions (Lundberg and Lee, 2017). In strengthening fairness, a localized approach is particularly useful for conducting individualized audits of algorithm decisions that may have a discriminatory impact on certain users (Bhaumik et al., 2022).

Besides three main categories, some literature also proposes additional categories such as Outcome (the way the results are presented, whether in numbers, text, or visuals), Output Format, Interpretability Objectives, Users/Stakeholders, and Domain (Arrieta et al., 2020; Hu et al., 2021; Molnar, 2021; Speith, 2022). These categories reflect the complexity in the development of XAI systems and show that interpretability cannot be separated from the user context, intended use, and customized form of explanation presentation. For example, in a financial context, stakeholders such as regulators may be more interested in global aggregate visualizations, while end-users (customers) require simple and intuitive explanations of the reasons for rejecting their loan applications. Some literature also emphasizes the importance of categories such as Function, which defines the way the XAI method extracts information from the model (e.g., through perturbation or structure visualization), as well as Problem Type, which distinguishes whether the XAI method is used for classification or regression. This distinction is important to ensure the suitability of the method to the type of data and analytic objectives at hand. Overall, this taxonomy framework contributes to a more structured literature map and can serve as a basis for developing AI systems that can not only be explained but also fairly and responsibly audited (Speith, 2022; Vilone and Longo, 2020). This aligns with the main objectives of this research: identifying XAI methods that support the achievement of algorithmic justice, particularly in finance, and suggesting a more integrative classification framework as a basis for developing a more transparent and ethical AI-based financial system.

### Review of Literature and Examination of Practical Applications

Explainable Artificial Intelligence (XAI) methods have been widely applied in the financial sector, especially in credit scoring and fraud detection issues. However, the application of XAI for fraud detection is still relatively low compared to the emphasis on credit risk prediction. Generally, SHAP (SHapley Additive exPlanations) is the most widely used XAI method in this field. SHAP is model-independent and can provide both local and global explanations, with many studies utilizing it to assess the contribution of features to prediction, either directly or in combination with techniques such as clustering and decision trees (Maree et al., 2020). Interestingly, some studies integrate SHAP with counterfactual methods to generate more contextualized and personalized local explanations (Chaquet-Ulldemolins et al., 2022; Watson, 2022).

Besides SHAP, LIME (Local Interpretable Model-agnostic Explanations) is often used in research. LIME is usually applied at a local scale to explain a single instance of prediction, whereas SHAP is more flexible and can be used at both scales. Both utilize the concept of feature importance to explain variable contributions to the prediction outcome (Tyagi, 2022; Ullah et al., 2021). Counterfactual explanations add a dimension of interpretation through 'what-if' examples that show feature changes that can affect the prediction results. One prominent method is

PermuteAttack, a genetic algorithm-based approach that optimizes the minimum possible feature changes to generate counterfactual examples (Hashemi and Fathi, 2020). Other methods like PDP (Partial Dependence Plots) and DALE (Area of Local Effects) are applied for global interpretability through visualizing the relationship between features and target outputs. Additionally, methods such as PASTLE and CASTLE simplify the data space by identifying pivot points or behavioral clusters and extracting rule-based rules. Other alternative approaches that support local explanation include Anchors, MANE, and joint models such as TREPAN, which are used in various model scenarios.

Regarding model-specific methods, LTreeX and inTrees are two approaches that utilize decision trees to provide explanations. LTreeX operates locally by generating a surrogate of Random Forest (Dedja et al., 2022). In contrast, inTrees offers a global explanation by extracting rules from a collection of decision trees (Deng, 2019). The combination of hybrid methods, such as SHAP plus counterfactuals, and new approaches, such as Rational Shapley Values, shows that innovations continue to be made to improve the understanding of models in a financial context (Redelmeier et al., 2020). Research conducted by Hastie et al. (2009) and Zhang et al. (2022) emphasizes the importance of XAI in explaining financial risk prediction, especially through applying SHAP and counterfactual methods. Another practical implementation is integrating LIME and SHAP to explain stock prediction models and company data (Mandeep et al., 2022; Park et al., 2021). Additionally, the use of XAI is driven by regulatory factors, which seek to meet GDPR legal demands regarding transparency in automated decisions (Huynh et al., 2021).

Table 2. Classification of XAI Techniques

| XAI Method | Stage | Model | Scope | Description |
|---|---|---|---|---|
| SHAP (Kim and Woo, 2021) | Post-hoc | Model-agnostic | Global & Local | Using Shapley scores to explain individual and aggregate feature contributions. |
| LIME (Hadash et al., 2022; Wu and Wang, 2021). | Post-hoc | Model-agnostic | Local | Local model that mimics the main model to explain instance-specific predictions. Used for instance-level explanation, often paired with SHAP. |
| Counterfactual (Dastile et al., 2022; Guidotti et al., 2019; Hashemi and Fathi, 2020) | Post-hoc | Model-agnostic | Local | Provide feature-based explanations of changes that affect outcomes (what-if scenarios). |
| PDP (Friedman, 2001; Gkolemis et al., 2022; Zou et al., 2022) | Post-hoc | Model-agnostic | Global | Visualizing the effects of a single feature on model results is generally used to interpret key features. |
| DALE (Gkolemis et al., 2022) | Post-hoc | Model-agnostic | Global | Variation of PDP with a localized area-of-effect approach that focuses more on the contribution of features to the target. |
| Anchors (Ribeiro et al., 2018) | Post-hoc | Model-agnostic | Local | Explanations are based on if-then rules that are iteratively selected from the most relevant features. |
| PASTLE (La Gatta et al., 2021b) | Post-hoc | Model-agnostic | Local | Data space reduction to pivot points is then used for rule extraction. |

| XAI Method | Stage | Model | Scope | Description |
|---|---|---|---|---|
| CASTLE (La Gatta et al., 2021a) | Post-hoc | Model-agnostic | Local | Classify data into behavioural clusters before creating rule-based explanations. |
| MANE (Tian and Liu, 2020) | Post-hoc | Model-agnostic | Local | Designed for deep learning, it uses cross-feature extraction and linear regression for explanation. |
| LTreeX (Dedja et al., 2022) | Post-hoc | Model-specific | Local | Create a surrogate of Random Forest, extracting local rules for each instance. |
| inTrees (Deng, 2019) | Post-hoc | Model-specific | Global | Rule extraction from ensemble decision trees such as Random Forest and Boosted Trees. |
| TREPAN + NN (Craven and Shavlik, 1995; De et al., 2020) | Post-hoc | Hybrid | Local | Combining clustering from Neural Networks with decision trees to generate explanations. |
| Rational SHAP (Hadash et al., 2022; Watson, 2022) | Post-hoc | Model-agnostic | Local | Combination of SHAP and counterfactual for more robust and intuitive explanatory results. |

While various XAI approaches have been applied, most of the methods used in the financial sector are post-hoc methods, i.e., they are applied after the model provides predictions. This suggests that interpretability is still considered an additional process, rather than part of the initial design of the model. However, a posteriori method, such as decision trees, which are intrinsically explainable, have not been widely used in the reviewed research. This becomes an important opportunity for further exploration, especially for creating AI systems that are not only accurate but also explainable and auditable from the design stage. With the wide variety of methods and approaches to implementing XAI in the financial sector, it can be concluded that SHAP is the most widely used method, followed by LIME, counterfactual, PDP, and other new methods. A key challenge going forward is to ensure that the methods implemented not only provide technical transparency but are also able to meet the needs of users and stakeholders in a fair and responsible manner. Therefore, future development and evaluation of XAIs will need to pay more attention to the context of use, the type of users, and the ethical implications of the explanations provided.

## CONCLUSION

Based on the results of a literature review and systematic synthesis of the available literature, this study concludes that the use of Explainable Artificial Intelligence (XAI) in the finance domain has experienced rapid development, but is still faced with taxonomic challenges and limited standards. This study was initially conducted through a two-stage literature search: the first focused on reviews of XAI methods in general, and the second targeted practical applications of XAI in the financial sector. The analysis found that there is no consensus in the classification of XAI methods, with many authors proposing different taxonomies. This prompted this research to propose a unified taxonomy framework that is simple yet covers the main characteristics of existing XAI methods.

In practical applications, SHAP and LIME have emerged as the two most dominant explanation methods used in financial studies, especially for problems such as credit scoring and fraud detection. The main advantage of these two methods lies in their model-agnostic nature. It can be applied post-hoc, thus allowing integration with different types of predictive models and simultaneous application. Besides SHAP and LIME, new methods further expand the scope of XAI approaches, such as counterfactual explanations, partial dependence plots, and hybrid methods originally used for image classification but have been adapted for financial tabular data. This diversity of approaches shows that XAI in the financial context offers not only transparency but also great flexibility in responding to the various needs of users and regulators.

The main implication of this research is the importance of developing an XAI framework that is not only oriented towards technical explanations but also considers the principles of fairness and accountability, especially in sensitive sectors such as finance. With the rise of regulations such as GDPR and the demands of public ethics, the need for humanly explainable AI systems is becoming more urgent. Therefore, these findings can serve as a starting point for developing more inclusive and socially responsible AI systems. Future research is recommended to further explore the interaction between XAI taxonomy and fairness metrics in real practice and develop evaluation tools to holistically measure the quality of explanations produced by various XAI methods. This way, the future of XAI research will be increasingly oriented towards social impact, technological sustainability, and the integration of ethics in artificial intelligence.

## REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access,* 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Agu, E. E., Abhulimen, A. O., Obiki-Osafiele, A. N., Osundare, O. S., Adeniran, I. A., & Pelumi, C. (2024). Discussing ethical considerations and solutions for ensuring fairness in AI-driven financial services. *International Journal of Frontline Research in Multidisciplinary Studies*, *3*(2), 001–009. https://doi.org/10.56355/ijfrms.2024.3.2.0024

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5). https://doi.org/10.1002/widm.1424

Aninze, A. (2024). Artificial Intelligence Life Cycle: The Detection and Mitigation of Bias. *International Conference on AI Research*, *4*(1), 40–49. https://doi.org/10.34190/icair.4.1.3131

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. *Information Fusion*, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Atmaja, S. A. (2025). Ethical Considerations in Algorithmic Decision-making: Towards Fair and Transparent AI Systems. *Riwayat: Educational Journal of History and Humanities*, *8*(1), 620–627. https://doi.org/10.24815/jr.v8i1.44112

Attaran, M., & Deb, P. (2018). Machine Learning: The New "Big Thing" for Competitive Advantage. *International Journal of Knowledge Engineering and Data Mining*, *5*(1), 1. https://doi.org/10.1504/IJKEDM.2018.10015621

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How Cognitive Biases Affect XAI-assisted Decision-making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 78–91. https://doi.org/10.1145/3514094.3534164

Bhaumik, D., Dey, D., & Kayal, S. (2022). A Framework for Auditing Multilevel Models using Explainability Methods. http://arxiv.org/abs/2207.01611

Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, 70, 245–317. https://doi.org/10.1613/jair.1.12228

Chaquet-Ulldemolins, J., Gimeno-Blanes, F.-J., Moral-Rubio, S., Muñoz-Romero, S., & Rojo-Álvarez, J.-L. (2022). On the Black-Box Challenge for Fraud Detection Using Machine Learning (II): Nonlinear Analysis through Interpretable Autoencoders. *Applied Sciences*, *12*(8), 3856.

https://doi.org/10.3390/app12083856

Craven, M. W., & Shavlik, J. W. (1995). Extracting Tree-Structured Representations of Trained Networks. *NIPS 1995: Proceedings of the 8th International Conference on Neural Information Processing Systems*, 24–30.

Darias, J. M., Iaz-Agudo, B. en D., & Recio-Garcia, J. A. (2021). A Systematic Review on Model-agnostic XAI. *Workshops for the 29th International Conference on Case-Based Reasoning, September*, 28–29. https://doi.org/10.5281/zenodo.5838263

Dastile, X., Celik, T., & Vandierendonck, H. (2022). Model-Agnostic Counterfactual Explanations in Credit Scoring. *IEEE Access*, 10, 69543–69554. https://doi.org/10.1109/ACCESS.2022.3177783

De, T., Giri, P., Mevawala, A., Nemani, R., & Deo, A. (2020). Explainable AI: A Hybrid Approach to Generate Human-Interpretable Explanation for Deep Learning Prediction. *Procedia Computer Science*, 168, 40–48. https://doi.org/10.1016/j.procs.2020.02.255

Dedja, K., Nakano, F. K., Pliakos, K., & Vens, C. (2022). BELLATREX: Building Explanations through a LocaLly AccuraTe Rule EXtractor. *ArXiv Preprint*. http://arxiv.org/abs/2203.15511

Deng, H. (2019). Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics*, *7*(4), 277–287. https://doi.org/10.1007/s41060-018-0144-8

Deokar, R., Nanjundan, P., & Mohanty, S. N. (2024). Transparency in Translation: A Deep Dive into Explainable AI Techniques for Bias Mitigation. *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, 1–6. https://doi.org/10.1109/APCIT62007.2024.10673712

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Gkolemis, V., Dalamagas, T., & Diou, C. (2022). DALE: Differential Accumulated Local Effects for efficient and accurate global explanations. *ArXiv Preprint, Oktober*. http://arxiv.org/abs/2210.04542

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems*, *34*(6), 14–23. https://doi.org/10.1109/MIS.2019.2957223

Hadash, S., Willemsen, M. C., Snijders, C., & IJsselsteijn, W. A. (2022). Improving understandability of feature contributions in model-agnostic explainable AI tools. *CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3491102.3517650

Hanif, A., Zhang, X., & Wood, S. (2021). A Survey on Explainable Artificial Intelligence Techniques and Challenges. *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*, 81–89. https://doi.org/10.1109/EDOCW52865.2021.00036

Hashemi, M., & Fathi, A. (2020). PermuteAttack: Counterfactual Explanation of Machine Learning Credit Scorecards. *ArXiv Preprint*. http://arxiv.org/abs/2008.10138

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. *Springer New York*. https://doi.org/10.1007/978-0-387-84858-7

Hu, Z. F., Kuflik, T., Mocanu, I. G., Najafian, S., & Shulner Tal, A. (2021). Recent Studies of XAI - Review. Adjunct Proceedings of the 29th ACM Conference on User Modeling, *Adaptation and Personalization*, 421–431. https://doi.org/10.1145/3450614.3463354

Huynh, T. D., Tsakalakis, N., Helal, A., Stalla-Bourdillon, S., & Moreau, L. (2021). Addressing Regulatory Requirements on Explanations for Automated Decisions with Provenance—A Case Study. *Digital Government: Research and Practice*, *2*(2), 1–14. https://doi.org/10.1145/3436897

Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences*, *12*(3), 1353. https://doi.org/10.3390/app12031353

Jain. (2024). Transparency in AI Decision Making: A Survey of Explainable AI Methods and Applications. *Advances in Robotic Technology*, *2*(1), 1–10. https://doi.org/10.23880/art-16000110

Jiang, K., Zhao, C., Wang, H., & Chen, F. (2024). FEED: Fairness-Enhanced Meta-Learning for Domain Generalization. *2024 IEEE International Conference on Big Data (BigData)*, 949–958. https://doi.org/10.1109/BigData62323.2024.10825892

Kim, S., & Woo, J. (2021). Explainable AI framework for the financial rating models. *2021 10th International Conference on Computing and Pattern Recognition*, 252–255. https://doi.org/10.1145/3497623.3497664

La Gatta, V., Moscato, V., Postiglione, M., & Sperlì, G. (2021a). CASTLE: Cluster-aided space transformation for local explanations. *Expert Systems with Applications*, 179, 115045. https://doi.org/10.1016/j.eswa.2021.115045

La Gatta, V., Moscato, V., Postiglione, M., & Sperlì, G. (2021b). PASTLE: Pivot-aided space transformation for local explanations. *Pattern Recognition Letters*, 149, 67–74. https://doi.org/10.1016/j.patrec.2021.05.018

Lin, K.-Y., Liu, Y., Li, L., & Dou, R. (2021). A Review of Explainable Artificial Intelligence (pp. 574–584). https://doi.org/10.1007/978-3-030-85910-7_61

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 1–45. https://doi.org/10.3390/e23010018

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. https://doi.org/10.3390/app12199423

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 2017-Decembe*r(Section 2), 4766–4775.

Mandeep, Agarwal, A., Bhatia, A., Malhi, A., Kaler, P., & Pannu, H. S. (2022). Machine Learning Based Explainable Financial Forecasting. *2022 4th International Conference on Computer Communication and the Internet (ICCCI)*, 34–38. https://doi.org/10.1109/ICCCI55554.2022.9850272

Maree, C., Modal, J. E., & Omlin, C. W. (2020). Towards Responsible AI for Financial Transactions. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 16–21. https://doi.org/10.1109/SSCI47803.2020.9308456

Martins, T., de Almeida, A. M., Cardoso, E., & Nunes, L. (2023). Explainable Artificial Intelligence (XAI): A Systematic Literature Review on Taxonomies and Applications in Finance. *IEEE Access*, *12*, 618–629. https://doi.org/10.1109/ACCESS.2023.3347028

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, *55*(5), 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

Molnar, C. (2021). Interpretable Machine Learning A Guide for Making Black Box Models Explainable. *In Hands-On Machine Learning with R* (Vol. 04, pp. 305–342). Chapman and Hall/CRC. https://doi.org/10.1201/9780367816377-16

Official Journal of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons about the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da. *European Union Law*. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Oguntibeju, O. O. (2024). Mitigating Artificial Intelligence Bias in Financial Systems: A Comparative

Analysis of Debiasing Techniques. *Asian Journal of Research in Computer Science, 17*(12), 165–178. https://doi.org/10.9734/ajrcos/2024/v17i12536

Oyeniran, O. C., Adewusi, A. O., Adeleke, A. G., Akwawa, L. A., & Azubuko, C. F. (2022). Ethical AI: Addressing bias in machine learning models and software applications. *Computer Science & IT Research Journal, 3*(3), 115–126. https://doi.org/10.51594/csitrj.v3i3.1559

Park, M. S., Son, H., Hyun, C., & Hwang, H. J. (2021). Explainability of Machine Learning Models for Bankruptcy Prediction. *IEEE Access, 9*, 124887–124899. https://doi.org/10.1109/ACCESS.2021.3110270

Qureshi, N. I., Choudhuri, S. S., Nagamani, Y., Varma, R. A., & Shah, R. (2024). Ethical Considerations of AI in Financial Services: Privacy, Bias, and Algorithmic Transparency. *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, 1–6. https://doi.org/10.1109/ICKECS61492.2024.10616483

Redelmeier, A., Jullum, M., & Aas, K. (2020). Explaining Predictive Models with Mixed Features Using Shapley Values and Conditional Inference Trees (pp. 117–137). https://doi.org/10.1007/978-3-030-57321-8_7

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1). https://doi.org/10.1609/aaai.v32i1.11491

Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences, 3*(2), 272. https://doi.org/10.1007/s42452-021-04148-9

Saarela, M., & Podgorelec, V. (2024). Recent Applications of Explainable AI (XAI): A Systematic Literature Review. *Applied Sciences, 14*(19), 8884. https://doi.org/10.3390/app14198884

Sahakyan, M., Aung, Z., & Rahwan, T. (2021). Explainable Artificial Intelligence for Tabular Data: A Survey. *IEEE Access, 9*(June), 135392–135422. https://doi.org/10.1109/ACCESS.2021.3116481

Shuford, J. (2024). Exploring Ethical Dimensions in AI: Navigating Bias and Fairness in the Field. *Journal of Artificial Intelligence General Science (JAIGS) ISSN:3006-4023, 3*(1), 103–124. https://doi.org/10.60087/jaigs.vol03.issue01.p124

Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. 2022 *ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250. https://doi.org/10.1145/3531146.3534639

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Farina, M. (2021). A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access*, 9, 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

Tian, Y., & Liu, G. (2020). MANE: Model-Agnostic Non-linear Explanations for Deep Learning Model. *2020 IEEE World Congress on Services (SERVICES)*, 33–36. https://doi.org/10.1109/SERVICES48979.2020.00021

Tyagi, S. (2022). Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions. September. http://arxiv.org/abs/2209.09362

Ullah, I., Rios, A., Gala, V., & Mckeever, S. (2021). Explaining Deep Learning Models for Tabular Data Using Layer-Wise Relevance Propagation. *Applied Sciences, 12*(1), 136. https://doi.org/10.3390/app12010136

Vilone, G., & Longo, L. (2020). Explainable Artificial Intelligence: a Systematic Review. May.

http://arxiv.org/abs/2006.00093

Watson, D. (2022). Rational Shapley Values. 2022 ACM Conference on Fairness, Accountability, and Transparency, 1083–1094. https://doi.org/10.1145/3531146.3533170

Wojtas, M., & Chen, K. (2020). Feature Importance Ranking for Deep Learning. *Advances in Neural Information Processing Systems,* 2020-Decem(33), 5105–5114. http://arxiv.org/abs/2010.08973

Wu, T.-Y., & Wang, Y.-T. (2021). Locally Interpretable One-Class Anomaly Detection for Credit Card Fraud Detection. *2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, 25–30. https://doi.org/10.1109/TAAI54685.2021.00014

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness Beyond Disparate Treatment &amp; Disparate Impact. *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180. https://doi.org/10.1145/3038912.3052660

Zhang, Z., Wu, C., Qu, S., & Chen, X. (2022). An explainable artificial intelligence approach for financial distress prediction. *Information Processing & Management, 59*(4), 102988. https://doi.org/10.1016/j.ipm.2022.102988

Zou, Y., Gao, C., & Gao, H. (2022). Business Failure Prediction Based on a Cost-Sensitive Extreme Gradient Boosting Machine. *IEEE Access,* 10, 42623–42639. https://doi.org/10.1109/ACCESS.2022.3168857