Artificial Intelligence Review (2024) 57:216
https://doi.org/10.1007/s10462-024-10854-8

Explainable artificial intelligence (XAI) in finance: a systematic literature review

Jurgita Černevičienė[1] · **Audrius Kabašinskas**[1]

Accepted: 3 July 2024 / Published online: 26 July 2024

Abstract
As the range of decisions made by Artificial Intelligence (AI) expands, the need for Explainable AI (XAI) becomes increasingly critical. The reasoning behind the specific outcomes of complex and opaque financial models requires a thorough justification to improve risk assessment, minimise the loss of trust, and promote a more resilient and trustworthy financial ecosystem. This Systematic Literature Review (SLR) identifies 138 relevant articles from 2005 to 2022 and highlights empirical examples demonstrating XAI's potential benefits in the financial industry. We classified the articles according to the financial tasks addressed by AI using XAI, the variation in XAI methods between applications and tasks, and the development and application of new XAI methods. The most popular financial tasks addressed by the AI using XAI were credit management, stock price predictions, and fraud detection. The three most commonly employed AI black-box techniques in finance whose explainability was evaluated were Artificial Neural Networks (ANN), Extreme Gradient Boosting (XGBoost), and Random Forest. Most of the examined publications utilise feature importance, Shapley additive explanations (SHAP), and rule-based methods. In addition, they employ explainability frameworks that integrate multiple XAI techniques. We also concisely define the existing challenges, requirements, and unresolved issues in applying XAI in the financial sector.

Keywords Explainable artificial intelligence (XAI) · Finance · Financial data science · Explainability · Interpretability · Decision making

1 Introduction

Artificial Intelligence (AI) has made remarkable progress in recent years, with notable implementations by tech giants like Google, Meta (formerly Facebook), and Amazon. These companies have integrated AI into various services and products, such as Google's

✉ Jurgita Černevičienė
jurmark@ktu.lt

Audrius Kabašinskas
audrius.kabasinskas@ktu.lt

[1] Department of Mathematical Modelling, Faculty of Mathematics and Natural Sciences, Kaunas University of Technology, Kaunas, Lithuania

AI-powered search algorithms (Diamant 2017), Meta's recommendation systems (Hutson 2020), and Amazon's intelligent robotics applications (Diamant 2017). Subsequently, AI will have an even more significant impact in the future (Makridakis 2017). The AI market in finance is expanding globally and is predicted to continue growing (OECD 2021). Despite these achievements, the increasing complexity of AI models has resulted in the use of opaque black box models, which reduces their interpretability (Guidotti et al. 2018). As a result, it is pivotal to develop and evaluate solutions to address this problem, especially in sensitive areas such as finance, where high accuracy is not the only requirement for expanding the adoption of AI systems.

Machine Learning (ML) is classified as a subset of Artificial Intelligence (AI) due to its emphasis on algorithms that autonomously generate predictions and decisions by learning from data (Rius 2023; Yao and Zheng 2023; Pokhariya et al. (2022). AI has a wider range of applications than only ML. Some AI systems, for instance, may be built upon an ML model but not be entirely dependent on it; conversely, other systems may function solely on rule-based systems and not employ any machine-learnt models (Sprockhoff et al. 2023; Liubchenko 2022). Therefore, it is correct to assert that ML is a component of AI, but not all AI applications exclusively depend on ML methods, thereby demonstrating the variety and complexity within the wider domain of AI.

Explainable Artificial Intelligence (XAI) offers a means to explore, explain, and comprehend intricate systems. However, evaluating and determining the explicable nature of an explanation is a complex and challenging issue. Providing understandable and interpretable financial explanations is vital to establish trust and accountability in decision making. XAI research underscores the importance of explaining decisions in finance, and authors often compare finance and medicine to emphasise the need for explanations in both fields (Silva et al. 2019; Sovrano and Vitali 2022).

The ML community has not yet reached an agreement on the specific criteria that should be employed to determine explainability. Attempts have been made to clarify concepts such as trustworthiness, interpretability, explainability, and reliability (Gilpin et al. 2018; Ribeiro et al. 2016; Lipton 2018; Došilović et al. 2018). Although numerous studies distinguish between interpretability and explainability, the two terms are frequently used interchangeably (Linardatos et al. 2020). Explainable AI and interpretability are often characterised by imprecise definitions, which can lead to misleading conclusions (Rudin 2019). Interpretability is the degree to which a human can understand and explain the cause of a decision made by a model (Cartwright 2023; Zeng et al. 2023; Doshi-Velez and Kim 2017; Miller 2019). Explainability refers to the degree of transparency in the results produced by the model, while interpretability enhances its utility by providing users with comprehension of the underlying concepts of the model, thus enabling them to grasp the logic and algorithms governing the AI system (Linardatos et al. 2020). Conversely, explainability concerns the extent to which individuals can comprehend the internal reasoning and mechanisms of a machine learning system. The explainability of a model is positively correlated with the level of understanding exhibited during its training or decision-making processes regarding these internal processes (Linardatos et al. 2020). According to Gilpin et al. (2018), explainability is a fundamental requirement in addition to interpretability, which is inadequate on its own. According to research conducted by Linardatos et al. (2021) and Doshi-Velez and Kim (2017), this study posits that interpretability surpasses explainability as a more comprehensive concept. However, to avoid confining their applicability to particular contexts, we employed the terms interpretable and explainable interchangeably.

Relying on systems without transparent decision-making procedures is not trustworthy. In finance, XAI is of highest priority due to the intricate and complex structure of its

models, as well as the potential effects of incorrect or biased choices, where AI is used for tasks such as credit scoring, bankruptcy forecasting, fraud detection or portfolio optimisation. Other areas of interest include exploring how XAI can improve investment decision-making and portfolio management, or examining the ethical and legal implications of using opaque AI systems in financial services. Additionally, XAI techniques can enhance the comprehensibility of predictive models in various areas such as loan underwriting, insurance pricing, stock price prediction, and regulatory compliance.

The terms ML and AI are also used interchangeably to refer to AI methods used in research. This approach allows for a more comprehensive view of the field as it encompasses a wide range of techniques that belong to AI. Using these terms interchangeably ensures clarity and consistency throughout the study and allows researchers to communicate their findings to a broader audience more effectively.

We categorise the XAI methods employed in finance based on their specific financial applications, identify existing research gaps, and propose potential areas for future investigation. The purpose of this research is threefold: First, we report the review results that examine the current advances and trends in XAI for various financial tasks by summarising the techniques used; second, we help practitioners select the appropriate XAI technique that fits their particular financial use case by offering the categorisation; and third, to help researchers pinpoint the specific areas and applications using XAI methods that need further investigation. Our work also makes other significant contributions, as we review a large body of recent and related work and attempt to implement all explainability features previously explored. A total of 138 chosen studies were evaluated based on four review questions (RQs):

- Which financial AI applications have benefited from the implementation of XAI?
- How do the XAI methods differ depending on the application and task?
- How are new XAI methods developed and applied?
- What are the challenges, requirements, and unresolved issues in the literature related to the application of XAI in finance?

The structure of this article can be summarised as follows: Background section provides a concise introduction to the critical advancements in XAI and explains essential terminology. It is designed for readers who may need to gain a broad understanding of XAI. The methodology for conducting a systematic review can be found in Sect. 3. This section describes the search for relevant scientific articles, gathering necessary data, and conducting a thorough analysis. The systematic analysis of selected scientific articles based on the research questions outlined earlier results are presented in Sect. 4. Sections 6,7 and 8 contain the discussions, future work, recent developments, and conclusions.

2 Background

2.1 XAI in general

To promote the integration of AI, XAI strives to improve transparency by devising methods that empower end users to understand, place confidence in and efficiently control AI systems (Adadi and Berrada 2018; Arrieta et al. 2020; Saeed and Omlin 2023). Early work on rule-based expert systems identified the need for explanations (Biran and

Cotton 2017). Consequently, the advent of Deep Learning (DL) systems pushed XAI to the forefront of academic inquiry (Saeed and Omlin 2023). Many practical challenges now employ ML models that require higher predictive accuracy (e.g., stock price prediction (Carta et al. 2021), online banking fraud detection (Achituve et al. 2019), bankruptcy prediction (Cho and Shin 2023). Increasing the complexity of the model often leads to higher predictive accuracy, but also reduces its capacity to provide explainable predictions (Linardatos et al. 2020). Black-box models, including deep learning models such as deep neural networks (DNNs) or generative adversarial networks (GANs), as well as ensemble models like XGBoost and Random forests, achieve the highest level of performance. However, these models are difficult to explain, as noted by Yang et al. (2022a, b) and Linardatos et al. (2020). Conversely, white-box or intrinsic models, such as linear models, rule-based models, and decision trees, provide a straightforward structure that simplifies the interpretation of outcomes.

We will primarily focus on presenting the fundamental concepts of XAI, as several comprehensive surveys have been conducted to categorize and elucidate it (see Arrieta et al. 2020; Adadi and Berrada 2018; Belle and Papantonis 2021). There are a number of terms that are frequently used in academic discourse and discussions related to XAI systems to define the key characteristics of such systems. The methods vary according to their stage (ante-hoc, post-hoc), scope (local, global), type (model-specific, model-agnostic), and area of application (finance, medicine, education, transportation, ecology, agriculture, etc.).

Explanations consist of two stages: The *ante-hoc stage* involves the interpretation and analysis of data using explanatory data, specifying data sources, and constructing a model. Assessing the data at this point is essential as it provides significant insights and improves the understanding of the model (Saranya and Subhashini 2023). These techniques are known as white-box approaches because they are deliberately engineered to preserve a fundamental structure. They are intrinsic and do not require explanation methods. In the post-hoc stage, an explanation is generated after the ML model has been constructed and necessitates an additional explanation method (Hassija et al. 2024). Explainability is categorised as local or global based on whether the explanation is derived from a particular piece of data (local) or the entire model (global) (Vilone and Longo 2021; Angelov et al. 2021).

Model-agnostic methods function autonomously without being dependent on any specific ML model. Typically, these techniques are post-hoc and aim to tackle the problem of comprehending intricate models such as Convolutional Neural Networks (CNN). They provide explanations that are not limited to a particular model structure, making them adaptable and broadly applicable to various types of model (Zolanvari et al. 2021). On the contrary, the applicability of the model-specific method is dependent on the structure of the specific model and is restricted to that particular model (Kinger and Kulkarni 2024).

Ethical, legal, and practical factors drive the desire for explainability (Eschenbach 2021). The European Union (EU) regulates AI through the AI Act. The regulation requires the implementation of specific criteria for the development and use of XAI (European Commission 2020). Achieving explainability through the creation of advanced and intricate XAI models is a challenging undertaking, and several barriers impede this process. Frasca et al. (2024) highlight the trade-off between performance and explainability, stating that higher accuracy often requires complex models. They address challenges such as finding pivotal characteristics, scaling issues, and model acceptability, highlighting the need for significant computational resources, and aligning explanations with user intuitions.

2.2 Artificial intelligence in finance

The significance of AI becomes evident when one observes the concerted efforts of governments to regulate its usage and practical implementation. One notable initiative in this regard is establishing the European Commission's High-Level Expert Group on AI(HLEG) in 2020. HLEG suggests a comprehensive definition of AI and sets ethical standards for developing and deploying robust artificial intelligence systems. The primary objective of HLEG is to propose recommendations for policy enhancements and address AI's social, ethical, and legal dimensions. As defined by HLEG, AI encompasses both software and hardware components, which collaborate to collect and analyse data from the environment. The AI system acquires knowledge through this analysis and formulates decisions to achieve specific objectives. The adaptability of the AI system is formed by evaluating past actions and their consequences within the operational environment. This evaluation can be accomplished using either symbolic rules or numerical models (Samoili et al. 2020). Overall, the efforts undertaken by governments and expert groups like HLEG exemplify the recognition of AI's significance and the commitment to ensuring its responsible and beneficial integration into various domains.

AI has become a game-changing and innovative tool in many industries, including finance. The ability to predict bankruptcy is crucial for financial institutions that use artificial intelligence approaches. This task requires careful attention. Several studies (Verikas et al. 2010; Balcaen and Ooghe 2006; Dimitras et al. 1996) have demonstrated the immense potential of AI in revolutionising decision-making, mitigating risks, and enhancing profitability within the financial domain. Leveraging the capabilities of AI empowers financial institutions to gain a competitive advantage and elevate their customer service offerings. The work of Cao (2021) provided a thorough, multidimensional, and problem-oriented economic-financial overview of recent research on AI in finance, which formed the basis for the financial domain categories.

Despite the potential of AI to improve and develop financial products, there are challenges to taking full advantage of innovative algorithms (Harkut and Kasat 2019). In their recent study, Eluwole and Akande (2022) identified the main difficulties in using AI in the financial sector. These challenges include aspects such as accuracy, consistency, transparency, trust, ethics, legal considerations, governance regulations, competence gaps, localisation issues, and the intricacies of ML design and integration.

Researchers often employ several types of ML techniques to evaluate and improve the efficiency of their models (Dastile et al. 2022; Zhu et al. 2022; Wang et al. 2019). Therefore, we employed the term "multi-approach AI technique" to denote the utilisation of multiple AI methods to address one financial task.

2.3 XAI in finance

The Royal Society (2019) states that as the use of AI technologies in decision-making processes increases, individuals need to understand how AI works. This need arises from concerns about bias, compliance with policies and regulations, and the ability of developers to understand AI systems. Miller (2019) specifies that improved explainability of AI can increase confidence in the results. Arrieta et al. (2020) suggest that interpretability can also ensure fairness, reliability, and truthfulness in AI decision making

and that XAI aims to create more understandable models while maintaining efficiency and allowing humans to trust AI systems and control them.

Adadi and Berrada (2018) have identified different explanations, perspectives, and motives for using them, including justification, control, discovery, and improvement of classification or regression tasks. These perspectives encompass an expansive spectrum, ranging from justifying and controlling AI/ML approaches to discovering new insights and improving the accuracy of classification or regression tasks. Ensuring the interpretability of AI/ML approaches becomes crucial in identifying the input features that significantly influence the outcomes. Once fully understood, a model can be combined with specialised knowledge to generate an advanced model with enhanced capabilities. Credit scoring and risk management are frequently researched topics in XAI finance research (Demajo et al. 2020; Chlebus et al. 2021; Misheva et al. 2021; Bussmann et al. 2020).

2.3.1 Explainability categories

The comprehensive evaluation of XAI methods, encompassing numerous techniques, procedures, and performance measurements, has been investigated and reported in published surveys such as Doshi-Velez and Kim (2017) and Zhou et al. (2021). To categorise explanatory approaches, Belle and Papantonis (2021) developed a classification system widely used to analyse and compare different XAI techniques and practices and provides valuable information on the advantages and limits of each approach.

Feature relevance explanation: The main goal of feature importance explanation is to evaluate the influence of a feature on the model outcome, and one of the significant contributions to this field, particularly in XAI, is the Shapley Additive Explanations (SHAP) method (Lundberg and Lee 2017). In this method, a linear model is constructed around the instance to be explained, and the coefficients are interpreted as the significance of the features. However, this approach is seen as an indirect means of generating explanations as it only focusses on the individual contribution of a feature and does not provide information about the dependencies between features. In cases where there are strong correlations between features, the resulting scores may indicate inconclusive or contradictory data. When conducting an analysis, it is imperative to consider this aspect. For example, the study by Torky et al. (2024) presents a unique XAI model that has the ability to independently identify the underlying reasons behind financial crises and clarify the process of selecting relevant features. The authors used the pigeon optimiser to enhance the feature selection process. Subsequently, a Gradient Boosting classifier is utilised, employing a subset of the most influential attributes, to determine the sources of financial crises.

Explanation by simplification: Explanations by simplification approximate complex models to simpler ones. The main explanation for the simplification challenge is to ensure that the simpler model is flexible enough to accurately approximate the complex model and increase its effectiveness by comparing the accuracy of classification problems. Rule-based learners and decision tree techniques are simplification methods for model-agnostic explanations. Model-specific explanations can also use distillation, rule-based learners, and decision trees to provide simplified explanations. An example of such an explanation category is the Weighted Soft Decision Forest (WSDF). This method combines multiple soft decision trees by assigning weights to each tree's output. The purpose is to simulate the credit scoring process. Another example is credit risk assessment decision support systems that utilise the recursive rule extraction algorithm

and decision tree. These systems generate interpretable rules for machine learning-based credit evaluation techniques (Hayashi 2016; Zhang et al. 2020a, b, c).

Local explanation: Explaining model predictions in a specific case at a statistical unit level is known as local explainability. This approach can be helpful for companies and individuals in identifying the crucial factors that contribute to their financial challenges (Hadash et al. 2022). Local explanation methods include rule-based learning, linear approximations, and counterfactuals. The Local Interpretable Model-Agnostic Explanations (LIME) technique developed by Ribeiro et al. (2016) is a commonly used ML tool that can help understand black-box model behaviour and identify the key features used to make predictions by eliminating perturbations induced on the inputs. Although LIME is a locally linear model, its accuracy may be limited due to dependence on an external model. Counterfactual explanations describe causal scenarios where Event A's absence would result in Event B's absence, providing an understanding of model predictions and recommendations for future action (Johansson et al. 2016). Park et al. (2021) examined the problem of predicting bankruptcy by showcasing the ability to reproduce the measurement of feature importance in tree-based models using LIME in bankruptcy datasets. They highlighted the potential to extract significant feature importance from other models that lack built-in feature measurement capabilities but demonstrate higher accuracy.

Visual explanation: The purpose of visual explanation is to create visual representations. This method simplifies understanding how a model works, even when dealing with complex dimensions. These methods grasp the decision boundary and the interaction among features, making visualisations a helpful tool, particularly for individuals without professional expertise. The proposed decision support tool for lending decisions, developed by Chen et al. (2022), employs interactive visualisations to clarify the structure and behaviour of time series models. It allows users to investigate the model and gain a precise understanding of how conclusions are derived.

Transparent models: Models that are transparent in their decision-making process are valuable in improving understanding. Using transparent models can improve understanding of how inputs relate to outputs. It allows to comprehend the correlation between the data provided and the outcomes generated by the model. Linear regression models illustrate transparent models because the model coefficients indicate the significance or weight of each input feature in producing the outcome and explain how the additional input unit linearly affects the output. Decision trees and rule-based learners are also transparent models because they use a transparent set of decision-making rules that can be examined to determine how the output was generated. K-nearest-neighbour models can also be transparent, depending on the simplicity of the distance metric used to determine nearest neighbours and allow the model to provide a transparent representation of its decision-making process (Clement et al. 2023).

Frameworks: A framework is commonly understood as a methodology incorporating multiple XAI methods to complete a single task (Linardatos et al. 2020). To assess a model's performance or demonstrate its approximation capabilities, researchers frequently use multiple XAI methods. Using various methods can improve the explanation and make comparing different methods easier. For example, Chen et al. (2022) created a machine learning model that is globally interpretable. This model includes an interactive visualisation and provides different summaries and explanations for each decision made. It offers case-based reasoning explanations by considering similar past cases, identifies the key features that influence the model's prediction, and provides customised concise explanations for specific lending decisions.

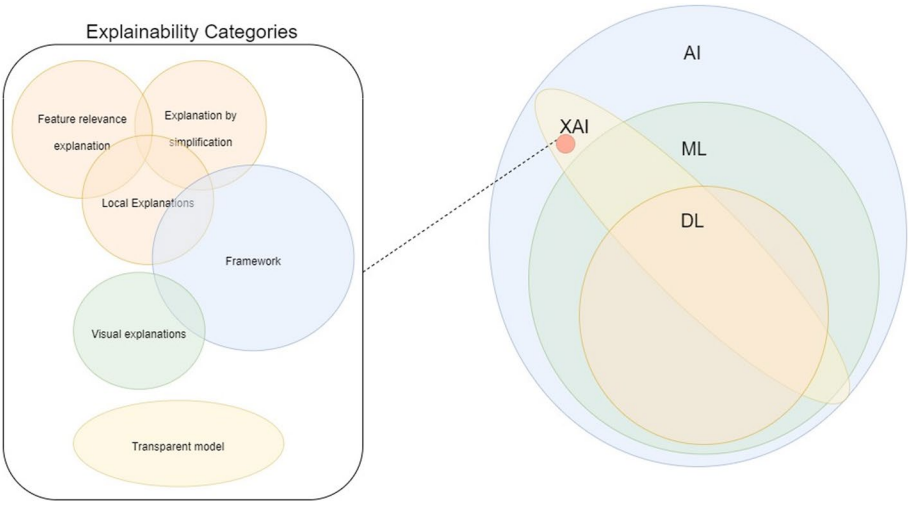

Fig. 1 Venn diagram illustrates the relationship between AI and XAI

The summary of the different categories and their connections in the XAI landscape within the finance industry is illustrated in Fig. 1. The visual representation aims to improve comprehension of the intricate XAI techniques and their uses in finance.

3 SLR methodology

The systematic review followed the principles of Kitchenham and Charters (2007), which ensured a credible and verified framework for performing a thorough examination. The review process consisted of three main phases: planning, conducting, and preparing the report, as shown in Fig. 2.

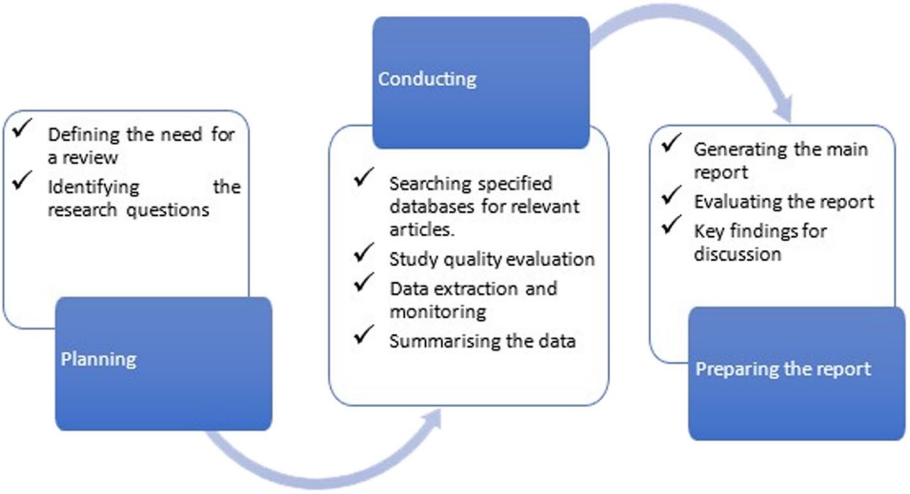

Fig. 2 SLR process stages

3.1 Identifying the research questions

A systematic literature review is proposed to examine the different methods used to provide XAI in the financial industry. The primary objective of this review is to provide valuable information on the various techniques employed across multiple domains and tasks and their corresponding evaluations. As shown in Table 1, four research questions were formulated to guide this literature review.

These questions are designed to systematically address the key aspects of XAI in finance, which include the financial tasks to which these methods are applied, the novel XAI approaches that have been developed and implemented, and the obstacles encountered during their execution.

After identifying the research question, we provide a detailed explanation of the methodology used to search the relevant literature. This initial step provides comprehensive details on the databases employed in the search process, the search strategies and terms implemented, and the criteria applied to evaluate the relevance and value of each article chosen for review. An analysis of the primary article metadata, containing bibliometric information about the articles, is also provided. These data contain a variety of essential information, such as the title of the article, the author's name, the publication date, the journal where the article is published, the number of citations obtained by the article, the keywords used, and the country of the authors (Donthu et al. 2021).

3.2 Searching specified databases for relevant articles

The initial starting point for a literature search is to identify suitable digital sources and literature databases. This approach facilitates a targeted keyword search, leading to the discovery of pertinent studies (Levy and Ellis 2006). An extensive search of the Scopus and Web of Science (WOS) databases from 2005 to 2022 was conducted during the research due to their comprehensive coverage of finance and AI topics (Zhu and Liu 2020; Pranckutė 2021) applications. We have selected these databases for their rich content and because they include highly respected financial journals (Singh et al. 2021). These databases ensure that peer-reviewed articles published in leading international journals and conference proceedings are included, which helped us maintain higher quality standards. These databases offer a comprehensive range of bibliometric analysis tools, enabling users to access and export bibliographic data customised to their research requirements.

We conducted queries on the two databases outlined earlier to collect articles, specifically targeting sources recognised for their comprehensive coverage of relevant literature (Table 2). We performed a comprehensive literature search using meticulously

Table 1 List of the research questions

Research question	Description
RQ1	Which financial AI applications have benefited from the implementation of XAI?
RQ2	How do the XAI methods differ depending on the application and task?
RQ3	How are the new XAI methods developed and applied?
RQ4	What are the challenges, requirements, and unresolved issues in the literature related to the application of XAI in finance?

Table 2 Article collection query

Database	Searching query	Number of articles
Scopus	(TITLE-ABS-KEY ("Asset evaluation" OR "banking" OR "corporate performance" OR "credit risk" OR "finance" OR "financial performance" OR "portfolio optimisation" OR "Portfolio selection" OR "capital budgeting" OR "financial planning" OR "bankruptcy" OR "credit scoring" OR "financial distress" OR "credit" OR "loan" OR "biotech" OR "Financial Advice")) AND (KEY ("explainable artificial intelligence" OR "xai" OR "interpretable machine learning" OR "explainable ai" OR "explainable machine learning" OR "explainable artificial intelligence (xai)" OR "interpretability" OR "ai transparency")) AND (EXCLUDE (SUBJAREA, "CENG")) AND (EXCLUDE (SUBJAREA, "MATE") OR EXCLUDE (SUBJAREA, "ARTS") OR EXCLUDE (SUBJAREA, "MEDI")) AND (EXCLUDE (DOCTYPE, "re") OR EXCLUDE (DOCTYPE, "ch")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "MATH") OR LIMIT-TO (SUBJAREA, "ENGI") OR LIMIT-TO (SUBJAREA, "DECI") OR LIMIT-TO (SUBJAREA, "BUSI") OR LIMIT-TO (SUBJAREA, "ECON")) AND (EXCLUDE (SUBJAREA, "ENER") OR EXCLUDE (SUBJAREA, "NEUR") OR EXCLUDE (SUBJAREA, "PHYS")) AND (LIMIT-TO (PUBSTAGE, "final"))	221
WOS	"financ*" OR "Asset evaluation" OR "banking" OR "corporate performance" OR "credit risk" OR "finance" OR "financial performance" OR "portfolio optimisation" OR "Portfolio selection" OR "capital budgeting" OR "financial planning" OR "bankruptcy" OR "credit scoring" OR "financial distress" OR "credit" OR "loan" OR "biotech" OR "Financial Advice"(All Fields) and "explainable artificial intelligence" OR "xai" OR "interpretable machine learning" OR "explainable ai" OR "explainable machine learning" OR "explainable artificial intelligence (xai)" OR "interpretability" OR "ai transparency" OR "interpretable" (Author Keywords)	552

chosen keywords and filters to guarantee the relevance of the articles' content. We carefully chosen the keywords to cover a broad spectrum of studies that are pertinent to our study issue. This comprehensive search was designed to find all relevant research while limiting the inclusion of irrelevant literature.

A total of 773 papers were discovered in the initial search results. Subsequently, we applied a screening approach to assess the suitability of these results for inclusion in our review.

When conducting an SLR various inclusion and exclusion criteria were used to filter relevant articles (Table 3). The inclusion criteria consisted of identifying articles that contained keywords from Table 2 and were relevant to directly assessing XAI applications in finance. Only peer-reviewed articles in English and with full text availability in databases were considered for inclusion. However, the review excluded articles related to the philosophy of XAI, technical reports, review articles, and duplicates. This was made to ensure that the review articles provided concrete examples and empirical evidence for XAI applications in finance. Technical reports often contain highly specialised information, and review articles summarise existing research that serves a purpose other than the review. Finally, editorials and opinions or viewpoints on a particular topic were also excluded from the review.

The process of conducting a review comprises four primary stages: identification, screening, eligibility determination, and categorisation of potential research articles. The PRISMA diagram (Page et al. 2021) systematically documented the procedure, with Fig. 3 illustrating each phase of the literature search. The identification phase involved searching the Scopus and Web of Science databases for related articles from 2005 to 2022 using specific keywords relevant to the topic. The screening phase then involved reviewing the titles and abstracts to exclude unrelated studies. During the eligibility phase, we conducted full-text articles to determine whether they met the established inclusion criteria. Finally, the sorting phase involved selecting the final articles for the review. The PRISMA chart provided an explicit and transparent overview of the entire process, allowing for easy analysis and replication of the study.

By following the steps mentioned above, we can provide a comprehensive and well-substantiated evaluation, eventually contributing to the advancement of expertise and understanding in the field of financial data science.

3.3 Descriptive analysis of the data

This section presents the results of the quantitative data analysis performed by the selected publications. The research was carried out mainly by analysing bibliometric data using R Biblioshiny, which is a graphical web-based user interface for the Bibliometrix software

Table 3 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
The application of XAI methods has been proposed as a means of achieving various finance-related tasks; Articles in English; The work must be an original article in a peer-reviewed journal or conference proceedings;	Articles related to the XAI philosophy, technical reports, a book, book chapter, review articles; Editorials and opinions or viewpoints; Duplicated articles

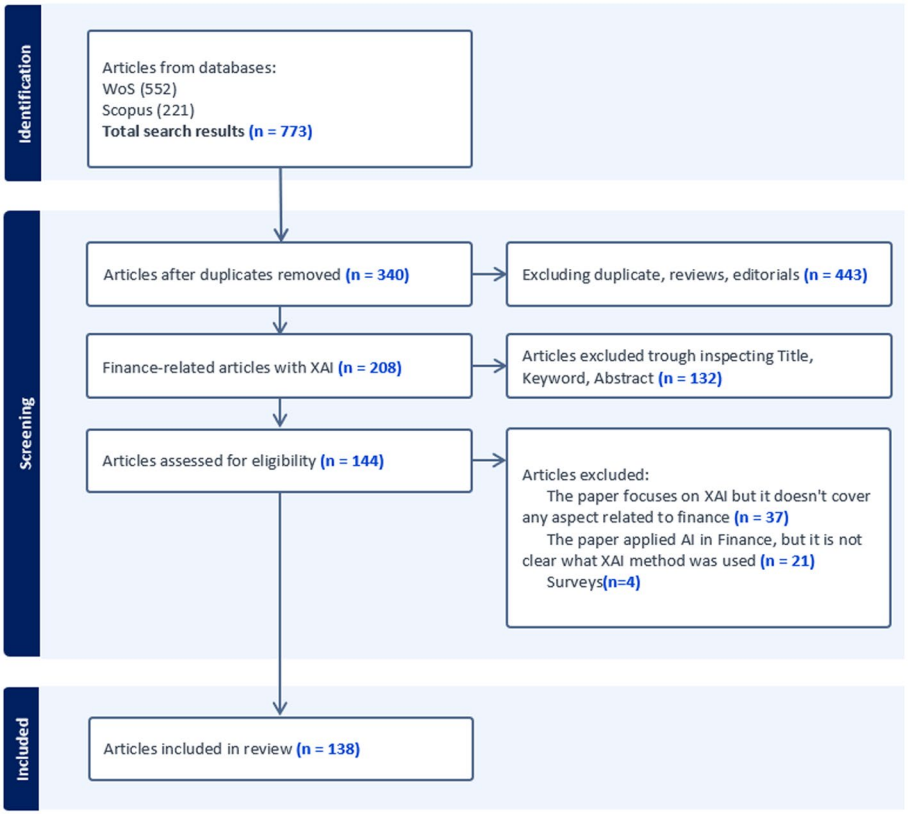

Fig. 3 The PRISMA flowchart provides a detailed checklist and flow diagram of how the literature search was conducted

version. This application enables the visual representation of statistical data from articles, making it easier to visually analyse the data (Aria and Cuccurullo 2017).

Table 4 presents a concise overview of the bibliometric data collected from 138 published works by 397 authors. 77 were published in peer-reviewed journals, and the remainder were presented in conference papers and proceedings. Based on the data, there has been a significant increase in the number of studies concentrating on XAI in finance, with 80% of the reviewed works published between 2020 and 2022. This upward trend highlights the growing interest in implementing XAI in the financial sector and underscores the need for further research to explore its possibilities and address its challenges.

All articles that were reviewed in this research used author-defined keywords to aid in the indexing process within bibliographic databases. The word cloud technique was used to compare the keywords with the ones extracted from the abstracts. Figure 4 highlights the frequency of words with different font sizes and colours.

The word cloud clearly highlights the author-defined keywords, with explainable AI, credit scoring, deep learning, and credit risk being the most prevalent. Conversely, the less prominent keywords include cost-sensitivity, financial forecasting, explainability, and credit score prediction. The statement demonstrates that research publications prioritise the advancement of deep learning, with a particular focus on credit evaluation and risk

Table 4 Summary statistics of the collected articles

Main information about the data	
Time interval	2005:2022
Sources	94
Papers	138
The average age of the paper	2.95
Average citations per paper	7.121
Keywords	
Keywords Plus	694
Author's Keywords	389
Authors	
Authors	391
Authors of single-authored papers	7
Authors collaboration	
Co-Authors per paper	3.38
International co-authorships %	10
Document types	
Article	77
Conference paper	61

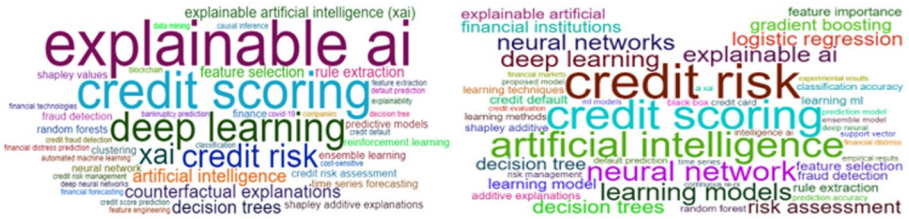

Fig. 4 A word cloud displaying **a** author-defined keywords and **b** keywords extracted from abstracts

management. The word cloud containing abstract-extracted keywords reveals that the terms credit risk, credit scoring, artificial intelligence, and neural networks are the most prevalent, while random forest, prediction accuracy, and financial markets are less prevalent.

A keywords analysis has helped us better understand our chosen subjects. The trend topics plotted in Fig. 5 visually represent the progress of the XAI approach and relevant finance topics. The graph showcases each term's year and frequency of usage through bubbles. The size reflects the relative usage frequency for the respective term. This analytical tool is invaluable for professionals keen to explore their field of study. Researchers can pinpoint areas that require further investigation by identifying recent trends and issues. Furthermore, this resource highlights research areas that link XAI in finance with other pertinent topics, such as deep learning and risk management. These articles offer critical insights and are at the forefront of current research, serving as a valuable foundation for future exploration in this field.

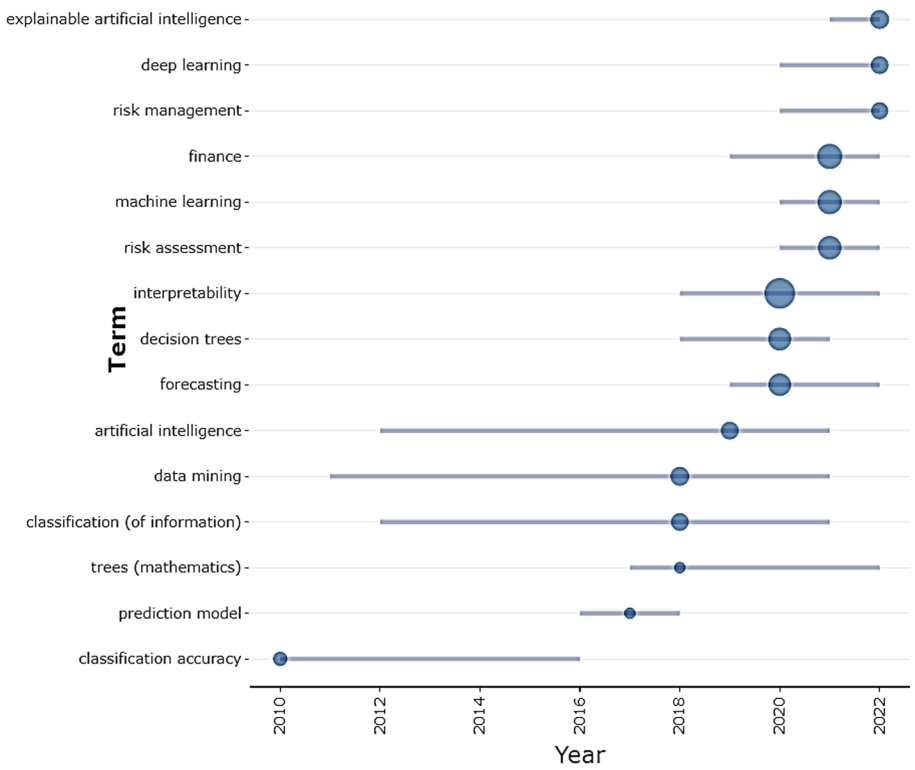

Fig. 5 Trend Topics plot

The geographical distribution of the number of publications on a global scale is illustrated on the map in Fig. 6. The grey regions indicate areas that do not have any publications during the specified time frame.

China produced the most publications (n = 33), followed by Italy (n = 17) and Germany (n = 11). According to the findings of this study, China emerged as the leading contributor,

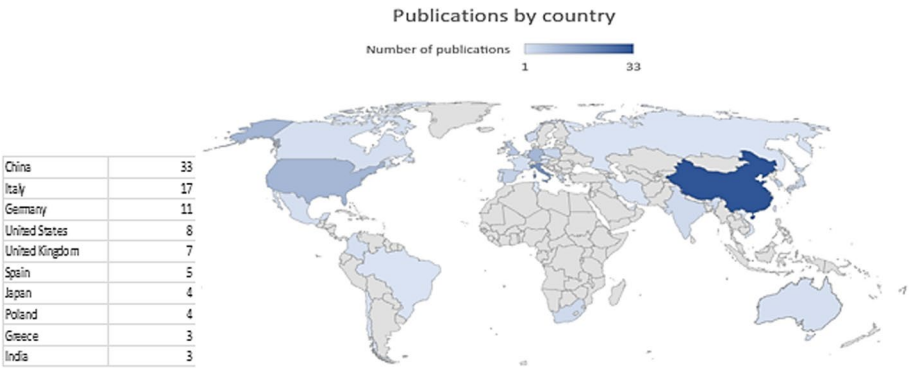

Fig. 6 Publications by country of the first author and top ten countries in terms of publication number

with 33 published articles that showcase their significant involvement and contribution to the respective research domain. Italy and Germany follow in second and third place with 17 and 11 releases, respectively. These results indicate that China has a strong presence of XAI in finance, while Italy and Germany also create essential contributions to academic discourse.

4 Review results and main findings

After reviewing some relevant features of the publications, we will next present the results of a comprehensive study of 138 articles that address the aforementioned questions. The results represent the state of XAI research in the financial sector and are based on (i) which financial AI applications have benefited from the implementation of XAI?, (ii) how the XAI methods used differ depending on the application and task, (iii) how are new XAI methodologies being developed and applied, and (iv) what are the challenges, requirements, and unresolved issues in the literature related to the application of XAI in finance.

4.1 RQ1: which financial AI applications have benefited from the implementation of XAI?

We employ a categorisation approach, rooted in the paper of Cao (2021), to analyse the objectives of AI in the selected papers and its widespread application in finance. This categorisation includes several significant domains, including financial market analysis and forecasting, agent-based economics and finance, intelligent investment, optimisation, and management, innovative credit, loan and risk management, intelligent marketing analysis, campaign and customer care, and smart blockchain. A brief overview of the finance domains explored in the articles is presented in Fig. 7, along with the specific AI techniques used and the XAI approaches developed in response. The figure summarises the information provided in the articles. It gives a visual representation of the relationships

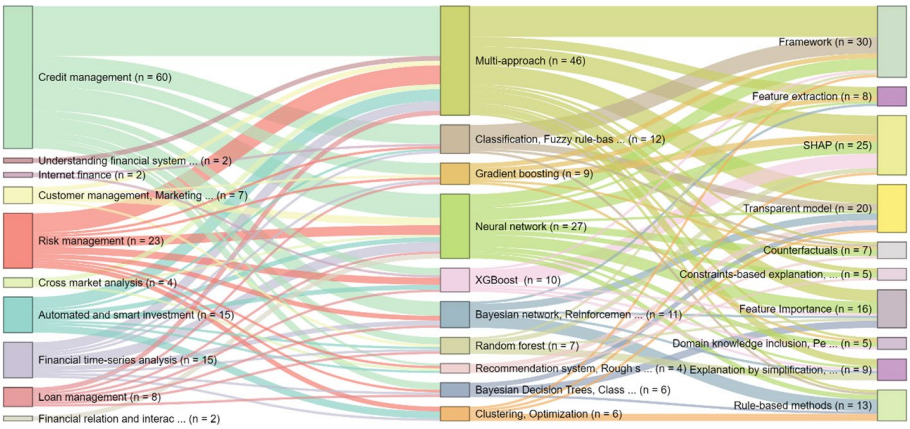

Fig. 7 The selected articles from different application domains were divided into clusters based on the finance area, the AI method, and the explanation form. The number of articles that fall under each of these categories is given in parentheses

between the various finance domains, AI techniques, and XAI methods used in the research. This enables a quick and efficient understanding of the essential findings and implications of financial AI and XAI research. Appendix A contains a complete table that provides a detailed analysis of the financial domains examined in the research, the specific AI methods used, and the corresponding references.

The selected articles are evaluated as structured, with the initial focus on the financial area, the specific applied methodology of artificial intelligence, and the corresponding prediction task. Our findings indicate that 44% of the selected articles were published in the domain of credit management. Additional topics covered by financial tasks include managing risks, automated and intelligent investments, financial time-series analysis, loan management, cross-market analysis, and other related tasks.

Identifying XAIs in finance domains has highlighted the importance of credit management as a critical undertaking, considering its substantial impact in several sectors of the financial industry. According to Moscato et al. (2021), good credit management is essential not only in the banking industry but also in other financial sectors. It ensures financial institutions' stability and profitability, while also facilitating the flow of capital for businesses and individuals. Conducting a credit management assessment requires many important steps, one of which is to incorporate XAI into the machine learning system as described in the articles. These tasks include assessing credit scores (62%), conducting risk assessments (35%), making well-informed credit decisions, and identifying suitable credit classifications.

A total of 23 academic papers have been published on the subject of risk management, encompassing a diverse array of financial risks. These risks include fraudulent activities, bankruptcies, financial risks, potential bank failures, and the need for precise identification of banknotes. Regarding the specific risks examined, fraudulent activities accounted for 39%, bankruptcies accounted for 13%, financial risks accounted for 34%, and the possibility of bank failure and the need for accurate banknote identification each accounted for 4%.

Automated and intelligent investing encompasses a wide range of financial investment decisions, the most prominent of which are portfolio management and optimisation (47%), and recommendation engine evaluation (13%). The remaining 40% comprises alternative methods of selecting financial investments (Fig. 8).

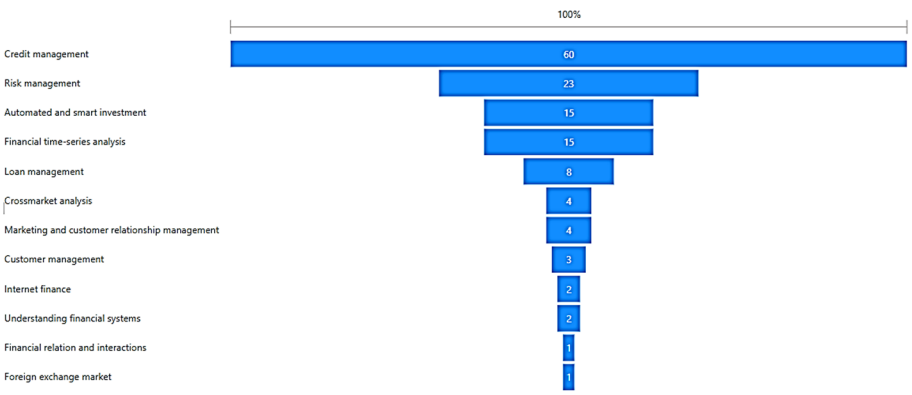

Fig. 8 Distribution of financial application

4.2 RQ2: How do the XAI methods differ depending on the application and task?

This section presents a simple classification system based on the explainability categories as they relate to different financial domains. Figure 9 visually represents the main types of explanation used in explanation design and documented in the literature (Belle and Papantonis 2021), specifically feature-relevance explanations, local explanations, visual explanations, as well as additional instances of simplification-based explanations and feature-relevance explanations. Several XAI techniques used in the literature have used various methods that fall into different XAI categories. Through our research, we have classified specific methodologies as frameworks. Our findings indicate that within the financial industry, most studies involving XAI focus on three key areas: Feature relevance explanation, Explanation by simplification, and Local explanation. Specifically, 45% of studies concentrate on feature relevance explanation, 18% on explanation by simplification, and 15% on local explanation.

Table 5 presents a detailed examination of several XAI techniques. The table provides comprehensive details about the scope, type, and specific association of each method with AI methodologies. According to the definition in Sect. 2.1, ante-hoc XAI methods are inherently designed to be interpretable and explainable. In contrast, post-hoc methods involve generating explanations after the ML model has been constructed and require an additional method for explanation. Moreover, explanations can be categorised as either local or global, depending on whether they are derived from a particular data point (local) or the whole model (global).

4.2.1 Feature relevance explanation

To explain the decision-making process of a particular method, feature-relevance explanations attempt to measure the impact of each input variable through quantification. Variables

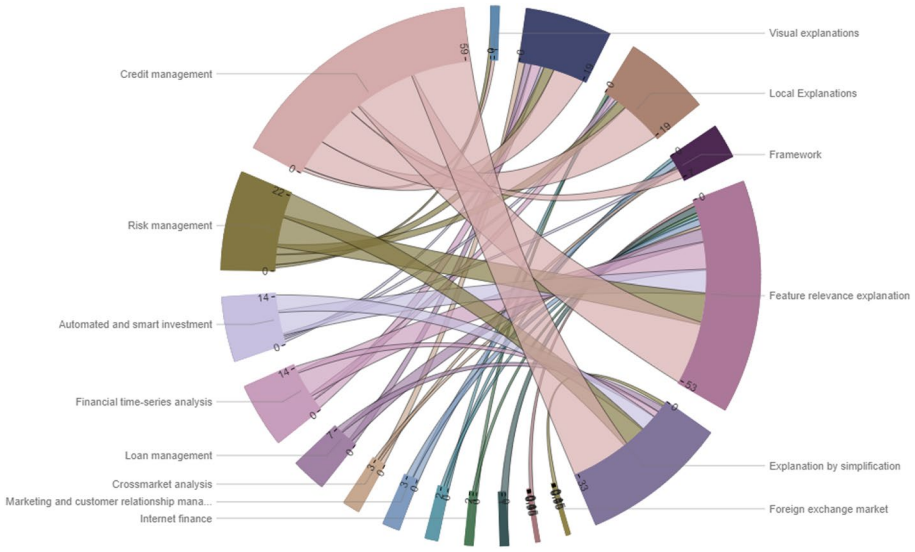

Fig. 9 The number of articles on the categories of XAI and their application in finance

Table 5 XAI techniques and the specific AI methods to which they relate

XAI method	Aim	AI method	Type	Scope
Constraints-based explanation	Loss forecasting and scenario stress testing	Neural Network Chen (2020)	Ad-hoc	Global
Counterfactuals	Credit scoring	Gradient boosting Bueff et al. (2022), Neural Network Pawelczyk et al. (2020)	Post-hoc	Local
	Credit scoring	Multi-approach Dastile et al. (2022), Neural network Crupi et al. (2022), Mohammadi et al. (2021)	Ad-hoc	Local
	Credit risk scoring	Regression Szepannek and Lübke (2021)	Ad-hoc	Global
	Risk performance	Support Vector Machine Gomez et al. (2020)	Ad-hoc	Local
Domain knowledge inclusion	Option pricing: predicting implied volatility surface	Neural Network Zheng et al. (2021)	Ad-hoc	Global
Explanation by simplification	Personal credit evaluation	Decision Tree Chen et al. (2020)	Ad-hoc	Global
	Credit, peer-to-peer platform	XGBoost Bussmann et al. (2021)	Post-hoc	Local
Feature extraction	Risk management—prediction of company listing status	Decision Tree Zhou et al. (2017)	Ad-hoc	Global
	Credit card fraud detection	Gradient boosting Tian and Liu (2020)	Post-hoc	Local
	Credit scoring	Gradient Boosting Liu et al. (2021a, b)	Post-hoc	Local/global
	Credit default	Multi-approach Li et al. (2020a, b)	Post-hoc	Local
	Credit risk assessment	Multi-approach Liu et al. (2021a, b)	Ad-hoc	Global
	Financial fraud	Multi-approach Hsin et al. (2021)	Post-hoc	Local
	Risk-averse portfolio selection (RPS) problem	Multi-approach Zhu et al. (2022)	Post-hoc	Local
	Financial forecasting of currencies	Neural Network Xu et al. (2021a, b)	Post-hoc	Local
Feature importance	Credit Risk Assessment for small and micro-sized enterprises	Classification Wang and Zhang (2020), Multi-approach Chen and Li (2006)	Post-hoc	Local
	Stock price prediction: The S&P stocks	Decision Tree Carta et al. (2021)	Post-hoc	Local
	Credit fraud detection	Multi-approach Chaquet-Ulldemolins et al. (2022a, b)	Post-hoc	Global
	Investment Strategy	Multi-approach Wang et al. (2019)	Post-hoc	Local

Table 5 (continued)

XAI method	Aim	AI method	Type	Scope
	Risk-return trade-offs	Multi-approach Anis and Kwon (2021)	Post-hoc	Local
	Stock market prediction	Multi-approach Zhu et al. (2020)	Ad-hoc	Global
	Credit ratings of enterprises	Neural Network Guo et al. (2023)	Post-hoc	Local
	Credit risk	Ensemble model Xu et al. (2021a, b), neural network Kellner et al. (2022); Random Forest Uddin et al. (2022)	Post-hoc	Global
	Detecting online banking fraud	Neural Network Achituve et al. (2019)	Ad-hoc	Global
	Portfolio management	Neural Network Guan and Liu (2021)	Post-hoc	Local
	Stock price forecasting	Neural Network Zhou et al. (2020)	Post-hoc	Local
	Credit scoring	Random Forest Zhang et al. (2020a, b, c)	Ad-hoc	Global
	Bankruptcy prediction	XGBoost Carmona et al. (2022)	Post-hoc	Local/global
Framework	Corporate credit rating prediction	Classification Rodríguez et al. (2022)	Ad-hoc	Local/global
	Credit classification	Fuzzy rule-based Gorzałczany (2016)	Ad-hoc	Global
	Credit risk management	Gradient Boosting Bastos and Matos (2022)	Ad-hoc	Local
	Successful reverse factoring	Gradient boosting Liang et al. (2022)	Post-hoc	Local
	Credit scoring	LightGBM Liu et al. (2022a, b, c), Recursive-Rule eXtraction algorithm (Re-RX) algorithm Hayashi and Oishi (2018)	Post-hoc	Local
	Anomaly detection in financial transactions	Multi-approach Kiefer and Pesch (2021)	Post-hoc	Local
	Churn prediction models in the banking industry	Multi-approach Cao (2021), Random Forest Marín Díaz et al. (2022)	Post-hoc	Local/global
	Credit fraud detection	Multi-approach Chaquet-Ulldemolins et al. (2022a, b)	Post-hoc	Global
	Credit risk	Multi-approach Silva et al. (2019)	Post-hoc	Global
	Credit risk assessment in P2P lending	Multi-approach Amato et al. (2022)	Post-hoc	Local/global
	Credit score prediction in P2P	Multi-approach Moscato et al. (2021)	Post-hoc	Local
	Credit scoring	Multi-approach Bücker et al. (2022)	Post-hoc	Local/global

Table 5 (continued)

XAI method	Aim	AI method	Type	Scope
	Financial decision	Multi-approach La Gatta et al. (2021a, b)	Post-hoc	Local
	Financial distress prediction (FDP)	Multi-approach Zhang et al. (2022a, b)	Post-hoc	Local/global
	Financial transaction data to predict High and Low levels on the traits	Multi-approach Ramon et al. (2021)	Post-hoc	Local/global
	Loan applications	Multi-approach Hadash et al. (2022)	Post-hoc	Local
	Predict the liquidity ratio of mutual funds—portfolio decomposition	Multi-approach Kong et al. (2020)	Ad-hoc	Global
	Risk of fraud in P2P lending	Multi-approach Li et al. (2020a, b)	Post-hoc	Local
	Credit decision-making	Neural network Tyagi (2022)	Post-hoc	Local/global
	Micro-segmentation of customers in the finance sector	Neural network Maree and Omlin (2022)	Post-hoc	Local
	Stock price prediction	Neural network Freeborough and van Zy (2022)	Post-hoc	Local
	Transaction categorization model	Neural network Kotios et al. (2022)	Post-hoc	Local
	Bankruptcy prediction	Optimisation Cho and Shin (2023)	Post-hoc	Local
	Bitcoin prices prediction	Regression Giudici and Raffinetti (2021)	Ad-hoc/post-hoc	Local/global
	Credit risk management	Regression Nagl et al. (2022)	Ad-hoc	Local/global
	Stock performance	Regression Guo et al. (2021)	Ad-hoc	Global
	Credit risk assessment	XGBoost Gramegna and Giudici (2021)	Post-hoc	Local
Knowledge-graph-based	Credit Approval System	Multi-approach Sovrano and Vitali (2022)	Post-hoc	Local
	Fund recommendation	Recommendation system Hsu et al. (2022)	Post-hoc	Local/global
LIME	Bankruptcy prediction	Multi-approach Park et al. (2021)	Post-hoc	Local
	Stock price prediction	Neural network Gite et al. (2021)	Post-hoc	Local
	Credit scoring	Random forest Patron et al. (2020), Walambe et al. (2021)	Post-hoc	Local
	Stock price prediction	Random forest Nicosia et al. (2022)	Ad-hoc	Local
Permutation feature importance (PI)	Stock price prediction	Multi-approach Carta et al. (2022)	Post-hoc	Global

Table 5 (continued)

XAI method	Aim	AI method	Type	Scope
Rule-based methods	Credit investigation	Bayesian network Wu and Han (2021)	Post-hoc	Global
	Credit Card Portfolio Management	Classification Sun et al. (2011)	Post-hoc	Local/global
	Financial decision	Clustering La Gatta et al. (2021a, b)	Post-hoc	Local/global
	Fraud detection	Clustering Irarrázaval et al. (2021)	Ad-hoc	Global
	Credit risk	Decision Tree Xu et al. (2017)	Ad-hoc	Global
	Loan evaluation	Ensemble model Dong et al. (2021)	Ad-hoc	Local/global
	Bank failure prediction	Multi-approach Wang et al. (2016)	Ad-hoc	Global
	Credit rating	Neural network de Campos Souza et al. (2021)	Ad-hoc	Global
	Credit scoring	Neural network Tsakonas and Dounias (2005)	Post-hoc	Local
	Stock price prediction	Optimisation Ghandar and Michalewicz (2011)	Ad-hoc	Global
	Credit assignment	Reinforcement learning Dinu et al. (2022)	Ad-hoc	Global
	Credit risk assessment	Rule system Hayashi (2016)	Post-hoc	Local
	Credit scoring, automation of loan lending process	Rule system Sachan et al. (2020)	Ad-hoc	Global
Sensitivity	Decision explainability	Neural network Zhang et al. (2022a, b)	Ad-hoc	Global
SHAP	Predict subjective financial well-being	Gradient boost Madakkatel et al. (2019)	Post-hoc	Local
	Stock prices prediction	Gradient boosting Ohana et al. (2021)	Post-hoc	Local/global
	Credit approval	Multi-approach Lusinga et al. (2021)	Post-hoc	Local
	Credit default prediction	Multi-approach Alonso Robisco and Carbo Martinez (2022)	Post-hoc	Local
	Credit rating	Multi-approach Kim and Woo (2021)	Post-hoc	Local
	Credit scoring	Multi-approach Hickey et al. (2021)	Post-hoc	Local
	Crude oil price prediction	Multi-approach Gao et al. (2022), Jabeur et al. (2021)	Post-hoc	Local
	Investment recommendations suggestion	Multi-approach Petersone et al. (2022)	Post-hoc	Local
	Stocks: the volatility of healthcare stocks	Multi-approach Weng et al. (2022)	Post-hoc	Local

Table 5 (continued)

XAI method	Aim	AI method	Type	Scope
	Credit risk assessment	Neural Network Cardenas-Ruiz et al. (2022), Gradient boosting de Lange et al. (2022)	Post-hoc	Local/global
	Detecting accounting anomalies in financial statement audits	Neural network Müller et al. (2022)	Ad-hoc	Global
	Financial transaction classification	Neural network Maree et al. (2020)	Post-hoc	Local
	Predict economic growth rates and crises	Neural network Park and Yang (2022)	Post-hoc	Local
	Relationships in capital flows	Neural network Nakamichi et al. (2022)	Post-hoc	Local
	Portfolio construction	Optimisation Papenbrock et al. (2021)	Post-hoc	Local
	Lending to Small and Medium Enterprises (SME)	Random forest Babaei et al. (2023)	Post-hoc	Local
	Robo-advisor, portfolio management; 8 crypto	Random forest Babaei et al. (2022)	Post-hoc	Local
	Credit, peer-to-peer platform	XGBoost Bussmann et al. (2020)	Post-hoc	Local
	Cryptocurrency trading	XGBoost Fior et al. (2022)	Post-hoc	Local/global
	Fake news on crowd-sourced platforms for financial markets	XGBoost Zhang et al. (2020a, b, c)	Post-hoc	Local
	Financial fraud detection	XGBoost Fukas et al. (2022)	Post-hoc	Local
	Loan	XGBoost Stevens et al. (2020)	Post-hoc	Local
	Portfolio construction	XGBoost Jaeger et al. (2021)	Post-hoc	Local/global
Stakeholder-oriented explanations	Risk prediction	Multi-approach Golbin et al. (2019)	Post-hoc	Local/global
Transparent model	Stock price prediction	Bayesian Decision Trees Nuti et al. (2021)	Ad-hoc	Global
	Credit scoring	Classification Kamalloo and Abadeh (2010), Multi approach (Gramespacher and Posth (2021), Obermann and Waack, (2016), Zhang and Dai (2020), Masyutin and Kashnitsky (2017), Pintelas et al. (2020), Regression Dumitrescu et al. (2022)	Ad-hoc	Global
	Investment decision	Clustering Zhang et al. (2020a, b, c)	Ad-hoc	Global
	Loan classification	Decision Tree Zurada (2010)	Ad-hoc	Global

Table 5 (continued)

XAI method	Aim	AI method	Type	Scope
	Option pricing	Decision Tree Ciocan and Mišić (2022)	Ad-hoc	Global
	Credit risk assessment and bankruptcy prediction	Ensemble model Florez-Lopez and Ramon-Jeronimo (2015)	Ad-hoc	Global
	Classification of loan applications	Multi-approach Szwabe and Misiorek (2018)	Ad-hoc	Global
	Financial distress prediction system	Multi-approach Chou (2019)	Ad-hoc	Local/global
	Stock market prediction	Multi-approach Bouktif and Awad (2013)	Ad-hoc	Global
	Credit card default prediction	Neural network De et al. (2020)	Ad-hoc	Local
	Assess cyber risk	Regression Giudici and Raffinetti (2022)	Ad-hoc	Local
	Bank credit risk	Rough set theory Chiua et al. (2010)	Ad-hoc	Global
	Financial statement fraud detection	Rule system Hajek (2019)	Ad-hoc	Global
	Financial distress prediction (FDP)	XGBoost Liu et al. (2022a, b, c)	Ad-hoc	Global
TreeSHAP	Credit scoring	Gradient boosting Liu et al. (2022a, b, c)	Ad-hoc	Global
	Non-life insurance coverage	XGBoost Gramegna and Giudici (2020)	Post-hoc	Local
Visual explanation	Credit scores prediction	Multi-approach Yang et al. (2022a, b)	Ad-hoc	Global
	Banknote recognition and counterfeit detection	Neural network Han and Kim (2019)	Post-hoc	Local
Weight of evidence (WOE)	fraud detection and credit risk assessment	Multi-approach Raymaekers et al. (2022)	Post-hoc	Global

with higher values are considered more critical to the model. Within the domain of ensemble models, including Random Forest, XGboost, and Neural Network algorithms, Feature relevance methods are the most commonly used approach, as shown in Fig. 10.

The Shapley Additive Explanations (SHAP) method is widely used in the field of XAI (Lundberg and Lee 2017) and is also frequently mentioned in this review. The main goal of this approach, based on a game-theoretic concept, is to create a linear model that focusses on the specific case to be explained. The coefficients of this model are then interpreted as the importance of the features. Gradient boosting methods, such as LightGBM and gradient boosting decision trees, have been employed to predict credit default (de Lange et al. 2022), large price drops in the S&P 500 index (Ohana et al. 2021), subjective levels of financial well-being (Madakkatel et al. 2019), and solve practical marketing problems (Nakamichi et al. 2022). The SHAP method has been incorporated into these models to enable the identification of crucial explanatory variables. Credit risk assessment (Cardenas-Ruiz et al. 2022), economic growth rates and crisis prediction (Park and Yang 2022), relationships in capital flows (Nakamichi et al. 2022), and detection of accounting anomalies in financial statement audits (Müller et al. 2022) utilise various types of neural networks, including Convolutional Neural Network (CNN), long short-term memory (LSTM) network-based models, and Autoencoder Neural Networks (AENNs). SHAP are used to improve the accuracy and interpretability of these neural network models. The authors used a combination of Shapley values and Random Forest models to obtain both accurate predictions and reasonable explanations. Shapley values are used with a dynamic Markowitz portfolio optimisation model to improve the trustworthiness of robotic advisors in making investment decisions and to explain how the robotic advisor chooses portfolio weights (Babaei et al. 2022). The Random Forest method for small and medium-sized enterprises (SME) financial performance prediction demonstrated that a reduced number of balance sheet indicators could accurately forecast and clarify SME default and expected return (Babaei et al. 2023). The Shapley value (Shapley (1953)) method was incorporated to increase interpretability, gradually removing variables with the least explanatory power.

Several authors implement the feature importance method to explain the interpretability of their approach, which is based on the same principle as the SHAP approach, by emphasising the characteristics that significantly impact the model outcome. A combination of

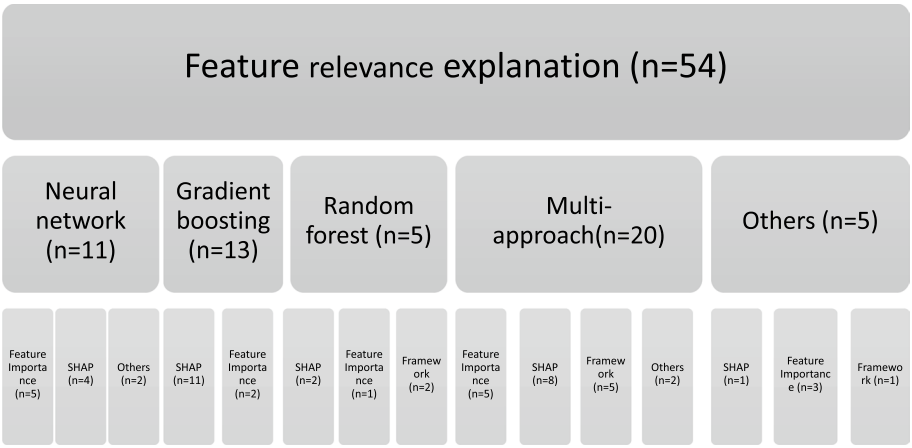

Fig. 10 Feature relevance methods used with artificial intelligence methods and specific XAI methods

linear quantile regression and neural networks was used to predict bank loan losses with a new feature importance measurement method to estimate the strength, direction, interaction and other nonlinear effects of different variables on the model (Kellner et al. 2022). To address the lack of interpretability in enterprise credit rating models, Guo et al. (2023) used Convolutional Neural Networks with attribute and sequence attention modules. Zhou et al. (2020) include a generic time-series predicting framework that uses deep neural networks and a triple attention mechanism to gain satisfactory performance on multi-modality and multi-task learning problems in financial time series analysis. For credit risk assessment, a novel machine learning method, Least Squares Support Feature Machine (LS-SFM), is proposed, which introduces a single-feature kernel and a sampling method to reduce the cost of misclassification and provide interpretable results (Chen and Li 2006). Another research proposes a novel BWSL (Investment Buying Winners and Selling Losers) strategy for stock investing called AlphaStock, which uses enhanced learning and interpretable deep-attention networks to address quantitative trading challenges and achieve an investment strategy that balances risk and reward to reach (Wang et al. 2019). Carta et al. (2021) introduced a graphical user interface that includes generating phrases from global articles to identify high-impact words in the market, creating functions for a decision tree classifier, and predicting stock prices above or below a certain threshold.

In their empirical study, Wang et al. (2024) compared the performance of SHAP-value-based feature selection and importance-based feature selection in the context of fraud detection. They found that traditional feature importance methods typically yield a single score that represents the overall importance of each feature throughout the entire data set. On the other hand, SHAP values provide a more detailed explanation by giving significance values for each feature for individual predictions. This allows for the capture of interaction effects and a better understanding of the model's behaviour.

4.2.2 Explanation by simplification

Simplifying models is commonly achieved through feature extraction and rule-based methods. These techniques are widely regarded as effective ways to provide a set of interpretable features that explain decisions (Speith 2022). In Fig. 11, the XAI methods discussed

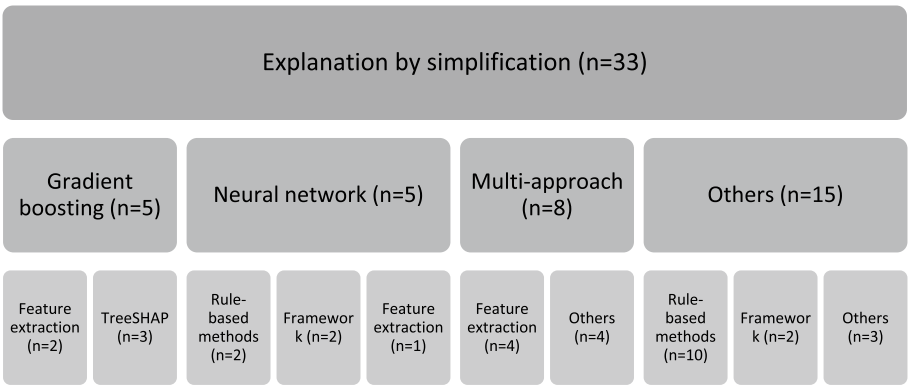

Fig. 11 Explanation by simplification methods used with AI techniques and specific XAI methods

in this article are typically used alongside boosting methods, neural networks, and other approaches that involve multiple techniques.

Various simplification models have been proposed to detect credit card fraud and provide accurate and personalised financial recommendations and risk assessments. One such model is the non-linear model GBDT (Gradient Boosting Decision Tree), which uses cross features and trains a linear regression model for each transaction with MANE (Model-Agnostic Non-linear Explanations) (Tian and Liu 2020). Another model, the Weighted Soft Decision Forest (WSDF), aggregates multiple soft decision trees using a weighting mechanism of each tree's output to simulate the credit scoring process (Zhang et al. 2020a, b, c). Credit risk assessment decision support systems like Continuous Re-RX, Re-RX with J48graft, and Sampling Re-RX use the recursive rule extraction algorithm and decision tree to generate interpretable rules for machine learning-based credit assessment techniques (Hayashi 2016). The TreeSHAP method groups borrowers based on their financial characteristics to measure credit risk related to peer-to-peer lending platforms (Bussmann et al. 2021). Furthermore, a personalised mutual fund recommender system was proposed using a knowledge graph structure and embedding functions that provide both general and complex explanations, which was evaluated using a mutual fund transaction dataset (Hsu et al. 2022). All these models aim to provide accurate, interpretable, and personalised financial recommendations and risk assessments, and they can be applied in Big Data environments.

4.2.3 Local explanation

Local explanations are methods used to explain the decision-making process of machine learning models at the individual instance or statistical unit level by breaking down a complex problem or model into smaller, more manageable, and understandable parts to provide insight and the key factors affecting a specific prediction (Arrieta et al. 2020). These explanations can be generated through various techniques that focus on explaining part of the system function, such as LIME, Counterfactuals, and others (Fig. 12). Using local explanations improves the transparency and interpretability of machine learning models, thereby increasing confidence in their results.

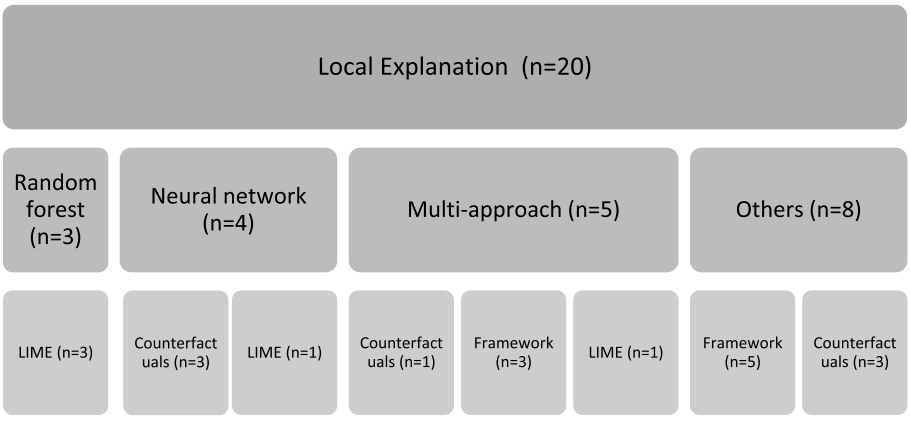

Fig. 12 Local explanation methods used with AI methods and specific XAI methods

A local explanation method that addresses the opacity issue in ML models for credit assessment involves a counterfactual-based interpretability assessment technique. This technique utilises counterfactuals to assess the model's robustness under different scenarios by calculating the decision boundary while accounting for variations in the dataset (Bueff et al. 2022). In the realm of financial news prediction, the LIME method is used to provide investors with clear explanations of how stock prices are predicted using financial news headlines. Future research directions are also identified, such as multilingual predictions, automated predictions from financial news websites, and the integration of emotion-based GIFs (Gite et al. 2021). The emphasis is placed on the importance of feature engineering in finance, and a feature selection approach is proposed to improve predictive performance by identifying relevant features for each stock. Additionally, a genetic algorithm-based approach has been proposed to generate feature-weighted, multiobjective counterfactuals to enhance the interpretability of bankruptcy prediction, which has been experimentally shown to outperform other counterfactual generation methods (Cho and Shin 2023).

4.2.4 The Relationship between XAI, AI methods, and financial problems

We employed heatmaps to visually depict the correlation between XAI, AI approaches, and the specific financial tasks they address. These maps illustrate the values of the primary variables by using a grid of colored squares positioned along each axis. Analysing the hue and position of these squares deduces a relationship between the two variables (Ferreira et al. 2013). Figs. 13 and 14 use heatmaps to show the relationship between two variables: the XAI method used and the AI method; the XAI method used and the goal of the problem. Plotting these variables on the two axes of the heatmap and coloring the squares to

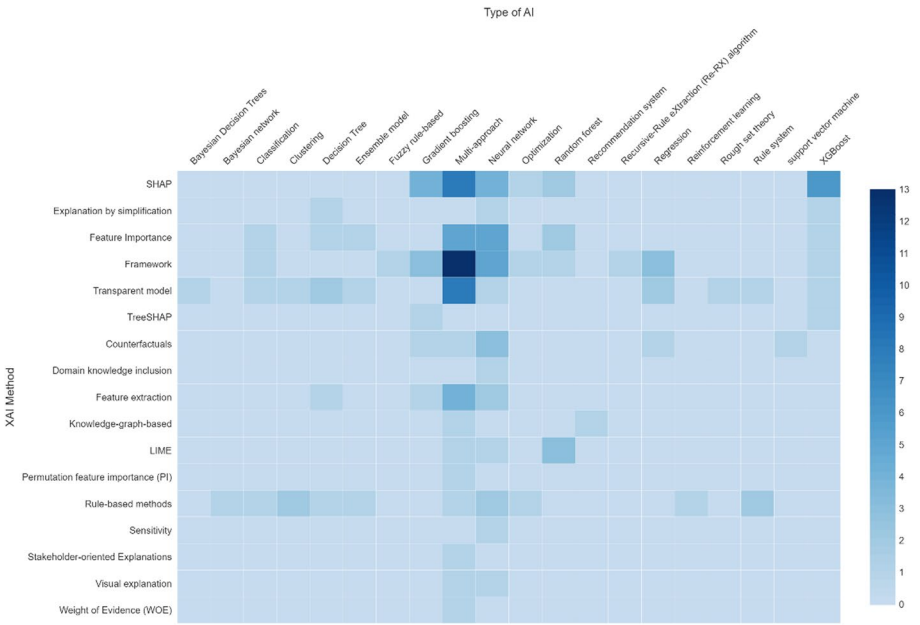

Fig. 13 Relationship between the XAI method and the type of AI

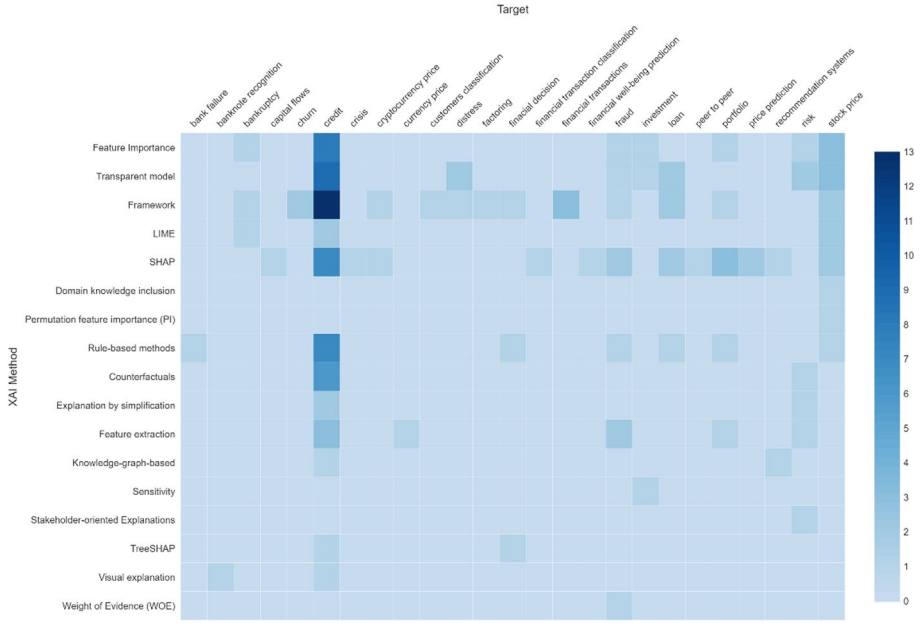

Fig. 14 Relationship between XAI Methods and Problem Types

represent the values of the main variables (in this case, the XAI method) makes it possible to see how the two variables are related and how commonly they were used together.

Using a feature-expression heatmap, a group of converging associations has been successfully identified between XAI, the specific type of AI implemented, and the particular problem type at hand. As stated previously, for assessing credit problems, the most commonly used XAI methods have been feature importance and transparent methods. Additionally, many authors have employed frameworks of XAI methods. It is evident that there is a correlation between frameworks and multi-approach AI techniques. The SHAP technique is employed in conjunction with boosting methods, neural networks, and multi-approach AI techniques.

4.3 RQ3: how are new XAI methods developed and applied?

The development of new XAI methods is often guided by an intuitive understanding of what makes a valuable explanation. Despite ongoing efforts to develop new techniques to extract information from AI systems to provide explanations for their outcomes, the current state of research is still insufficient to fully assess the impact of these explanations on user experience and behaviour regarding their validity and reliability. Although research in finance has seen a significant increase, progress in establishing techniques or metrics to assess the effectiveness of explanation generation methods and the quality of the resulting explanations has been comparatively slow. Only 16 reviewed articles evaluated new XAI techniques or frameworks (Fig. 15).

Several innovative techniques have been introduced in XAI to enhance the interpretability of complex models. One such technique (Han and Kim 2019) represents the first

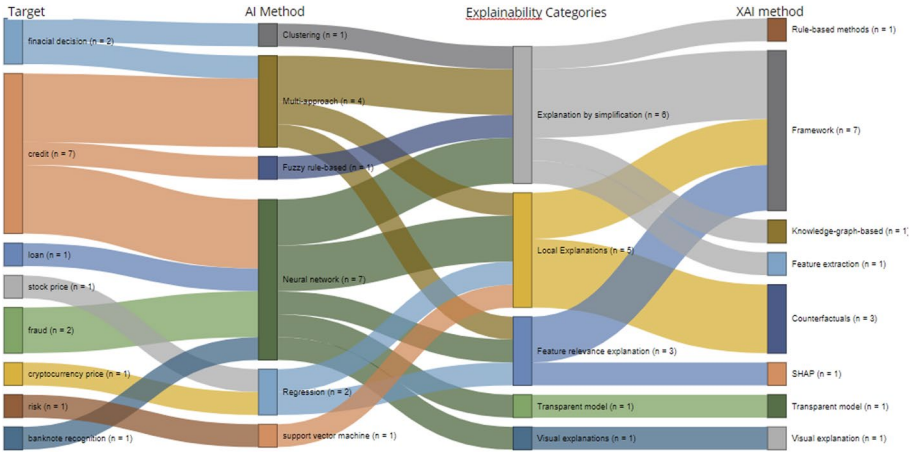

Fig. 15 The selected articles were divided into clusters based on financial task, the AI method, the XAI category and the created new XAI method. The number of articles that fall under each of these categories is given in parentheses.

attempt to ensure the accuracy and interpretability of a banknote detection and counterfeit detection system using XAI. To achieve this goal, the authors proposed a new visualisation approach that addresses the limitations of an existing method. Known as pGrad-CAM, this novel method overcomes the problem of blank activation maps that can occur when using the Grad-CAM technique. Unlike Grad-CAM, which relies on weighted sums of feature maps determined by the average gradient values, pGrad-CAM evaluates positive gradient areas pixel-by-pixel even when the average gradient value is negative, resulting in richer and more comprehensive results in informative activation cards. Another technique, ViCE (Gomez et al. 2020), is a new method that provides counterfactual explanations for model predictions and allows users to understand the decision process by presenting the minimal changes required to modify a decision for each sample through an interactive interface with a use case the effectiveness of the tools on a credit dataset and their potential for future improvement can be customised through a modular black-box design. Meanwhile, the new approach, called Contrafactual Explanations as Interventions in Latent Space (CEILS) (Crupi et al. 2022), has the dual aim of incorporating causality into the creation of counterfactual explanations and using them to make practical suggestions for recourse, while the same time being a method that is convenient can be added to existing counterfactual generator. This is the first step towards giving end users realistic explanations and actionable recommendations on achieving their desired outcome in automated decision-making processes. The RESHAPE (Reconstruction Error SHapley Additive exPlanations Extension) (Müller et al. 2022) creates attribute-level explanations for AENN and a thorough evaluation framework for benchmarking XAI methods in financial audit environments. A new black-box high-fidelity explanatory method called MANE (Model-Agnostic Nonlinear Explanations) (Tian and Liu 2020) explains why a deep learning model classifies a transaction as fraudulent and identifies the key characteristics that contribute to the decision. The approach involves using an aggregation strategy to extract cardholder behaviour patterns from transaction data and Gradient Boosting Decision Tree (GBDT) to discover cross-features that approximate the non-linear local boundary of the deep learning model and improve interpretability. The study shows that the proposed approach outperforms

state-of-the-art techniques on a real financial data set. It introduces a highly accurate explanatory model that characterises behavioural patterns and cross-traits. It allows linear regression models to fit the local nonlinear limit of all deep learning models. Finally, LaGatta et al. (2021) introduced two new XAI methods: CASTLE and PASTLE. The CASTLE approach integrates global knowledge learned by a black box model during its training into local explanations, which is achieved through a clustering algorithm that identifies instances that share similar characteristics and are classified homogeneously by the model. Alternatively, the PASTLE method improves on the standard explanations of the feature importance by providing information about the modifications required in the target instance to increase or decrease the likelihood of the label assigned by the model.

Two studies present new metrics to evaluate and improving XAI systems. Sovrano and Vitali (2022) introduce the DoX Pipeline, a model-agnostic approach to assess the explainability of AI systems by knowledge graph extraction and a new metric based on Achinstein's theory of Explanations. The proposed metric is tested on two real-world Explainable AI-based systems in healthcare and finance using artificial neural networks and TreeSHAP, and the results demonstrate the feasibility of quantifying explainability in natural language information. Gorzaczany and Rudziski (2016) proposed a new approach to the design of fuzzy rule-based classifiers (FRBCs) for financial data classification that addresses interpretability issues by optimising both accuracy and interpretability requirements during the design phase of classifiers. The approach includes an original measure of complexity-dependent interpretability, an efficient implementation of strong fuzzy partitions for attribute domains, a simple and computationally efficient representation of the classifier's rule base, and original genetic operators for processing rule bases. The approach outperformed 24 alternative methods in interpretability while remaining highly competitive in accuracy and computational efficiency.

Several frameworks have been proposed to improve the transparency, auditability, and explainability of complex credit scoring models, including TAX4CS, the three Cs of interpretability, and a lending decision framework. The Transparency, Auditability, and Explainability for Credit Scoring (**TAX4CS**) framework proposes a structured approach to perform explanatory analysis of complex credit scoring models. The framework emphasises the importance of assessing the suitability of the model rather than focussing solely on interpretability (Bücker et al. 2022). A framework called 'the **Three Cs of Interpretability**" (Silva et al. 2019) encompasses the concepts of completeness, correctness and compactness developed in an ensemble model that satisfies the requirements of correctness and diversity and generates comprehensive and concise explanations for predictive models. Evaluation of the model in three datasets, including financial and biomedical, has demonstrated its superior predictive performance compared to individual models such as Deep Neural Network, Scorecard, and Random Forest while providing accurate, compact, and diverse explanations consistent with expert analysis. A lending decision framework (Chen et al. 2022) includes an interpretable machine learning model, an interactive visualisation tool, and various types of summaries and explanations for each decision. The proposed lending decision framework (Chen et al. 2022) includes an interpretable machine learning model, an interactive visualisation tool, and various types of summaries and explanations for each decision. It is a two-layer additive risk model that can be decomposed into meaningful subscales, and the online visualisation tool allows the model to be explored to understand how it arrived at its decision. The framework also offers three explanations: case-based reasoning, essential features, and custom sparse explanation. A new general-purpose framework, the Counterfactual Conditional Heterogeneous Autoencoder (C-CHVAE), also allows the generating of counterfactual feature sets without requiring a specific distance or

cost function for tabular data. The framework can be applied to various autoencoder architectures that can model heterogeneous data and estimate conditional log-likelihoods of the changing and unchanging inputs. In addition, C-CHVAE does not require a distance function for the input space and is not restricted to specific classifiers (Pawelczyk et al. 2020).

4.4 RQ4: What are the challenges, requirements, and unresolved issues in the literature related to the application of XAI in finance?

Based on the literature review, we present the challenges that need to be resolved in the field of explainability of AI models. We have identified three main challenges in conducting this comprehensive review: difficulties arise when determining the intended audience for the explanation, such as whether it is targeted towards technical specialists or end users, the lack of new XAI techniques or explainable machine learning methods were used to tackle the financial problem, the absence of explainability valuation metrics that play a significant role in building trust in AI models in finance, and the security of information provided to the XAI model.

4.4.1 The audience for the method of explanation?

The rationale for employing XAI is that the outcomes generated by AI models are more likely to be perceived as credible by end users when accompanied by an accessible explanation in nontechnical language (Hagras 2018).

Some authors specify the intended audience for the method of explanation, such as the RESHAPE method (Müller et al. 2022), proposed as a tool to facilitate the application of complex neural networks in financial audits by presenting auditors with detailed but understandable explanations, thereby increasing their confidence in opaque models, indicating that the degree of interpretability may only be understandable to experts in the field.

Mohammadi et al. (2021) found that by handling different types of constraints, black-box encoding can enable stakeholders and explanation providers to consider individual-specific situations where a person may request that their personal limitations be considered in the provided explanation. Access to reliable and efficient tools to explain algorithmic decisions becomes crucial as stakeholders adopt more complex neural models to make decisions.

4.4.2 New XAI technique or new explanation for machine learning method?

It was difficult to discern whether some authors (Obermann and Waack 2016; Cao 2021; Chaquet-Ulldemolins et al. 2022a, b; Liu et al. 2021a, b; Zhou et al. 2020; Dong et al. 2021) had introduced a new explainable AI method or whether they were instead presenting an explainable machine learning method that aimed to improve the ML method itself rather than a new XAI method to develop.

As an example of a non-new XAI method, we can present the proposed interpretable personal credit evaluation model DTONN that combines the interpretability of decision trees with the high prediction accuracy of neural networks (Chen et al. 2020). This example shows that the researchers sought interpretability using various methods rather than inventing new XAI methods.

4.4.3 Explainability evaluation metrics?

One of the limitations of XAI techniques is the insufficient availability of evaluation metrics that determine the quality of the explanations provided. The process of choosing the most appropriate explanation method can be complex one, as there is no single, universally accepted definition of what constitutes a satisfactory explanation. This can vary depending on the individual and the specific circumstances involved. Further investigation is required to establish suitable standards for selecting and evaluating explanatory techniques in various contexts.

Some researchers use central tendency measures of mode and median to assess the effectiveness of an explanation method and the most recent XAI technique. According to the results, CASTLE outperformed Anchors by 6% in terms of accuracy, indicating that explanations can help to understand black box models (La Gatta et al. 2021a, b). The accuracy of MANE, the XAI method proposed for credit card fraud detection, was assessed using the RMSE local approximation accuracy metric and PCR (positive classification rate) to validate the accuracy of the selected features by the explanatory model (Tian and Liu 2020). To assess the efficacy of integrating knowledge of the financial domain into the design and training of neural networks, Zheng et al. (2021) used the mean average percentage error in both the training and test set as a metric for performance evaluation. It can be inferred that the evaluation of the machine learning method is prioritised over the effectiveness of the XAI method in most cases.

4.4.4 Information security

The security of the XAI model is paramount, as any breach of confidentiality could expose it to malicious exploitation. This could lead to attackers deceiving the model and risking the integrity of its results. The model can be manipulated to produce a different output by injecting specific information by attackers (Ghorbani et al. 2019). This risk associated with XAI is a critical consideration, as it enables even minor modifications to the original AI model, which can significantly alter the outcome. Such attacks on loan or credit management systems can be catastrophic and have significant economic consequences.

4.4.5 Impact of XAI on the finance area

Transparency and regulatory compliance are paramount in the finance industry (Zheng et al. 2019). In this regard, the use of XAI can greatly benefit financial institutions. It enables them to achieve both objectives with the utmost efficiency and accuracy. XAI offers superior transparency and traceability compared to traditional AI, making it a more suitable choice for financial institutions aiming to comply with regulations and establish trust with their stakeholders.

The evaluated research has yielded empirical evidence highlighting XAI's potential for improving transparency and predictive performance. Ghandar and Michalewicz (2011) investigated whether model interpretability could benefit financially intelligent systems by realising valuable properties. The experimental results demonstrated that interpretability objectives could yield rules with superior prediction capability while avoiding excessive complexity, thus underscoring the significance of interpretability in financial modelling.

XAI is crucial for financial professionals to understand the variables that affect risk predictions. This, in turn, leads to more accurate risk assessments and builds confidence

among decision makers in the information provided by AI models. Lange et al. (2022) created an XAI model to predict credit defaults, utilising a unique data set of unsecured consumer loans. The authors combined SHAP and LightGBM to interpret how explanatory variables affect predictions by implementing XAI methods in banking to improve the interpretability and reliability of AI models. These XAI-implementing models usually outperform state-of-the-art models (Bussmann et al. 2020).

Ensuring the trustworthiness of recommendations in financial XAI is a critical challenge that demands our attention. Misleading explanations can cause users to rely on inaccurate results, eroding their confidence and trust. Legal regulations also drive the requirement to validate machine learning outcomes. The European Union has achieved a significant breakthrough in regulating AI technology by introducing the AI Act. This pioneering legal framework is designed to encourage the responsible and ethical use of AI in diverse industries and applications. The enactment of this act represents a significant step forward in ensuring that AI is used in a safe and beneficial manner.

5 Discussion and areas for further investigation

The main goal of this review was twofold. First, to question and organise the tasks encompassed by XAI in finance and highlight the challenges in generating explainable models. Second, to suggest potential avenues for future studies on implementing XAI in finance. To develop XAI in the field of finance, some challenges and future work for further research are presented:

(1) To fully exploit the potential of the XAI, it is essential to focus on explaining the decisions made by the XAI algorithm. Providing a clear and concise explanation of each variable is a critical aspect. When using numeric coding, it is crucial to indicate the meaning of each variable. It is highly recommended to avoid confusion instead of abbreviating it. For instance, the variable "RM" can be written as a "risk measure". It is worth noting that variables can sometimes be presented in numeric coding without any legend or abbreviation, which can lead to confusion and misinterpretation.

(2) Research shows that explainable AI is vital for credit assessment and risk management. However, other sensitive areas, such as portfolio optimisation, internet financing, and fraud detection, are comparatively less explored than credit scoring and risk management. Therefore, further research using XAI methods in these less studied areas is crucial.

(3) There is currently a lack of definitive guidance regarding selecting an appropriate XAI method for a given scenario, considering the multitude of options available. Furthermore, a degree of ambiguity surrounds the formulation of suitable explanations and how contextual factors may influence them (Ben David et al. 2021). Researchers frequently use multiple XAI methods with multiple datasets to assess a model's performance or demonstrate its approximation capabilities. Using diverse datasets can effectively demonstrate the model's robustness and flexibility in handling various data distributions (Pawelczyk et al. 2020). In a recent study by Tian and Liu (2020), an assessment of the effectiveness of AI techniques in finance was proposed to evaluate the prediction of Chinese companies' listing statuses. The study examined decision cost and classification accuracy as critical metrics for assessing the model's performance.

This new perspective can be applied to future XAI in finance to evaluate the effectiveness of machine learning techniques.

(4) To grasp the application of XAI in finance, it is beneficial to examine the frameworks used by researchers who implemented XAI techniques to improve machine learning explainability is beneficial. The integration of LIME and SHAP techniques can substantially enhance the interpretability of XAI by effectively elucidating both the global and local contexts of the model. Further investigation is needed to determine the most commonly used combined XAI methods and their specific purposes.

(5) Feature relevance explanations are commonly used XAI techniques in finance. However, it is possible that once we identify "significant" or "relevant" characteristics, we could encounter a situation where we have high-dimensional feature vectors that require additional analysis. This analysis involves examining correlations or employing various metrics to determine similarities or closeness between the data points. Although the interpretation may simply act as a mathematical depiction, using these methods to restructure a black-box model internally is feasible.

(6) Decision Trees are a commonly utilised intrinsic tool commonly used in financial XAI. Decision trees are a highly adaptable tool that can act both as an intrinsic method and a post-hoc method for approximating more complicated models. They possess the ability to function as a model and an explanation in and of themselves, as well as approximate complex models such as neural networks or gradient-boosting machines. We recommend avoiding complex models with unclear operations and no decision rule explanations. Although decision trees with multiple levels can theoretically be comprehensible, their extended length of decision rules along a path makes them not easily understandable. The explainability task cannot be realised due to model complexity. We propose to expand the meaning of the term XAI, not only to indicate that the model is inherently explainable, but also to explain the model's behaviour clearly and concisely—either through diagrams or visuals—so that anyone, regardless of their level of knowledge, can easily and quickly understand the behaviour of the model.

Close collaboration between financial professionals and AI developers is crucial for deploying a financial AI model to improve the financial decision process. Financial experts possess the necessary expertise to determine whether a model's behaviour conforms to appropriate standards. Calculating bias and fairness metrics is imperative to ensure that the models employed do not display discriminatory behaviour. This is vital to maintaining ethical and moral standards in utilising such models and must be considered for the sake of all concerned parties. It is of utmost importance to adhere to explanatory methods that can be demonstrated faithfully and reliably to the model. Use of any other techniques must be avoided. The factors mentioned above make it necessary to use XAI methods while implementing financial AI. Increasing the use of explainable techniques in financial AI research would significantly enhance its financial relevance.

6 Recent developments

We conducted a final search in May 2024 to ensure the completeness of this review and identify any recently published studies. This search led to new studies on XAI measurement metrics by Giudici and Raffinetti (2023) and the Key Risk Indicators for AI Risk Management (KAIRI) framework by Giudici et al. (2024). It also found a new Python

package called Aletheia that was made to make it easier to figure out what deep ReLU networks are (Sudjianto et al. 2020) and PiML, a Python toolbox that was made just for interpretable machine learning (Sudjianto et al. 2023).

As stated in Sect. 5.4.3, one drawback of XAI approaches is the lack of measurement criteria that accurately assess the quality of the explanations given. To ensure AI is trustworthy and to properly control its use in the financial sector, Giudici and Raffinetti (2023) have increased the use of Lorenz Zonoids to create measuring tools for the four main trustworthiness criteria: S.A.F.E. (Sustainability, Accuracy, Fairness, and Explainability). They used the algorithm to predict bitcoin prices, showcasing its superiority over state-of-the-art techniques. Giudici et al. (2024) have developed the Key Risk Indicators for AI Risk Management (KAIRI) framework. This framework consists of statistical tools that can analyse the prediction output of any machine learning model, regardless of the data structure and model used. KAIRI can be used to measure the level of safety in any AI application. Sudjianto et al. (2020) introduce Aletheia, an innovative toolkit that incorporates practical case studies of credit risk to enhance the understanding of decision-making processes in deep ReLU networks through network analysis. This allows researchers to have a direct understanding of the logic of the network, detect any issues, and optimise it. In an additional academic publication, Sudjianto et al. (2023) provide PiML, an original Python toolkit that not only provides an expanding selection of interpretable models but also includes an improved collection of diagnostic tests. The widespread adoption of PiML by several financial organisations since its introduction in May 2022, highlights its significant potential in the banking industry.

Advancements in XAI include the use of measurement tools such as S.A.F.E. and KAIRI, as well as interpretable models such as Aletheia and PiML. These advances aim to enhance the reliability and risk control of AI systems, with a specific emphasis on their practical use in the financial industry.

7 Conclusions

In finance, it is imperative to carefully evaluate and consider the potential limitations and drawbacks of AI technologies before making any decisions. It is recommended to integrate the XAI methodology while using these technologies to achieve optimal results. This methodology is a highly effective tool for achieving financial success by instilling confidence in the decisions made with AI technology. It provides benefits that extend beyond a mere understanding of the implemented decisions. By implementing this methodology, financial institutions can effectively validate the outcomes of their decisions, thereby ensuring their credibility and trustworthiness.

This review aims to provide a comprehensive analysis of the XAI methods used in the financial industry, focussing on evaluating their efficacy in diverse domains and tasks. In the financial sector, it is imperative to exercise caution when relying solely on sophisticated AI methodologies, as they can lead to unforeseen biases and uncommon problems. The study revealed that traditional machine learning methods are used more frequently in finance than complex multilevel AI algorithms. Additionally, the analysis presents insightful trends in the field of XAI research, highlighting the prevalence of LIME and SHAP as the most commonly adopted techniques.

The review emphasises the challenges and unresolved issues that arise when implementing XAI techniques in finance. XAI has the potential to facilitate the practical application of AI techniques. However, it is important to consider the obstacles that may hinder its implementation. These include the challenge of identifying the target audience, the lack of new XAI techniques, the absence of evaluation metrics for explainability, and the security of the data used in the XAI models. Addressing these challenges will be crucial to ensuring the successful integration of XAI in real-world scenarios. Furthermore, the review presents a comprehensive overview of the potential applications and future research directions of XAI in the field of finance. Some of these are portfolio optimisation, internet financing, and fraud detection.

Moreover, the collaboration between financial professionals and AI developers can improve XAI to be more specific to finance by enhancing the integration of not only model-agnostic techniques but also including more particular finance methods.

This review highlights empirical examples that demonstrate the potential benefits of XAI in the financial industry. It provides valuable insights into the potential of XAI in finance and contributes to the ongoing discussion on the use of XAI in this industry. The adoption of AI algorithms in the financial sector has been on the rise and it is expected that the use of XAI is expected to have a significant impact on their decision-making process in the future. However, the challenges mentioned above must be addressed, and this presents an opportunity for researchers to work in the field of financial XAI.

Author contributions Both authors contributed to the article and approved the submitted version. Material preparation, data collection and analysis were performed by J.Č and A.K.. A.K. critically revised the manuscript. Both authors read and approved the final manuscript.

Funding This work was partially supported by the COST ACTION CA19130, Fintech and Artificial Intelligence in Finance—Towards a transparent financial industry (FinAI).

Declarations

Competing interest The authors declare no competing interests.

References

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

Alonso Robisco A, Carbo Martinez JM (2022) Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. Financ Innov 8(1):70

Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM (2021) Explainable artificial intelligence: an analytical review. Wiley Interdisc Rev 11(5):e1424

Anis HT, Kwon RH (2021) A sparse regression and neural network approach for financial factor modelling. Appl Soft Comput 113:107983

Aria M, Cuccurullo C (2017) bibliometrix: An R-tool for comprehensive science mapping analysis. J Informet 11(4):959–975
Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Herrera F (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Inform Fusion 58:82–115
Babaei G, Giudici P, Raffinetti E (2022) Explainable artificial intelligence for crypto asset allocation. Financ Res Lett 47:102941
Babaei G, Giudici P, Raffinetti E (2023) Explainable fintech lending. J Econ Bus 125:106126
Balcaen S, Ooghe H (2006) 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. Br Account Rev 38(1):63–93
Bastos JA, Matos SM (2022) Explainable models of credit losses. Eur J Oper Res 301(1):386–394
Belle V, Papantonis I (2021) Principles and practice of explainable machine learning. Front Big Data 39:78
Bücker M, Szepannek G, Gosiewska A, Biecek P (2022) Transparency, auditability, and explainability of machine learning models in credit scoring. J Oper Res Soc 73(1):70–90
Bueff AC, Cytryński M, Calabrese R, Jones M, Roberts J, Moore J, Brown I (2022) Machine learning interpretability for a stress scenario generation in credit scoring based on counterfactuals. Expert Syst Appl 202:117271
Bussmann N, Giudici P, Marinelli D, Papenbrock J (2020) Explainable AI in fintech risk management. Front Artif Intell 3:26
Bussmann N, Giudici P, Marinelli D, Papenbrock J (2021) Explainable machine learning in credit risk management. Comput Econ 57:203–216
Cao L, Yang Q, Yu PS (2021) Data science and AI in FinTech: an overview. Int J Data Sci Anal 12:81–99. https://doi.org/10.1007/s41060-021-00278-w
Carmona P, Dwekat A, Mardawi Z (2022) No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure. Res Int Bus Financ 61:101649
Carta SM, Consoli S, Piras L, Podda AS, Recupero DR (2021) Explainable machine learning exploiting news and domain-specific lexicon for stock market forecasting. IEEE Access 9:30193–30205
Carta S, Consoli S, Podda AS, Recupero DR, Stanciu MM (2022) Statistical arbitrage powered by explainable artificial intelligence. Expert Syst Appl 206:117763
Chaquet-Ulldemolins J, Gimeno-Blanes FJ, Moral-Rubio S, Muñoz-Romero S, Rojo-Álvarez JL (2022a) On the black-box challenge for fraud detection using machine learning (I): linear models and informative feature selection. Appl Sci 12(7):3328
Chaquet-Ulldemolins J, Gimeno-Blanes FJ, Moral-Rubio S, Muñoz-Romero S, Rojo-Álvarez JL (2022b) On the black-box challenge for fraud detection using machine learning (II): nonlinear analysis through interpretable autoencoders. Appl Sci 12(8):3856
Chen H (2020) An interpretable comprehensive capital analysis and review (CCAR) neural network model for portfolio loss forecasting and stress testing. J Credit Risk 17(3):89
Chen C, Lin K, Rudin C, Shaposhnik Y, Wang S, Wang T (2022) A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations. Decis Support Syst 152:113647
Cho SH, Shin KS (2023) Feature-weighted counterfactual-based explanation for bankruptcy prediction. Expert Syst Appl 216:119390
Ciocan DF, Mišić VV (2022) Interpretable optimal stopping. Manage Sci 68(3):1616–1638
Clement T, Kemmerzell N, Abdelaal M, Amberg M (2023) XAIR: a systematic metareview of explainable AI (XAI) aligned to the software development process. Mach Learn Knowl Extract 5(1):78–108
Crupi R, Castelnovo A, Regoli D, Gonzalez SM, B. (2022) Counterfactual explanations as interventions in latent space. Data Min Knowl Discov 5:1–37
Dastile X, Celik T, Vandierendonck H (2022) Model-agnostic counterfactual explanations in credit scoring. IEEE Access 10:69543–69554
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 Comput Sci 168:40–48
de Campos Souza PV, Lughofer E, Guimaraes AJ (2021) An interpretable evolving fuzzy neural network based on self-organized direction-aware data partitioning and fuzzy logic neurons. Appl Soft Comput 112:107829
de Lange PE, Melsom B, Vennerød CB, Westgaard S (2022) Explainable AI for credit assessment in banks. J Risk Financ Manag 15(12):556
Diamant E (2017) Advances in artificial intelligence: are you sure, we are on the right track? Trans Netw Commun 5(4):23
Dimitras AI, Zanakis SH, Zopounidis C (1996) A survey of business failures with an emphasis on prediction methods and industrial applications. Eur J Oper Res 90(3):487–513

Dong LA, Ye X, Yang G (2021) Two-stage rule extraction method based on tree ensemble model for interpretable loan evaluation. Inf Sci 573:46–64
Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM (2021) How to conduct a bibliometric analysis: an overview and guidelines. J Bus Res 133:285–296
Dumitrescu E, Hué S, Hurlin C, Tokpavi S (2022) Machine learning for credit scoring: improving logistic regression with non-linear decision-tree effects. Eur J Oper Res 297(3):1178–1192
Ferreira N, Poco J, Vo HT, Freire J, Silva CT (2013) Visual exploration of big spatio-temporal urban data: a study of new york city taxi trips. IEEE Trans Visual Comput Graphics 19(12):2149–2158
Fior J, Cagliero L, Garza P (2022) Leveraging explainable AI to support cryptocurrency investors. Future Internet 14(9):251
Florez-Lopez R, Ramon-Jeronimo JM (2015) Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment: a correlated-adjusted decision forest proposal. Expert Syst Appl 42(13):5737–5753
Frasca M, La Torre D, Pravettoni G, Cutica I (2024) Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. Discov Artif Intell 4(1):15
Freeborough W, van Zyl T (2022) Investigating explainability methods in recurrent neural network architectures for financial time series data. Appl Sci 12(3):1427
Gao X, Wang J, Yang L (2022) An explainable machine learning framework for forecasting crude oil price during the COVID-19 pandemic. Axioms 11(8):374
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 Comput Sci 7:e340
Giudici P, Raffinetti E (2021) Shapley-Lorenz eXplainable artificial intelligence. Expert Syst Appl 167:114104
Giudici P, Raffinetti E (2022) Explainable AI methods in cyber risk management. Qual Reliab Eng Int 38(3):1318–1326
Giudici P, Raffinetti E (2023) SAFE artificial intelligence in finance. Financ Res Lett 56:104088
Giudici P, Centurelli M, Turchetta S (2024) Artificial Intelligence risk measurement. Expert Syst Appl 235:121220
Gorzałczany R, Gorzałczany MB, GRudzinshi R (2016) A multi-objective genetic optimization for fast, fuzzy rule-Based credit classification with balanced accuracy and interpretability. Appl Soft Comput 40:206–220
Gramegna A, Giudici P (2020) Why to buy insurance? An explainable artificial intelligence approach. Risks 8(4):137
Gramegna A, Giudici P (2021) SHAP and LIME: an evaluation of discriminative power in credit risk. Front Artif Intell 4:752558
Gramespacher T, Posth JA (2021) Employing explainable AI to optimize the return target function of a loan portfolio. Front Artif Intell 4:693022
Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM Comput Surv (CSUR) 51(5):1–42
Guo M, Xu Z, Zhang Q, Liao X, Liu J (2021) Deciphering feature effects on decision-making in ordinal regression problems: an explainable ordinal factorization model. ACM Trans Knowl Discov Data (TKDD) 16(3):1–26
Guo W, Yang Z, Wu S, Wang X, Chen F (2023) Explainable enterprise credit rating using deep feature crossing. Expert Syst Appl 220:119704
Hagras H (2018) Toward human-understandable, explainable AI. Computer 51(9):28–36
Han M, Kim J (2019) Joint banknote recognition and counterfeit detection using explainable artificial intelligence. Sensors 19(16):3607
Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, Hussain A (2024) Interpreting black-box models: a review on explainable artificial intelligence. Cognit Comput 16(1):45–74
Hayashi Y (2016) Application of a rule extraction algorithm family based on the Re-RX algorithm to financial credit risk assessment from a Pareto optimal perspective. Operat Res Perspect 3:32–42
Hayashi Y, Oishi T (2018) High accuracy-priority rule extraction for reconciling accuracy and interpretability in credit scoring. N Gener Comput 36:393–418
Hsu PY, Chen CT, Chou C, Huang SH (2022) Explainable mutual fund recommendation system developed based on knowledge graph embeddings. Appl Intell 4:1–26
Hutson M (2020) Core progress in AI has stalled in some fields. Science. https://doi.org/10.1126/SCIENCE.368.6494.927
Irarrázaval ME, Maldonado S, Pérez J, Vairetti C (2021) Telecom traffic pumping analytics via explainable data science. Decis Support Syst 150:113559

Jabeur SB, Khalfaoui R, Arfi WB (2021) The effect of green energy, global environmental indexes, and stock markets in predicting oil price crashes: Evidence from explainable machine learning. J Environ Manage 298:113511
Kellner R, Nagl M, Rösch D (2022) Opening the black box–Quantile neural networks for loss given default prediction. J Bank Finance 134:106334
Kinger S, Kulkarni V (2024) Demystifying the black box: an overview of explainability methods in machine learning. Int J Comput Appl 46(2):90–100
Kotios D, Makridis G, Fatouros G, Kyriazis D (2022) Deep learning enhancing banking services: a hybrid transaction classification and cash flow prediction approach. J. Big Data 9(1):100
La Gatta V, Moscato V, Postiglione M, Sperli G (2021a) CASTLE: Cluster-aided space transformation for local explanations. Expert Syst Appl 179:115045
La Gatta V, Moscato V, Postiglione M, Sperlì G (2021b) PASTLE: Pivot-aided space transformation for local explanations. Pattern Recogn Lett 149:67–74
Levy Y, Ellis TJ (2006) A systems approach to conduct an effective literature review in support of information systems research. Inform Sci 9:89
Li Z, Chi G, Zhou Y, Liu W (2020b) Research on listed companies' credit ratings, considering classification performance and interpretability. Journal of Risk Model Validation 15(1):8
Liang Z, Xie T, Yi Z (2022) Core-firm financial structure on reverse factoring with machine learning models. Appl Math Model Comput Simul. https://doi.org/10.3233/ATDE221065
Linardatos P, Papastefanopoulos V, Kotsiantis S (2020) Explainable ai: a review of machine learning interpretability methods. Entropy 23(1):18
Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57
Liu W, Fan H, Xia M (2021b) Step-wise multi-grained augmented gradient boosting decision trees for credit scoring. Eng Appl Artif Intell 97:104036
Liu W, Fan H, Xia M (2022a) Credit scoring based on tree-enhanced gradient boosting decision trees. Expert Syst Appl 189:116034
Liu W, Fan H, Xia M, Pang C (2022b) Predicting and interpreting financial distress using a weighted boosted tree-based tree. Eng Appl Artif Intell 116:105466
Liu W, Fan H, Xia M, Xia M (2022c) A focal-aware cost-sensitive boosted tree for imbalanced credit scoring. Expert Syst Appl 208:118158
Liubchenko VV (2022) Some aspects of software engineering for AI-based systems. Probl Program 3–4:99–106
Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Adv Neural Inform Process Syst 30:78
Makridakis S (2017) The forthcoming Artificial Intelligence (AI) revolution: its impact on society and firms. Futures 90:46–60
Marín Díaz G, Galán JJ, Carrasco RA (2022) XAI for churn prediction in B2B models: a use case in an enterprise software company. Mathematics 10(20):3896
Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38
Moscato V, Picariello A, Sperlí G (2021) A benchmark of machine learning approaches for credit score prediction. Expert Syst Appl 165:113986
Nagl M, Nagl M, Rösch D (2022) Quantifying uncertainty of machine learning methods for loss given default. Front Appl Math Stat 120:78
Nakamichi T, Yoshida R, Tanaka R, Suzuki T (2022) Visualization of nonlinear relationship in capital flows of Japanese mutual funds. Nonlinear Theory Its Appl IEICE 13(2):221–226
Nuti G, Jiménez Rugama LA, Cross AI (2021) An explainable Bayesian decision tree algorithm. Front Appl Math Stat 3: 28-96
Obermann L, Waack S (2016) Interpretable multiclass models for corporate credit rating capable of expressing doubt. Front Appl Math Stat 2:16
Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 88:105906
Park S, Yang JS (2022) Interpretable deep learning LSTM model for intelligent economic decision-making. Knowl-Based Syst 248:108907
Park MS, Son H, Hyun C, Hwang HJ (2021) Explainability of machine learning models for bankruptcy prediction. IEEE Access 9:124887–124899
Pintelas E, Livieris IE, Pintelas P (2020) A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. Algorithms 13(1):17
Pranckutė R (2021) Web of Science (WoS) and Scopus: the titans of bibliographic information in today's academic world. Publications 9(1):12

Ramon Y, Farrokhnia RA, Matz SC, Martens D (2021) Explainable AI for psychological profiling from behavioral data: an application to big five personality predictions from financial transaction records. Information 12(12):518
Raymaekers J, Verbeke W, Verdonck T (2022) Weight-of-evidence through shrinkage and spline binning for interpretable nonlinear classification. Appl Soft Comput 115:108160
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
Sachan S, Yang JB, Xu DL, Benavides DE, Li Y (2020) An explainable AI decision-support-system to automate loan underwriting. Expert Syst Appl 144:113100
Saeed W, Omlin C (2023) Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowl-Based Syst 263:110273
Saranya A, Subhashini R (2023) A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decis Anal J 5:100230
Shapley LS (1953) A value for n-person games. In: Kuhn H, Tucker A (eds) Contributions to the Theory of Games II. Princeton University Press, Princeton, pp 307–317
Singh VK, Singh P, Karmakar M, Leta J, Mayr P (2021) The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. Scientometrics 126:5113–5142
Sun D, Liu L, Zhang P, Zhu X, Shi Y (2011) Decision rule extraction for regularized multiple criteria linear programming model. Int J Data Warehous Min (IJDWM) 7(3):88–101
Szepannek G, Lübke K (2021) Facing the challenges of developing fair risk scoring models. Front Artif Intell 4:681915
Uddin MS, Chi G, Al Janabi MA, Habib T (2022) Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability. Int J Financ Econ 27(3):3713–3729
Verikas A, Kalsyte Z, Bacauskiene M, Gelzinis A (2010) Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. Soft Comput 14:995–1010
Vilone G, Longo L (2021) Classification of explainable artificial intelligence methods through their output formats. Mach Learn Knowl Ext 3(3):615–661
Von Eschenbach WJ (2021) Transparency and the black box problem: why we do not trust AI. Philos Technol 34(4):1607–1622
Wang Y, Zhang Y (2020) Credit risk assessment for small and microsized enterprises using kernel feature selection-based multiple criteria linear optimization classifier: evidence from China. Complexity 2020:1–16
Wang D, Quek C, Ng GS (2016) Bank failure prediction using an accurate and interpretable neural fuzzy inference system. AI Commun 29(4):477–495
Wang H, Liang Q, Hancock JT, Khoshgoftaar TM (2024) Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. J Big Data 11(1):44
Weng F, Zhu J, Yang C, Gao W, Zhang H (2022) Analysis of financial pressure impacts on the health care industry with an explainable machine learning method: China versus the USA. Expert Syst Appl 210:118482
Wu HD, Han L (2021) A novel reasoning model for credit investigation system based on Fuzzy Bayesian Network. Procedia Computer Science 183:281–287
Yang G, Ye Q, Xia J (2022a) Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Inform Fusion 77:29–52
Yang K, Yuan H, Lau RY (2022b) PsyCredit: An interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing. Expert Syst Appl 198:116847
Yao K, Zheng Y (2023) Fundamentals of machine learning. Nanophotonics and machine learning: concepts, fundamentals, and applications. Springer International Publishing, Cham, pp 77–112
Zhang Z, Dai Y (2020) Combination classification method for customer relationship management. Asia Pac J Mark Logist 32(5):1004–1022
Zhang M, Sun J, Wang J (2022a) Which neural network makes more explainable decisions? An approach towards measuring explainability. Autom Softw Eng 29(2):39
Zhang Z, Wu C, Qu S, Chen X (2022b) An explainable artificial intelligence approach for financial distress prediction. Inf Process Manage 59(4):102988
Zheng XL, Zhu MY, Li QB, Chen CC, Tan YC (2019) FinBrain: when finance meets AI 2.0. Front Inform Technol Electron Eng 20(7):914–924
Zhou L, Si YW, Fujita H (2017) Predicting the listing statuses of Chinese-listed companies using decision trees combined with an improved filter feature selection method. Knowl-Based Syst 128:93–101
Zhou J, Gandomi AH, Chen F, Holzinger A (2021) Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics 10(5):593

Zhu J, Liu W (2020) A tale of two databases: The use of Web of Science and Scopus in academic papers. Scientometrics 123(1):321–335

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 Things J 10(4):2967–2978

Achituve I, Kraus S, Goldberger J (2019) Interpretable online banking fraud detection based on a hierarchical attention mechanism. In: 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1–6). IEEE

Amato F, Ferraro A, Galli A, Moscato F, Moscato V, Sperlí G (2022) Credit score prediction relying on machine learning.

Ben David D, Resheff YS, Tron T (2021) Explainable AI and adoption of financial algorithmic advisors: an experimental study. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp 390–400)

Biran O, Cotton C (2017) Explanation and justification in machine learning: a survey. In: IJCAI-17 workshop on explainable AI (XAI) (Vol. 8, No. 1, pp. 8–13).

Bouktif S, Awad MA (2013) Ant colony based approach to predict stock market movement from mood collected on Twitter. In: Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining (pp. 837–845).

Cao N (2021) Explainable artificial intelligence for customer churning prediction in banking. In: Proceedings of the 2nd International Conference on Human-centered Artificial Intelligence (Computing4Human 2021) (pp. 159–167)

Cao N (2021) Explainable artificial intelligence for customer churning prediction in banking.

Cardenas-Ruiz C, Mendez-Vazquez A, Ramirez-Solis LM (2022) Explainable model of credit risk assessment based on convolutional neural networks. In: Advances in Computational Intelligence: 21st Mexican International Conference on Artificial Intelligence, MICAI 2022, Monterrey, Mexico, October 24–29, 2022, Proceedings, Part I (pp 83–96). Cham: Springer Nature

Cartwright H (2023) Interpretability: Should—and can—we understand the reasoning of machine-learning systems? In: Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research, OECD Publishing, Paris

Chen B, Wang X, Wang Y, Guo W (2020) An interpretable personal credit evaluation model. In: Data science: 6th international conference of pioneering computer scientists, engineers and educators, ICPCSEE 2020, Taiyuan, September 18–21, 2020, Proceedings, Part II 6 (pp. 521–539). Springer, Singapore.

Chen Z, Li J (2006) Least squares support feature machine. In 2006 International conference on computational intelligence and security vol 1, pp 176–179. IEEE

Chiua JY, Yan Y, Xuedongb G, Chen RC (2010). a new method for estimating bank credit risk. In 2010 International Conference on Technologies and Applications of Artificial Intelligence (pp 503–507). IEEE.

Chou TN (2019) A practical grafting model based explainable AI for predicting corporate financial distress. In: Business information systems workshops: BIS 2019 international workshops, Seville, June 26–28, 2019, Revised Papers 22 (pp. 5–15). Springer International Publishing.

Chlebus M, Gajda J, Gosiewska A, Kozak A, Ogonowski, D, Sztachelski J, Wojewnik P (2021) Enabling machine learning algorithms for credit scoring—explainable artificial intelligence (XAI) methods for clear understanding complex predictive models (No. 2104.06735).

Demajo LM, Vella V, Dingli A (2020) Explainable ai for interpretable credit scoring. arXiv preprint arXiv:2012.03749.

Dinu MC, Hofmarcher M, Patil VP, Dorfer M, Blies PM, Brandstetter J, Hochreiter S (2022) XAI and Strategy Extraction via Reward Redistribution. In: xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Revised and Extended Papers (pp. 177–205). Cham: Springer International Publishing.

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

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

Eluwole OT, Akande S (2022) Artificial intelligence in finance: possibilities and threats. In 2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT) (pp 268–273). IEEE

European Commission (2020) On artificial intelligence:a European approach to excellence and trust. White Paper.

Explainable AI (2019). The basics policy briefing. Available at Royal Society. org/ai-interpretability.

Fukas P, Rebstadt J, Menzel L, Thomas O (2022) Towards explainable artificial intelligence in financial fraud detection: using shapley additive explanations to explore feature importance. In: Advanced Information Systems Engineering: 34th International Conference, CAiSE (2022) Leuven, Belgium, June 6–10, 2022, Proceedings. Springer International Publishing, Cham, pp 109–126

Ghandar A, Michalewicz Z (2011) An experimental study of multi-objective evolutionary algorithms for balancing interpretability and accuracy in fuzzy rulebase classifiers for financial prediction. In: 2011 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr) (pp 1–6). IEEE.

Ghorbani A, Abid A, Zou J (2019) Interpretation of neural networks is fragile. In: Proceedings of the AAAI conference on artificial intelligence vol 53(1), pp 3681–3688.

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp 80–89). IEEE

Golbin I, Lim KK, Galla D (2019) Curating explanations of machine learning models for business stakeholders. In: 2019 Second International Conference on Artificial Intelligence for Industries (AI4I) (pp 44–49). IEEE.

Gomez O, Holter S, Yuan J, Bertini E (2020) Vice: visual counterfactual explanations for machine learning models. In: Proceedings of the 25th International Conference on Intelligent User Interfaces (pp 531–535)

Guan M, Liu XY (2021) Explainable deep reinforcement learning for portfolio management: an empirical approach. In: Proceedings of the Second ACM International Conference on AI in Finance (pp 1–9).

Hadash S, Willemsen MC, Snijders C, IJsselsteijn WA (2022). Improving understandability of feature contributions in model-agnostic explainable AI tools. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp 1–9)

Hajek P (2019) Interpretable fuzzy rule-based systems for detecting financial statement fraud. In: Artificial Intelligence Applications and Innovations: 15th IFIP WG 12.5 International Conference, AIAI 2019, Hersonissos, Crete, Greece, May 24–26, 2019, Proceedings 15 (pp 425–436). Springer International Publishing.

Harkut DG, Kasat K (2019) Introductory chapter: artificial intelligence challenges and applications. Artificial Intelligence-Scope and Limitations.

Hickey JM, Di Stefano PG, Vasileiou V (2021) Fairness by explicability and adversarial SHAP learning. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III (pp 174–190). Springer International Publishing.

Hsin YY, Dai TS, Ti YW, Huang MC (2021) Interpretable electronic transfer fraud detection with expert feature constructions. In: CIKM Workshops.

Jaeger M, Krügel S, Marinelli D, Papenbrock J, Schwendner P (2021) Interpretable machine learning for diversified portfolio construction. Markus Jaeger, Stephan Krügel, Dimitri Marinelli, Jochen Papenbrock and Peter Schwendner. J Finan Data Sci Summer.

Johansson F, Shalit U, Sontag D (2016) Learning representations for counterfactual inference. In: International conference on machine learning (pp 3020–3029). PMLR

Kamalloo E, Abadeh MS (2010) An artificial immune system for extracting fuzzy rules in credit scoring. In: IEEE Congress on Evolutionary Computation (pp 1–8). IEEE

Kiefer S, Pesch G (2021) Unsupervised Anomaly Detection for Financial Auditing with Model-Agnostic Explanations. In KI 2021: Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event, September 27–October 1, 2021, Proceedings 44 (pp. 291–308). Springer International Publishing.

Kim S, Woo J (2021) Explainable AI framework for the financial rating models: Explaining framework that focuses on the feature influences on the changing classes or rating in various customer models used by the financial institutions. In 2021 10th International Conference on Computing and Pattern Recognition (pp 252–255).

Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Technical Report.

Kong K, Liu R, Zhang Y, Chen Y (2020) Predicting liquidity ratio of mutual funds via ensemble learning. In: 2020 IEEE International Conference on Big Data (Big Data) (pp 5441–5450). IEEE

Li L, Zhao T, Xie Y, Feng Y (2020) Interpretable machine learning based on integration of nlp and psychology in peer-to-peer lending risk evaluation. In: Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9 (pp 429–441). Springer International Publishing.

Liu Q, Liu Z, Zhang H, Chen Y, Zhu J (2021) Mining cross features for financial credit risk assessment. In: Proceedings of the 30th ACM international conference on information & knowledge management (pp 1069–1078)

Lusinga M, Mokoena T, Modupe A, Mariate V (2021) Investigating statistical and machine learning techniques to improve the credit approval process in developing countries. In: 2021 IEEE AFRICON (pp 1–6). IEEE

Madakkatel I, Chiera B, McDonnell MD (2019) Predicting financial well-being using observable features and gradient boosting. In AI 2019: advances in artificial intelligence: 32nd Australasian Joint Conference, Adelaide, SA, Australia, December 2–5, 2019, Proceedings 32 (pp 228–239). Springer International Publishing.

Maree C, Omlin CW (2022) Understanding spending behavior: recurrent neural network explanation and Interpretation. In: 2022 IEEE symposium on computational intelligence for financial engineering and economics (CIFEr) (pp 1–7). IEEE.

Maree C, Modal JE, Omlin CW (2020) Towards responsible AI for financial transactions. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI) (pp 16–21). IEEE

Masyutin A, Kashnitsky Y (2017) Query-based versus tree-based classification: application to banking data. In Foundations of Intelligent Systems: 23rd International Symposium, ISMIS 2017, Warsaw, Poland, June 26–29, 2017, Proceedings 23 (pp 664–673). Springer International Publishing

Misheva BH, Osterrieder J, Hirsa A, Kulkarni O, Lin SF (2021) Explainable AI in credit risk management. arXiv preprint arXiv:2103.00949.

Mohammadi K, Karimi AH, Barthe G, Valera I (2021) Scaling guarantees for nearest counterfactual explanations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp 177–187)

Müller R, Schreyer M, Sattarov T, Borth D (2022) RESHAPE: explaining accounting anomalies in financial statement audits by enhancing SHapley Additive exPlanations. In: Proceedings of the Third ACM International Conference on AI in Finance (pp 174–182)

Nicosia G, Ojha V, La Malfa E, La Malfa G, Jansen G, Pardalos PM, Umeton R (Eds.) (2022) Machine learning, optimization, and data science: 7th International Conference, LOD 2021, Grasmere, UK, October 4–8, 2021, Revised Selected Papers, Part II (Vol. 13164). Springer Nature

OECD (2021) Artificial intelligence, machine learning and big data in finance: opportunities, challenges, and implications for policy makers, https://www.oecd.org/finance/artificial-intelligence-machine-learningbig-data-in-finance.htm.

Ohana JJ, Ohana S, Benhamou E, Saltiel D, Guez B (2021) Explainable AI (XAI) models applied to the multi-agent environment of financial markets. In: Explainable and Transparent AI and Multi-Agent Systems: Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers 3 (pp 189–207). Springer International Publishing.

Papenbrock J, Schwendner P, Jaeger M, Krügel S (2021) Matrix evolutions: synthetic correlations and explainable machine learning for constructing robust investment portfolios. Jochen Papenbrock, Peter Schwendner, Markus Jaeger and Stephan Krügel The Journal of Financial Data Science Spring.

Patron G, Leon D, Lopez E, Hernandez G (2020) An Interpretable Automated Machine Learning Credit Risk Model. In Applied Computer Sciences in Engineering: 7th Workshop on Engineering Applications, WEA 2020, Bogota, Colombia, October 7–9, 2020, Proceedings 7 (pp. 16–23). Springer International Publishing.

Pawelczyk M, Broelemann K, Kasneci G (2020) Learning model-agnostic counterfactual explanations for tabular data. In: Proceedings of The Web Conference 2020 (pp. 3126–3132).

Petersone S, Tan A, Allmendinger R, Roy S, Hales J (2022) A data-driven framework for identifying investment opportunities in private equity. arXiv preprint arXiv:2204.01852.

Pokhariya J, Mishra PK, Kandpal J (2022) Machine learning for intelligent analytics. In: Advances in Cyber Security and Intelligent Analytics (pp 219–234). CRC Press.

Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp 1135–1144)

Rius ADDM (2023) Foundations of artificial intelligence and machine learning. In: Artificial Intelligence in Finance (pp 2–18). Edward Elgar Publishing

Rodríguez M, Leon D, Lopez E, Hernandez G (2022). Globally Explainable AutoML Evolved Models of Corporate Credit Risk. In: Applied Computer Sciences in Engineering: 9th Workshop on Engineering Applications, WEA 2022, Bogotá, Colombia, November 30–December 2, 2022, Proceedings (pp 19–30). Cham: Springer Nature Switzerland.

Samoili S, López CM, Gómez E, De Prato G, Martínez-Plumed F, Delipetrev B (2020) AI watch, defining artificial intelligence. EUR 30117 EN. Publications Office of the European Union, Luxembourg

Silva, W., Fernandes, K., & Cardoso, J. S. (2019, July). How to produce complementary explanations using an ensemble model. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1–8). IEEE.

Sovrano F, Vitali F (2022) How to quantify the degree of explainability: experiments and practical implications. In: 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp 1–9). IEEE.

Speith T (2022) A review of taxonomies of explainable artificial intelligence (XAI) methods. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp 2239–2250).

Sprockhoff J, Lukic B, Janson V, Ahlbrecht A, Durak U, Gupta S, Krueger T (2023) Model-based systems engineering for AI-based systems. In AIAA SCITECH 2023 Forum (p 2587)

Stevens A, Deruyck P, Van Veldhoven Z, Vanthienen J (2020) Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI) (pp 1241–1248). IEEE

Sudjianto A, Knauth W, Singh R, Yang Z, Zhang A (2020) Unwrapping the black box of deep ReLU networks: interpretability, diagnostics, and simplification. arXiv preprint arXiv:2011.04041

Sudjianto A, Zhang A, Yang Z, Su Y, Zeng N (2023) PiML toolbox for interpretable machine learning model development and validation. arXiv preprint arXiv:2305.04214.

Szwabe A, Misiorek P (2018). Decision trees as interpretable bank credit scoring models. In Beyond databases, architectures and structures. facing the challenges of Data Proliferation and Growing Variety: 14th International Conference, BDAS 2018, Held at the 24th IFIP World Computer Congress, WCC 2018, Poznan, Poland, September 18–20, 2018, Proceedings 14 (pp 207–219). Springer International Publishing

The Royal Society (2019) Explainable AI: the basics policy briefing. https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf

Tian Y, Liu G (2020) MANE: model-agnostic non-linear explanations for deep learning model. In 2020 IEEE World Congress on Services (SERVICES) (pp 33–36). IEEE

Torky M, Gad I, Hassanien AE (2024) Explainable AI model for recognizing financial crisis roots based on pigeon optimization and gradient boosting model (Retraction of Vol 16, art no 50, 2023).

Tsakonas A, Dounias G (2005) An architecture-altering and training methodology for neural logic networks: application in the banking sector. ICINCO

Tyagi S (2022) Analyzing machine learning models for credit scoring with explainable AI and optimizing investment decisions. arXiv preprint arXiv:2209.09362

Walambe R, Kolhatkar A, Ojha M, Kademani A, Pandya M, Kathote S, Kotecha K (2021) Integration of explainable AI and blockchain for secure storage of human readable justifications for credit risk assessment. In Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part II 10 (pp. 55–72). Springer Singapore.

Wang J, Zhang Y, Tang K, Wu J, Xiong Z (2019) Alphastock: a buying-winners-and-selling-losers investment strategy using interpretable deep reinforcement attention networks. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp 1900–1908)

Xu P, Ding Z, Pan M (2017).An improved credit card users default prediction model based on RIPPER. In: 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) (pp 1785–1789). IEEE

Xu R, Meng H, Lin Z, Xu Y, Cui L, Lin J (2021) Credit default prediction via explainable ensemble learning. In: 5th International Conference on Crowd Science and Engineering (pp 81–87)

Xu YL, Calvi GG, Mandic DP (2021) Tensor-train recurrent neural networks for interpretable multi-way financial forecasting. In: 2021 International Joint Conference on Neural Networks (IJCNN) (pp 1–5). IEEE.

Zeng EZ, Gunraj H, Fernandez S, Wong A (2023) Explaining explainability: towards deeper actionable insights into deep learning through second-order explainability. arXiv preprint arXiv:2306.08780.

Zhang R, Yi C, Chen Y (2020) Explainable machine learning for regime-based asset allocation. In: 2020 IEEE International Conference on Big Data (Big Data) (pp 5480–5485). IEEE

Zhang X, Du Q, Zhang Z (2020) An explainable machine learning framework for fake financial news detection.

Zhang Z, Liu X, Gao Z, Qu Y (2020) Interpretable weighted soft decision forest for credit scoring. In: Recent trends in decision science and management: proceedings of ICDSM 2019 (pp 87–95). Springer Singapore.

Zheng Y, Yang Y, Chen B (2021) Incorporating prior financial domain knowledge into neural networks for implied volatility surface prediction. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp 3968–3975)

Zhou D, Zheng L, Zhu Y, Li J, He J (2020) Domain adaptive multi-modality neural attention network for financial forecasting. In Proceedings of The Web Conference 2020 (pp 2230–2240).

Zhu M, Wang Y, Wu F, Yang M, Chen C, Liang Q, Zheng X (2022) WISE: Wavelet based Interpretable Stock Embedding for Risk-Averse Portfolio Management. In: Companion Proceedings of the Web Conference 2022 (pp 1–11).

Zhu Y, Yi C, Chen Y (2020) Utilizing macroeconomic factors for sector rotation based on interpretable machine learning and explainable AI. In 2020 IEEE International Conference on Big Data (Big Data) (pp 5505–5510). IEEE.

Zurada J (2010) Could decision trees improve the classification accuracy and interpretability of loan granting decisions? In: 2010 43rd Hawaii International Conference on System Sciences (pp 1–9). IEEE.