Artificial Intelligence Review (2025) 58:189
https://doi.org/10.1007/s10462-024-11077-7

A comprehensive review on financial explainable AI

Wei Jie Yeo[1] · Wihan Van Der Heever[1] · Rui Mao[1] · Erik Cambria[1] · Ranjan Satapathy[2] · Gianmarco Mengaldo[3]

Accepted: 13 December 2024 / Published online: 29 March 2025

Abstract
The success of artificial intelligence (AI), and deep learning models in particular, has led to their widespread adoption across various industries due to their ability to process huge amounts of data and learn complex patterns. However, due to their lack of explainability, there are significant concerns regarding their use in critical sectors, such as finance and healthcare, where decision-making transparency is of paramount importance. In this paper, we provide a comparative survey of methods that aim to improve the explainability of deep learning models within the context of finance. We categorize the collection of explainable AI methods according to their corresponding characteristics, and we review the concerns and challenges of adopting explainable AI methods, together with future directions we deemed appropriate and important.

Keywords XAI · Explainable AI · Interpretable AI · Finance · FinXAI

✉ Gianmarco Mengaldo
mpegim@nus.edu.sg

Wei Jie Yeo
yeow0082@e.ntu.edu.sg

Wihan Van Der Heever
wihan001@e.ntu.edu.sg

Rui Mao
rui.mao@ntu.edu.sg

Erik Cambria
cambria@ntu.edu.sg

Ranjan Satapathy
satapathy_ranjan@ihpc.a-star.edu.sg

[1] College of Computing and Data Science, Nanyang Technological University, 50 Nanyang Ave, 639798 Singapore, Singapore

[2] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, 138632 Singapore, Singapore

[3] National University of Singapore, Asian Institute of Digital Finance at NUS, Singapore, 9 Engineering Drive 1, 117575 Singapore, Singapore

1 Introduction

Finance is a constantly evolving sector that is deeply rooted in the development of human civilization. One of the main tasks of finance is the efficient allocation of resources, with a chief example being the handling of capital flows between various entities with different needs. These entities can be divided into individuals, companies, and countries, and lead to the common categorization of personal, corporate, and government finance. The sector can be traced back to 5000 years ago, in the agrarian societies that had been established and developed for some thousand of years at the time. Indeed, one of the first examples of banking, a central institution within finance, can be attributed to the Babylonian empire. Since then, societal development and technological advances have pushed the field to undergo several changes. In the past two decades, these changes have been particularly marked, due to the accelerating pace of technological development, especially in the context of AI. The latter has started spreading across multiple segments of finance, from digital transactions to investment management, risk management, algorithmic trading, and more (Team 2022). The use of novel AI- and non-AI technologies to automate and improve financial processes is now known as FinTech (Financial Technology), and its growth in the past two decades has been remarkable (Mroczkowska 2020). In this review, we focus on AI-based technologies and machine learning for financial applications.

Financial researchers and practitioners have been relying on supervised, unsupervised, and semi-supervised machine learning methods as well as reinforcement learning for tackling many different problems. Some examples include credit evaluation, fraud detection, algorithmic trading, and wealth management. In supervised-based machine learning methods it is common to use e.g., neural networks to identify complex relationships hidden in the available labeled data. The labels are usually provided by domain experts. For instance, one can think of building a stock-picking system, where a domain expert labels periods of positive and negative returns. The machine is then tasked to build the relationship between (possibly) high-dimensional data, and positive and negative returns of a given stock (or multiple stocks) and generalize to unseen data to e.g., predict the future stock's behavior (Ma et al. 2023, 2024). In unsupervised-based machine learning methods, the task is instead to identify data with similar characteristics that can therefore be clustered together (Aghabozorgi and Teh 2014), without domain-expert labeling. For example, one can think of identifying all stocks that have similar characteristics into clusters using some similarity metrics, such as valuation, profitability and risk. Semi-supervised learning is a middle ground between supervised and unsupervised learning, where only a portion of the data is labeled. Finally, reinforcement learning aims to maximize, through a set of actions, the cumulative reward specified by the practitioners. Reinforcement learning is used in finance for e.g., portfolio construction. Reinforcement learning is strictly related to Markov decision processes and substantially differs from both supervised and unsupervised learning.

Among supervised, unsupervised, and reinforcement learning methods, there is vast heterogeneity in terms of complexity. Some methods are considered easier to understand, hence to interpret by practitioners (also referred to as white-box methods), while others are considered not interpretable (also referred to as black-box methods). To this end, neural networks and deep learning strategies, that underpin the majority (albeit not the entirety) of recent machine learning methods for financial applications, are considered black-box methods - i.e., the reason for a given prediction is not of easy access when available). This

constitutes a critical issue, especially in risky and highly regulated sectors, such as healthcare and finance, where a wrong decision may lead to catastrophic loss of life (healthcare) or capital (finance).

Additionally, the utilization of large language models (LLMs) has experienced a significant surge, primarily fueled by the release of OpenAI's ChatGPT and GPT-4 models (Achiam et al. 2023). These models have demonstrated their versatility and efficacy across an extensive spectrum of applications, from commonsense reasoning to solving complex mathematical challenges (Mao et al. 2024). Researchers have rapidly embraced LLMs in the financial sector, notable for the development of BloombergGPT (Wu et al. 2023), a specialized 50-billion-parameter decoder-only transformer. This model is distinguished by its pre-training on a comprehensive dataset including general and domain-specific financial texts. Subsequent efforts in the field have either involved the refinement of base LLMs through fine-tuning (Li et al. 2023, 2024; Xie et al. 2024; Yang et al. 2023) or the direct application of LLMs for inferential purposes (Lopez-Lira and Tang 2023; Xie et al. 2023), showcasing the broad applicability and adaptability of these models. Despite these advancements, challenges remain, particularly regarding the transparency and reliability of the outputs generated by these models which are also categorized under black-box methods. A significant concern is the propensity of LLMs to produce responses that, while seemingly accurate, may be based on false premises or commonly referred to as "*hallucinations*". This complicates the task for end-users to distinguish between factual and fabricated information. This issue is exacerbated by the notable frequency of such occurrences of hallucination, as documented in recent studies (Huang et al. 2023; Alkaissi and McFarlane 2023), highlighting the critical need for vigilance.

Hence, it was deemed important to understand the reasons (i.e., the data and patterns) the machine used to make a given decision. This aspect encompasses the broad field of AI transparency (Cambria et al. 2023). The latter is composed of three pillars, (i) AI awareness, (ii) AI model explainability, and (iii) AI outcome explainability. The first is tasked to understand whether AI is involved in a given product. The second is responsible to provide a detailed explanation of the AI model, including its inputs and outputs. The third is responsible to provide a granular explanation of the inputs' contributions to the AI model's outcomes. To this last category, we find a vast array of post-hoc interpretability methods. In this review, we assume that AI awareness is achieved, i.e., we know that in a given financial process AI is involved, and focus on AI explainability, also referred to as eXplainable AI or simply XAI. A further distinction commonly made is between interpretability and explainability of an AI model. These two terms, frequently used interchangeably, have subtle differences. Interpretability refers to how and why a model works. Explainability refers to the ability of explaining the results in human terms.

While deep learning methods are considered black-boxes, many other methods in finance are considered white-box methods. The trade-off between complexity and interpretability is perhaps one of the most debated aspects in the field of financial AI. On one hand, white-box methods are highly interpretable but lack the ability to grasp complex relationships, frequently failing to meet the desired performance. On the other hand, black-box methods are not interpretable but usually (although not always) meet the desired performance. Therefore, it is not surprising that there are significant efforts being pushed forward in recent years to render black-box methods more interpretable, where the primary example is the field of deep learning.

1.1 Contribution

In this paper, we provide an extensive review of XAI methods for the financial field that we name FinXAI. Although there have been several surveys on XAI methods (Cambria et al. 2023; Sahakyan et al. 2021; Arrieta et al. 2020; Guidotti et al. 2018; Molnar 2020; Mueller et al. 2019; Rojat et al. 2021), these papers are targeted towards XAI in general domains and are not specific to finance. Other similar works in finance touches on minor aspects of explainability but are primarily focused on other topics such as sentiment analysis (Du et al. 2024) and financial sustainability (Ong et al. 2024). Existing FinXAI works such as Chen et al. (2023) conduct statistical analysis on the FinXAI trends by using NLP techniques on the lexical level, however, the work lacks a comprehensive study of finer details such as introducing the various ways FinXAI can be categorized. Weber et al. (2023) conduct a broader investigation but lacks important categorization details such as audiences, explanation type, or data structures. We note that this is important to differentiate the applicability of general XAI from FinXAI. Our work addresses the identified gaps by offering a detailed categorization of existing XAI methodologies, tailored specifically for the financial research community. This categorization is designed to facilitate a clearer understanding of which XAI approaches are most applicable to various financial applications.

To compile this review, we reviewed over 100 related papers, focusing mainly, though not exclusively on the third pillar, i.e., the explainability of the inputs' contributions to the AI model's outcomes. We ensure the quality of our review by de-duplicating similar FinXAI works or excluding XAI works that do not analyze financial tasks, see Fig. 1. To this end, we perform a detailed breakdown of 68 papers where we considered both post-hoc interpretability methods applied to black-box deep learning models, and inherently transparent models that do not require further post-hoc interpretability. Despite the relatively small number of collected papers in the field of XAI, it is important to note that our main objective is to focus specifically on XAI tech-

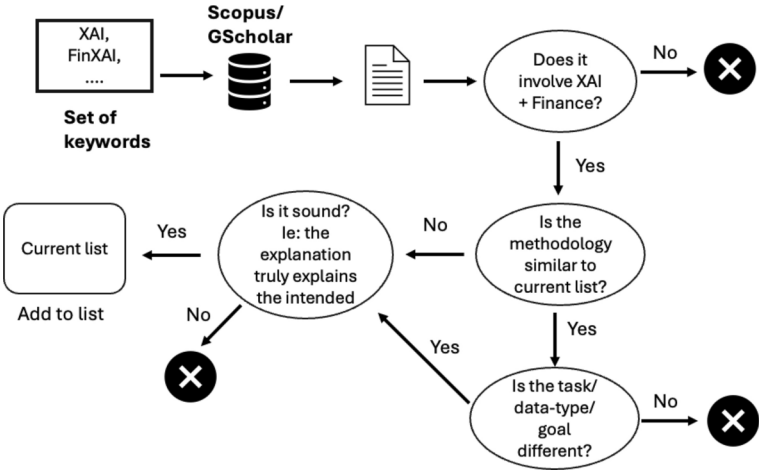

Fig. 1 We meticulously curated a list of high-quality FinXAI works that are likely to be of significant interest to the financial research community. The range of XAI methods applied in the financial sector is somewhat limited, predominantly focused on a narrow set of techniques. However, we included works that employ similar explainability methods but are applied to different tasks or aim to achieve distinct objectives

niques applicable to the financial industry. This targeted approach will provide valuable insights for researchers in related fields and will ultimately help drive innovation and progress in the financial industry. With the growing need for transparency and accountability of deep learning, the XAI community has seen increasing growth in the number of works published, we focus here instead only on works concerning financial use cases. Notably, FinXAI is but a small subset of the general field of XAI and, thus we take a holistic approach to assembling existing studies to keep up to date with the current approaches.

The reviewed articles were queried from both Google Scholar and Scopus where we searched using a set of keywords relating to works that have applied explainable AI techniques in financial use cases, the set of keywords include *"XAI, explainable AI, finance, financial sector, financial field, explainable ML, interpretable AI, credit evaluation, stock forecasting, financial explainable AI"*. Note that the keywords do not always include an overlap between 'finance' and 'XAI' since some works do not explicitly state the fusion of both in the title or keywords. Thus we manually filter out those that ultimately do not fit the mentioned criteria. We took a bottom-up approach as FinXAI is a relatively niche research area, with limited work compared to the general XAI field. We try to collect a diverse set of papers that covers each category sufficiently well, and summarized in Tables 1, 2, 3. In particular, we noticed a lack of counterfactual explainability works. Hence, we explicitly searched for XAI techniques that provide counterfactual explanations. Counterfactual explanations are deemed as a desirable form of explanation as the receiver tends to prefer understanding why a certain prediction was made instead of the opposing.

The main *contributions* of our work are as such:

- We provide an extensive study on consolidating XAI methods in the field of finance (FinXAI), for researchers interested in prioritizing transparency in their solutions.
- We frame the FinXAI process as a sequential flow of decision-making processes (see Fig 5), where we place importance on aligning the XAI technique with the target audience. The objective of this framework is to produce explanations that are both goal-oriented and audience-centric.
- We review current FinXAI techniques, analyze their technical contributions to ethical goals, and list down a number of key challenges faced in implementing XAI as well as important directions to be improved for the future.

The remainder of the review paper is organized as follows: Sect. 2 describes the definitions, reasons, and brief overview of FinXAI. Subsequently, we explain the methodology of FinXAI, starting from numerical in Sect. 3, textual in Sect. 4, hybrid analysis in Sect. 5 and ending with transparent models in Sect. 6. In Sect. 7, we analyze how the reviewed FinXAI methods contribute to ethical goals. Section 8 discusses key challenges of adopting explainable models and future directions for research. Finally, Sect. 9 offers concluding remarks.

2 FinXAI: definition, reason, approach

This section details the definition, purpose, and approaches that have been taken in improving the transparency of AI models. A collection of existing literature reviews are analyzed and collated to give the reader a better understanding of commonly used terminologies in

the XAI field, as well as the interlink between AI, linguistics, and social sciences which is essential to provide a solid understanding of the subject.

2.1 Definition of XAI

The general explainable AI methods is inherently linked to the broader concept of AI transparency. As mentioned in Sect. 1, this term encompasses under a single framework three key steps: AI awareness, AI model explainability, and AI outcome explainability. In this review, we focus on the latter two aspects, model and outcome explainability. Model explainability means that the inner workings of a given AI solution are interpretable, and therefore the results may be interpreted by humans. This is typically the case for models with reduced complexity (i.e., white-box models), such as linear and logistic regression, and decision trees.

Outcome explainability means that the inner workings of a given AI solution are not fully interpretable, and therefore the results may not be fully understandable by humans, unless some interpretability tools are applied to explain the AI outcomes. This is the case for complex models (i.e., black-box models), such as deep neural networks. In these cases, it is common to apply model agnostic post-hoc (and other) interpretability tools to understand the results the AI provided in human terms.

Correspondingly, XAI models may be cast into two broad categories: intrinsically explainable due to their highly interpretable nature (e.g., linear and logistic regression), and extrinsically explainable, hence requiring an external tool to make them interpretable. In turn, these two categories of models lead to different classes of model transparency: *simulatability*, *decomposability*, and *algorithmic transparency* (Arrieta et al. 2020). Each of these three classes inherits the preceding class' properties, that is, if a model is decomposable, it is also simulatable, and if a model is algorithmically transparent is also decomposable and simulatable. In simple terms, *simulatability* refers to the model's ability to allow a human observer to simulate a thought process over the inner workings of the model. *Decomposability* entails that interpretability is available at every segment of the model, including inputs, outputs as well as model inner workings and parameters. *Algorithmic transparency* largely deals with the human user being able to understand how the model reacts with varying inputs and more importantly the ability to reason about errors the model produces. An inherently transparent model exhibits the ability to provide human-understandable explanations without any additional layer of interpretability, albeit the criteria of validating what is perceived to be "human-understandable" is dynamic across various audiences.

2.2 Distinction between FinXAI and XAI

Since the focus of this work revolves around XAI in finance applications, it is important that a clear distinction is made. Although the above in general applies to FinXAI as well, there are several critical distinctions that set it apart from the broader XAI field. A key distinguishing feature of FinXAI is its multidisciplinary nature, requiring not only expertise in crafting accurate explanations but also the integration of insights from financial experts to ensure the explanations are both credible and persuasive to the target audience. In one regard, the areas of transparency not only concern the decision-making model itself but also the data and design process of the end-product (van den Berg and Kuiper 2020). For

Table 1 Classification of papers relating to *credit evaluation*

Paper	Transparency		Proximity		Explanation Procedure				
	Int	PH	Loc	Glo	Text	Vis	Ex	Simp	FR
Dikmen and Burns (2022)		✓	✓						✓
Gramespacher and Posth (2021); Chen and Ye (2022)	✓			✓					
Misheva et al. (2021); Serengil et al. (2022)		✓	✓	✓				✓	✓
Bussmann et al. (2021, 2020)		✓	✓	✓					✓
Rizinski et al. (2022)		✓	✓	✓					✓
Müller et al. (2022)		✓	✓	✓					✓
Zijiao et al. (2022)		✓	✓	✓		✓			✓
Biecek et al. (2021)		✓	✓	✓		✓			✓
Crosato et al. (2021)		✓	✓	✓		✓			✓
Davis et al. (2022)		✓	✓	✓			✓	✓	✓
Sudjianto and Zhang (2021); Dumitrescu et al. (2022)	✓		✓	✓					
Bueff et al. (2022)		✓		✓					✓
Fritz-Morgenthal et al. (2022); Tran et al. (2022)		✓		✓					✓
Srinivasan et al. (2019)	✓		✓		✓				
Grath et al. (2018)		✓	✓						✓
Adams and Hagras (2020)	✓		✓	✓					
Demajo et al. (2020)		✓	✓	✓			✓		✓
Luo et al. (2018)		✓	✓			✓			
Zhang et al. (2022a)		✓	✓			✓			

Paper	Audience				Data Analysis		Expl Type		Eval
	EU	Dev	DE	Reg	Num	Text	Fact	CF	
Dikmen and Burns (2022)	✓				✓		✓		
Gramespacher and Posth (2021); Chen and Ye (2022)				✓	✓	✓	✓		✓
Misheva et al. (2021); Serengil et al. (2022)		✓			✓		✓		
Bussmann et al. (2021, 2020)	✓	✓			✓		✓		
Rizinski et al. (2022)		✓			✓			✓	✓
Müller et al. (2022)				✓	✓		✓		

Table 1 (continued)

Paper	Audience				Data Analysis		Expl Type		Eval
	EU	Dev	DE	Reg	Num	Text	Fact	CF	
Zijiao et al. (2022)				✓	✓		✓	✓	
Biecek et al. (2021)	✓			✓	✓		✓		
Crosato et al. (2021)	✓				✓		✓		
Davis et al. (2022)		✓		✓	✓		✓		
Sudjianto and Zhang (2021); Dumitrescu et al. (2022)				✓	✓		✓		
Bueff et al. (2022)				✓	✓		✓	✓	
Fritz-Morgenthal et al. (2022); Tran et al. (2022)		✓	✓		✓		✓		
Srinivasan et al. (2019)	✓		✓			✓	✓		
Grath et al. (2018)	✓				✓			✓	✓
Adams and Hagras (2020)				✓	✓	✓	✓		
Demajo et al. (2020)	✓		✓	✓	✓		✓		✓
Luo et al. (2018)			✓			✓	✓		
Zhang et al. (2022a)	✓		✓		✓			✓	✓

The papers reviewed are split by task category and subsequently categorized by entailed properties. Missing options are either not stated or non-applicable. Intrinsic (Int), Post-hoc (PH), Local (Loc), Global (Glo), Textual (Text), Visual (Vis), By Example (Ex), Simplification (Simp), Feature relevance (FR), End-User (EU), Developer (Dev), Domain Expert (DE), Regulatory (Reg), Numerical (Num), Text, Factual (Fact), Counterfactual (CF), Explanation evaluation (Eval)

Table 2 Classification of papers relating to *financial prediction*

Paper	Transparency		Proximity		Explanation Procedure				
	Int	PH	Loc	Glo	Text	Vis	Ex	Simp	FR
Zhang et al. (2020)		✓		✓		✓			
Yang et al. (2018)		✓		✓		✓			
Deng et al. (2019)		✓	✓	✓		✓			
Ghosh and Sanyal (2021)		✓	✓	✓				✓	✓
Collaris et al. (2018)		✓	✓	✓				✓	✓
Benhamou et al. (2021); Fior et al. (2022)		✓	✓	✓					✓
Bracke et al. (2019)		✓	✓	✓					✓
Nazemi et al. (2022)	✓	✓		✓					
Carta et al. (2021)	✓			✓					
Cong et al. (2021)		✓		✓				✓	✓
Yasodhara et al. (2021)		✓		✓					✓
Park and Yang (2022); Islam et al. (2019); Weng et al. (2022); Wand et al. (2022); Vivek et al. (2022)		✓		✓					✓
Babaei et al. (2022)		✓		✓					✓
Lin et al. (2021)	✓		✓			✓			
Gite et al. (2021); Bandi et al. (2021)		✓	✓					✓	
Yan et al. (2019)		✓	✓					✓	
Carta et al. (2022)		✓	✓						✓
Kumar et al. (2022)		✓	✓						✓
Cho and Shin (2023)		✓	✓						✓
Shi et al. (2021); Kumar et al. (2017); Chen et al. (2020)		✓	✓			✓			
Achituve et al. (2019)		✓	✓			✓			
Farzad (2019)		✓		✓		✓			✓
Ong et al. (2023)	✓			✓					✓
Yuan and Zhang (2020); Koa et al. (2024)		✓	✓		✓				
Du et al. (2024)	✓			✓			✓		✓

Table 2 (continued)

Paper	Audience				Data Analysis		Expl Type		Eval
	EU	Dev	DE	Reg	Num	Text	Fact	CF	
Zhang et al. (2020)			✓		✓		✓		
Yang et al. (2018)		✓	✓			✓	✓		
Deng et al. (2019)		✓				✓	✓		
Ghosh and Sanyal (2021)	✓	✓	✓		✓	✓	✓		
Collaris et al. (2018)			✓		✓		✓		✓
Benhamou et al. (2021); Fior et al. (2022)			✓		✓		✓		
Bracke et al. (2019)	✓	✓	✓	✓	✓		✓		
Nazemi et al. (2022)			✓		✓		✓		
Carta et al. (2021)		✓	✓		✓	✓	✓		
Cong et al. (2021)			✓		✓	✓	✓		
Yasodhara et al. (2021)		✓			✓		✓		✓
Park and Yang (2022); Islam et al. (2019); Weng et al. (2022); Wand et al. (2022); Vivek et al. (2022)			✓		✓		✓		
Babaei et al. (2022)				✓	✓		✓		
Lin et al. (2021)			✓			✓	✓		
Gite et al. (2021); Bandi et al. (2021)		✓			✓	✓	✓		
Yan et al. (2019)				✓	✓		✓		
Carta et al. (2022)			✓		✓		✓		✓
Kumar et al. (2022)			✓✓		✓		✓		
Cho and Shin (2023)	✓				✓		✓	✓	
Shi et al. (2021); Kumar et al. (2017); Chen et al. (2020)	✓	✓			✓		✓		
Achituve et al. (2019)			✓		✓		✓		
Farzad (2019)	✓	✓	✓	✓	✓		✓		✓
Ong et al. (2023)			✓		✓	✓	✓		
Yuan and Zhang (2020); Koa et al. (2024)	✓					✓	✓		✓
Du et al. (2024)	✓		✓		✓		✓	✓	

The papers reviewed are split by task category and subsequently categorized by entailed properties. Missing options are either not stated or non-applicable

Table 3 Classification of papers relating to *financial analytics*

Paper	Transparency		Proximity		Explanation Procedure				
	Int	PH	Loc	Glo	Text	Vis	Ex	Simp	FR
Rallis et al. (2022)		✓	✓	✓					✓
Zhang et al. (2022)		✓	✓	✓		✓		✓	✓
Maree et al. (2020)	✓	✓		✓					✓
Maree and Omlin (2022a)		✓		✓					
Maree and Omlin (2022b)		✓		✓				✓	
Lachuer and Jabeur (2022)		✓		✓					✓
Gramegna and Giudici (2020)		✓		✓		✓			✓
Liu et al. (2020)		✓		✓		✓			
Yang et al. (2020)		✓	✓						✓
Ito et al. (2020)		✓	✓			✓			
Zhang et al. (2020)		✓	✓		✓				✓

Paper	Audience				Data Analysis		Expl Type		Eval
	EU	Dev	DE	Reg	Num	Text	Fact	CF	
Rallis et al. (2022)			✓		✓		✓		
Zhang et al. (2022)				✓	✓		✓	✓	
Maree et al. (2020)		✓			✓	✓	✓		
Maree and Omlin (2022a)	✓			✓	✓		✓		
Maree and Omlin (2022b)			✓		✓		✓		
Lachuer and Jabeur (2022)			✓		✓		✓		
Gramegna and Giudici (2020)		✓	✓		✓		✓		
Liu et al. (2020)				✓	✓	✓	✓		
Yang et al. (2020)				✓		✓		✓	
Ito et al. (2020)	✓		✓			✓	✓		
Zhang et al. (2020)		✓		✓	✓	✓	✓		

The papers reviewed are split by task category and subsequently categorized by entailed properties. Missing options are either not stated or non-applicable. There were no evaluation metrics present for these papers

example, the EU High-Level Expert Group on AI (HLEG 2019) states that the data the model interacted with should be traceable by human users at any given time. In addition, the design process of the system must be clear and explainable in a manner comprehensible to related stakeholders. The list of information types, regarded as explainable can even be extended to include principles and guidelines in the development of the AI system, as well as personnel involved in the implementation and development process (Kuiper et al. 2022).

A key goal for explainability is to gain the trust of affected stakeholders. Examples of such stakeholders include regulators, board members, auditors, end-users, and developers (Yeong Zee Kin 2023). To this end, the format and degree of explanation vary among audiences. The key message is usually conveyed in reports customized to the suitability of the receiving audience. It is common knowledge that financial service providers are regularly audited by supervisory authorities to ensure adherence to regulations and to prevent potential fraud from taking place. The level of scrutiny expected of the authorities is much higher than what the service providers expect. A study conducted by Kuiper et al. (2022) involves a preliminary investigation to identify the types of information that are deemed necessary, in the perspective of banks and supervisory authorities. The result was that supervisory authorities identify all forms of information types as relevant while banks only consider a subset of them. As such, there exists a gap between each organization's understanding of necessity, more often than not leading to the delay in approving the deployment of financial services.

As previously mentioned, the quality of an explanation is largely subjective and depends on the audience's needs. The level of required detail typically increases hierarchically, from end-users to regulatory authorities, as illustrated in Fig. 2. In this context, scrutiny refers to the depth of information deemed necessary. End-users generally require the least amount of explanation, focusing mainly on practical concerns such as the cause of an outcome and data security. Conversely, external regulators demand comprehensive explanations that cover every aspect of the end-product, including design guidelines, accountability, personnel involvement, deployment processes, and organizational training structures, in addition to addressing end-user requirements. Proximity refers to the region of explanation provided by the XAI technique and can be classified under local (reasons about a particular outcome) and global (view of the underlying reasoning and mechanics of the AI model). End-users tend to be concerned with how the outcome affecting them is provided (local proximity). For example, a person whose credit card application was rejected would want to know the underlying reason behind it. In contrast, the solution providers and regulators tend to focus on the internal operations and design workflow of the product, for reasons related to performance enhancement, fairness in the model's sense of judgment, and identification of biases in the prediction (global proximity). This necessity to tailor explanations to meet the specific needs of each audience is a defining characteristic of FinXAI, making it fundamentally **audience-centric**. Furthermore, the challenge of generating an appropriate explanation is intensified by the inherent preference biases of different audience types (Kuiper et al. 2022). Figure 5 illustrates our recommendations for incorporating this factor into the design of explainability solutions. While this challenge is not unique to financial applications, it is particularly pronounced in this domain due to the stringent need to comply with a comprehensive set of principles, which we discuss further in Sect. 2.3.

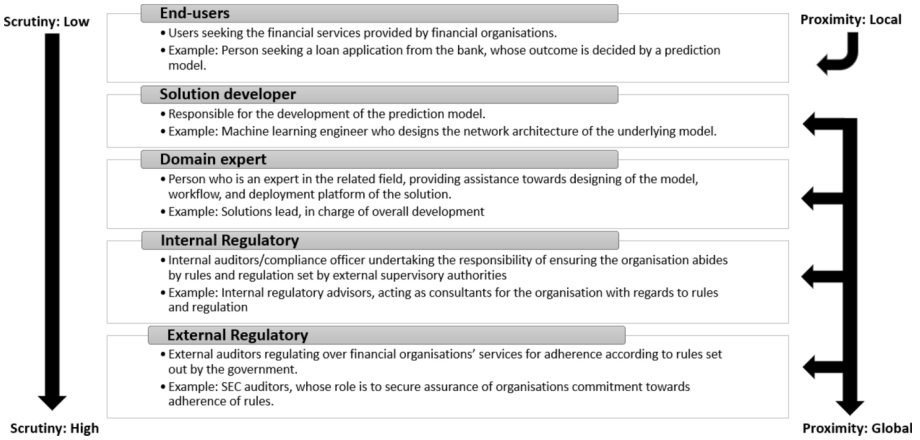

Fig. 2 Levels of explanation requirements by different audiences, categorized by explanation proximity, and ordered by scrutiny level. Local proximity refers to explanations concerned about a specific outcome. Global proximity refers to the underlying reasoning and mechanics of an AI model). End-users typically require are satisfied with local-proximity explanations, and the level of scrutiny is low. Developers, domain experts and regulatory authorities require global-proximity explanations instead, and the level of scrutiny is much higher

2.3 Reasons for FinXAI

As previously mentioned, various stakeholders lean towards different forms of explanation, naturally leading to different sets of goals the explanation can provide. Financial products typically undergo a rigorous verification process prior to deployment. Thus, a paramount reason for adopting explainable models is to ensure that financial solutions adhere to ethical standards outlined in the financial sector. The Monetary Authority of Singapore (MAS) stipulates that AI solutions should be developed in accordance with the Fairness, Ethics, Accountability and Transparency (FEAT) principles (of Singapore 2021). EU's General Data Protection Regulation (GDPR) (Goodman and Flaxman 2017) in 2018 announced a law referred as "right to explanation", dictating that individuals affected by automated decision-making solutions have a right to ask for an explanation of the outcome made for them.

The rising call for explainable models is mainly influenced by the rapid advancement of AI solutions and the increasing complexity surrounding them. More importantly, public cases of AI model displaying biases in their prediction magnifies the urge for explainable solutions. A famous example is Google's image recognition software, that accidentally labels dark-skin humans as gorillas (VINCENT 2018). Such biases can damage the company's reputation, and lead to profit losses.

The financial sector has its own set of ethics that should be upheld along with the desirable principles of AI ethics. These set of financial ethics often overlap with AI principles. An experiment involving 8 financial experts to investigate the relationship between the aforementioned sets was carried out in Rizinski et al. (2022). The results show that financial ethics (integrity, objectivity, competence, fairness, confidentiality, professionalism, diligence) has significant similarities with AI ethics (growth and sustainable development, human-centered values and fairness, transparency and explainability, safety and accountability). The strength of the links between each element was assessed, with integrity and fairness having

the strongest relationship with AI ethics. Indeed, this is understandable given that AI solutions should naturally embody these qualities, regardless of the industry taken into account.

As mentioned, the ethical goals set forth by XAI solutions differ among audiences, similar to the explanation types desired. Each audience is likely to be more affected by one than the other (Arrieta et al. 2020; Mohseni et al. 2021). Figure 3 list the ethical goals supported towards each set of audiences, and shows that there are some overlapping ethical goals across audiences. Referring to Arrieta et al. (2020), we provide a brief explanation of each ethical goal reported in Fig. 3, taking a financial perspective.

- *Trustworthiness*: Defined as instilling trust into users affected by the decisions of the AI model. Trustworthiness can be achieved when there is a high level of confidence in the model to constantly behave in the intended manner (Ribeiro et al. 2016). Trust is also sustained if the services provided are transparent and enables affected user to maintain their faith in the service providers. However, trust is a highly subjective quality and hence difficult to quantify. Judging if an explanation instills trust is mostly subjected to the affected users' opinions.
- *Fairness*: Refers to delivering AI solutions and explanations to every user and stakeholder equally, removing possible biases. Indeed, bias mitigation is a key constituent of fairness. The transparency of the AI model allows for a fair and socially ethical analysis, where any form of biases existing in the product chain are eliminated. In the financial markets, users tend to use the services provided by a firm if they are assured of fair and unbiased treatment.
- *Informativeness*: One important objective of AI models is to provide assistance to human counterparts in making decisions. Therefore, it is vital that the problem statement is made clear at all times. By providing explanations, the model benefits both from a

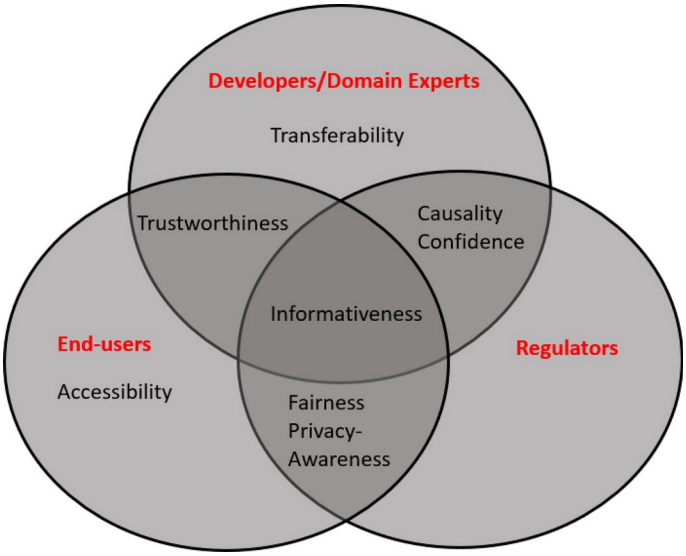

Fig. 3 Ethical goals are classified under three broad audiences: end-users, developers/domain experts, and internal/external regulatory authorities. Some ethical goals are shared by the three different audiences considered, such as informativeness (Arrieta et al. 2020)

social perspective as well as a performance standpoint, since knowing what is being done opens up opportunities for further refinement. Most of the papers in the literature dealing with this aspect aim at identifying relevant features which equates to highlighting parts of the input data the model is paying attention to. This can assist in debugging and allows pruning of unnecessary features which may cause overfitting.

- *Accessibility*: The main personnel interacting with algorithms are usually restricted to AI developers or domain experts, providing accessibility could allow for non-experts to get involved. This can be seen as an important stepping stone for making AI prevalent and well-accepted by the general society. Likewise, complicated algorithms deter financial companies from adopting such solutions, since extensive training is required while having to fear potential repercussions in the case of any unintended wrongdoings. If a model is able to relate its mechanisms in easily understandable terms, it can ease the fear of users and encourage more organizations to adopt such practices.
- *Privacy Awareness*: Not knowing the full limits of accessibility in the data can result in a breach of privacy. Likewise, such an issue triggers concerns within the overall design workflow. Accountable personnel in the designing process should ensure third parties are only allowed restricted access to the end-users data and prevent any misuse which can disrupt data integrity. Privacy awareness is especially important in the financial sector due to the amount and sensitivity of the information being captured.
- *Confidence*: The AI model should provide not only an outcome but also the confidence it has in the decision-making process, allowing domain experts to identify uncertainty in both model's results as well as the region of data captured. Stability in the prediction can be used to access a model's confidence while explanations provided by the model should only be trusted if it produces results that are consistent across different data inputs.
- *Causality*: It is usually in the interest of developers or experts to understand the causality between data features. However, proving it is a difficult task that requires extensive experimenting. Correlation can be involved in assessing causality, though it is frequently not representative of causality. Since AI models only discover correlations among the data they learn from, domain experts are usually required to perform a deeper analysis of causal relationships.
- *Transferability*: Allowing for the distillation of knowledge learned from AI models is an extensive area of research, a notable benefit is that it allows for the reusability of different models and averts endless hours of re-training. However, the complexity of the algorithms limits experts from deploying trained models in different domains. For example, a model trained to forecast future stock prices can likely be used to predict other financial variables such as bond price, market volatility, or creditworthiness, if the model behavior in these circumstances is known. Delivering an intuition of the inner workings can ease the burden of experts to facilitate adapting the knowledge learned, reducing the effort required for fine-tuning. Transferability is arguably one of the essential properties for the improvement of future AI models.

2.4 Approach of FinXAI

The review provided in this paper aims to give the readers an overall view of the XAI methodologies developed thus far in the financial industry. We note that explainability can

be injected across different stages of the development cycle. These stages include: *pre-modeling, modeling, and post-modeling* (Mellon 2021). Pre-modeling stage refers to the process chain before the designing stage of the AI model, this can include preliminary procedures which focus on identifying salient features by accessing readily available domain knowledge (Islam et al. 2019). The modeling phase includes any adjustment to the model's architecture or optimization objective. As a start, simpler transparent models should be preferred over complex black-box models if the problem at hand is not too complicated. Most of the papers in the review focus on the post-modeling stage, mainly due to the flexibility and ease of designing explainability techniques. Since the outcome is provided, it provides developers with more information to design an appropriate explanation method towards the form of data interacted (See Fig. 4). Most XAI techniques tend to focus on one stage of the modeling process, though it is possible to do so in two or more.

The focused regions of finance can be broadly categorized under three sections Bahrammirzaee (2010): *credit evaluation* (peer-to-peer lending, credit assessment, credit risk management, credit scoring, accounting anomalies), *financial prediction* (Asset allocation, stock index prediction, market condition forecasting, volatility forecasting, algorithmic trading, financial growth rate, economic crisis forecast, bankruptcy prediction, fraud detection (Athey et al. 2018), mortgage default) and *financial analytics* (financial text classification, spending behavior, financial corporate social responsibility (CSR), customer satisfaction). Following the task classification, we further differentiate the studies based on the underlying characteristics of the XAI technique as shown in Table 1, 2, 3. Specifically, we seek to answer questions such as "*What form of explanation is provided?*" (explanation procedure), "*Who is the explanation intended for?*" (audience), "*What kind of explanation is provided?*" (proximity, explanation type).

- *Transparency*: As mentioned in Sect. 2.4, interpretability of the model is either derived via interpreting the internal mechanisms of the AI model or through external techniques aimed at delivering some form of visualization or intuition of how the model works. Most of the reviewed papers focus on post-hoc explainability techniques, which we believe are preferred for a number of reasons. Intrinsic models usually under-perform

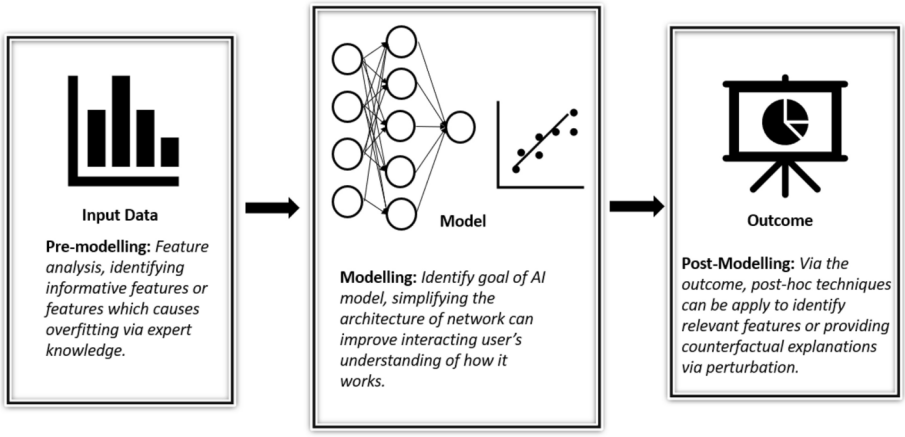

Fig. 4 Different stages where interpretability can be injected into the design workflow (Mellon 2021)

complex networks and as such, producing explanations for an inaccurate prediction is pointless. We additionally note that the method of conveying explanations for intrinsic models is by definition model-specific. This means the same method cannot be reused for a different model. While post-hoc techniques can be agnostic or specific towards any single model.

- *Proximity*: The explanations provided by XAI tools can seek to explain either the derivation of an outcome, known as local explanation, or how the model outputs on a global scale, referred to as global explanation. Global explanations tend to provide information on how the model makes decisions globally based on the learned weights, data features, and structure of the network. Producing an acceptable global explanation tends to be difficult in most cases (Molnar 2020) as opposed to just a region of the input data. On the other hand, local explanations focus on a specific region of the dataset and seek to assist the receiver in understanding how a particular prediction is made. Local explanation is more accurate for unique cases where the dependency on input features is rarely captured by the AI model, which can cause global explanations to ignore such dependency. End-users tend to prefer local explanations as their concern lies with the explanation surrounding their outcome. Regulators and financial experts, on the other hand, prefer global explanations in order to have a complete understanding of the model.
- *Explanation Procedure*: According to Arrieta et al. (2020), the various forms of post-hoc XAI techniques can be divided into several sections: text explanation (TE), visual explanation (VE), explanation by example (EE), explanation by simplification (ES) and feature relevance (FR). TE provides an explanation via text generation. Natural language tends to be easily understood by non-experts and is a common source of information in human society. VE enables visual understanding of the model's behavior, which may be preferable for image features (Selvaraju et al. 2017), such methods comprise graphical plots for both local and global explainability. EE captures a smaller subset of examples which represents the correlations modeled by the black-box model at a high level. ES techniques build a simpler surrogate model to approximate the underlying black-box model with high fidelity yet being interpretable. FR techniques aim to identify features deemed relevant for the model's prediction, by computing a relevance score for each feature. FR can account for explainability at both local and global levels and constitutes the largest share among the reviewed papers in our literature.
- *Audience*: Since the quality of explanations is subjective, it is very difficult to derive a one-fit-all explanation and hence, explanations should be customized towards one's needs. The examples of audiences are referenced from Fig 2, while we further merge internal and external regulators together. We highlight that aligning the objective of the explanation to the audience receiving it is important (Tomsett et al. 2018). Determining if an explanation is considered meaningful, is dependent on the target goals respective of each audience. Financial regulators, for example, would not be very concerned with understanding what sort of AI model or ML technique is used, but rather on the aspect of data privacy, model biases, or unfair treatment between affected end-users. It is uncommon for a single explanation to be deemed acceptable to audiences holding different positions in a financial company. An example is that the explanation produced for the developer tends to require additional customization before submitting to the immediate superior and the same applies to the proceeding higher-ups and external end-users.
- *Data Type*: The most commonly used forms of input data among the reviewed papers

consists of text, images, and numerical values. In terms of frequency among the forms of available data, numerical features are the most common source of information used in the financial industry. Images represent the least utilized source, as they tend to be storage intensive and contain a large amount of redundant information or are not applicable for most use cases. We only found a single work using image features. Chen et al. (2020) perform classification of eight different candlestick patterns and the explanation is delivered through monitoring changes in prediction after applying adversarial attacks. Surprisingly, textual information is not used as frequently as expected, albeit being a valuable source of information for deriving market sentiment or understanding consumers' emotions towards certain aspects of the business product. It is also possible to unify multiple sources of information, otherwise known as multi-model data. A boost in performance can be achieved, for instance by combining the patterns learned from time-series features and sentiment from textual features.

- *Explanation Type*: A single explanation can be conveyed in various forms, including factual, contrastive, and counterfactual explanations (Miller 2019). Factual delivers straightforward explanations that seek to answer the question *"Why does X lead to Y"* as opposed to contrastive *"Why does X lead to Y instead of Z"*. Counterfactual instead reasons how the consequent can be changed with respect to the antecedent, answering the question *"how to achieve Z by changing X"*. Humans tend to prefer contrastive rather than factual explanations since the latter can have multiple answers and referring to Miller (2019), explanations are selective. As humans tend to ignore a large portion of the explanations except for the important ones due to cognitive bias. For example, if Person A's loan application was rejected, there could be numerous reasons for this, such as *"Person A's income was too low for the past 6 months*", *"Person A's only have 1 existing credit card"*, *"Person A has had a credit default 3 months ago"* and so on. Whereas a contrastive explanation can instead involve comparing against another applicant whose outcome contrasts the target applicant's and an explanation can be made, highlighting the most significant factor. As argued by Lipton (1990), contrastive explanations are easier to deliver as one does not have to investigate the entire region of causes but rather a subset of it. Counterfactual explanations then seek to provide solutions for the contrastive explanation, commonly done by identifying the smallest changes to the input features, such that the outcome can be altered towards the alternative.
- *Explanation Evaluation*: Despite the extensive studies carried out to investigate what defines a good explanation, it is difficult to qualitatively compare among interpretations. The quality of an explanation is mostly subjective as a single explanation can be perceived with varying opinions among audiences. Nonetheless, there exist a number of studies that provides a quantitative approach to evaluating explanations. These measurements can be derived from human experts (Yang et al. 2020), referencing financial ethical goals (Adams and Hagras 2020) or through statistical methods (Müller et al. 2022). Islam et al. (2019) conducted a comparison between feature importance techniques in time series data and proposed a multivariate dataset that deals with the inability of techniques that identify salient time-series features. A vast majority of the reviewed papers focused only on evaluating the performance of the prediction model and consider it as a proxy for the quality of the explanation. We argue that such evaluation does not fully represent the quality of the explanation and even if so, it may not be suitable for every form of explanation procedure.

Selection Procedure: We design a framework shown in Fig. 5, framing the designing of the XAI solution as a sequential decision-making process. The selection categories can be referenced from Tables 1, 2, 3. The sequential structure of the framework ensures the explanation provided is tailored to the audience's needs while achieving the goal set out with respect to the target audience. We note that certain properties of the XAI technique have inner dependencies with each other, such as the relationship between explanation proximity and target audience. The quality of the explanation is evaluated and serves as feedback for any necessary adjustment, resulting in an audience-centric explanation.

3 XAI with numerical features

Numerical features are a common source of information across all aspects of data-driven methodologies. Financial tasks such as credit scoring of individuals/firms and financial market forecasts commonly use a collection of historical numerical features, such as stock price, trade volume, and volatility, and apply various forms of data-driven models to make predictions. These data-driven models may include supervised learning approaches, e.g., classification and regression tasks, and unsupervised learning approaches, e.g., clustering tasks. The use of numerical features within the context of finance is well established, hence it is not surprising that the majority of reviewed studies focus on this area. In the following, we outline the main approaches used for explainability in this context, namely visual explanation, explanation by simplification, feature relevance, and explanation by example, and conclude with a brief summary.

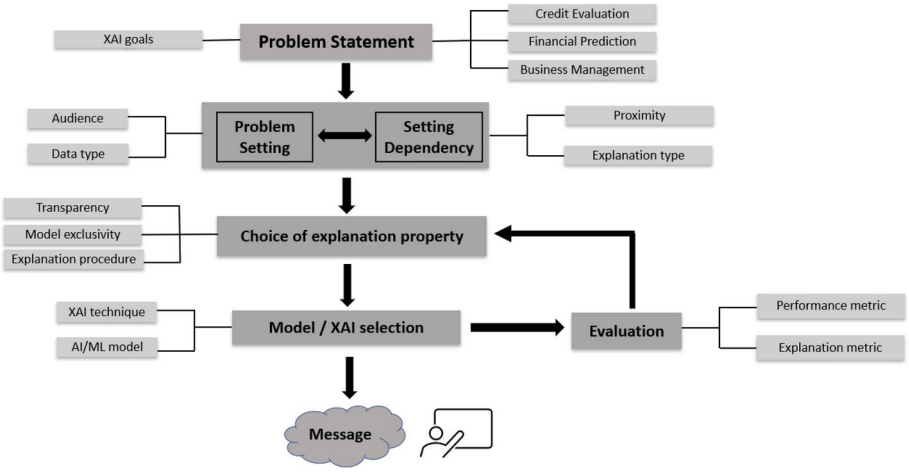

Fig. 5 XAI framework depicting a sequential flow of decision-making events. The proximity (local/global) and explanation type (factual/counterfactual) should be chosen in accordance with the target audience and data type available. The choice of explanation property is assessed by an iterative evaluation under the appropriate metric for both performance and explanation conveyed

3.1 Visual explanation

Visual explanatory (VE) techniques generate explanations of the underlying model in the form of visuals. VE techniques can be both model-specific and model-agnostic. The model-specific techniques reviewed are mainly constructed to interpret image-based networks such as convolutional neural networks (CNN). Kumar et al. (2017) propose to perform a deconvolution on the last layer preceding the output to extract a visual attentive map. The approach named CLass Enhanced Attentive Response (CLEAR) generates a graphical plot denoting the timeframe to which the stock-picking agent pays the most attention, along with a separate plot corresponding to the sentiment class of the stock. Chen et al. (2020) implement a CNN network to identify 8 common candlestick patterns which are widely used for technical analysis in stock market trading. The authors then perform an adversarial attack on regions of the feature space, to demonstrate that the model is focusing on regions similar to how a human would process the candlesticks. Shi et al. (2021) employ a reinforcement learning (RL) agent to optimize a portfolio of equities while using a temporal CNN as a feature extractor. The dynamic asset allocation is interpreted with Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al. 2017), improving over simple deconvolutions by producing class-discriminative explanations, and is applicable to any deep neural networks.

The proposed technique outputs a localization map using gradients corresponding to the target label. The right plot in Fig. 6 depicts a global map highlighting each asset's importance across the trading period. Interestingly in the left plot, the agent focuses on the worst-performing stock, GLID the most, rather than the high-performing stocks. Here, the agent predicts the stock decline and reduces the allocation proportion, and indirectly increases the weights of high-performing stocks which in this case is the target stock, NVDA. Achituve et al. (2019) propose to use an attention mechanism (Vaswani et al. 2017) to compute similarity scores of possibly fraudulent transactions on both feature and temporal levels and in return, allows for visualization at the top contributing features accounting for the model's prediction.

Model-agnostic VE techniques can be integrated with any form of model architecture and bear a similar resemblance with feature relevance techniques. Both investigate the effects on the model's output by adjusting the input features. Zijiao et al. (2022); Biecek

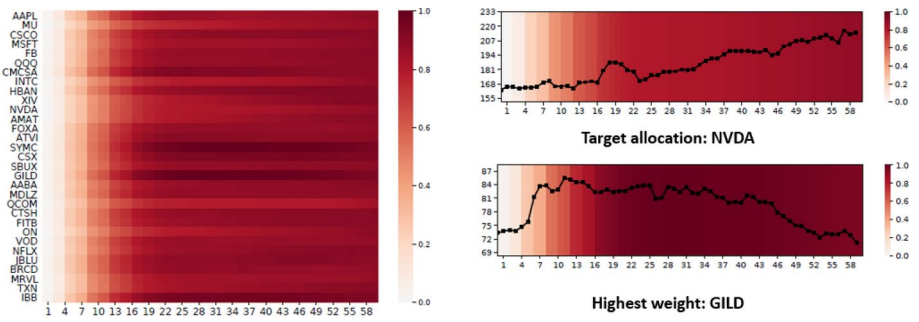

Fig. 6 [left] shows a heatmap denoting the global attentiveness of individual stocks in the overall portfolio. [right] correspondingly presents a heatmap of individual assets. The agent chooses to allocate the most weight of the portfolio to NVDA, while surprisingly focusing most on a declining stock, GILD. The agent reduces the weightage of GILD and allocates to NVDA. Adapted from Shi et al. (2021)

et al. (2021); Zhang et al. (2022); Farzad (2019) employ Partial Dependence Plots (PDP) to visualize the marginal effects of features relating to corporate distress, credit scoring, and detecting mortgage loans defaults. The generated plots can enable a way of inferring if the underlying input–output relationship is linear or complex. However, PDP has often been criticized for its assumption of independence between features, evaluating unrealistic inputs, and also conceals any heterogeneous effects of the input features. Accumulated Local Effects (ALE) (Apley and Zhu 2020) address the concerns of feature correlation by considering the conditional distribution rather than the marginal one. In particular, it accumulates differences between intervals within the feature set to account for individual feature effects. Crosato et al. (2021) employ ALE on top of a tree ensemble model, XGBoost (Chen and Guestrin 2016), as well as with global Shapley values (Shapley et al. 1953) for better scrutability. This work deduces that the increase in profit margin and solvency ratio leads to lower debt default rates of small enterprises.

Zhang et al. (2022) evaluate across an arsenal of XAI techniques, encompassing the aforementioned, and also include Individual Conditional Expectation (ICE) for financial auditing purposes. ICE differs subtly from PDP in that it considers instance-based effects rather than averaging across all instances, making it a local approach (see Fig. 7). Zhang et al. (2022a) generate counterfactual explanations on credit loan applications by coupling unsupervised VAE with a supervised probit regression. The combined model yields a discriminative latent state, corresponding to class labels of either delinquency or non-delinquency. The counterfactual is subsequently produced by a stepwise manipulation function towards the opposite class label. The authors evaluate the generated counterfactuals quantitatively using maximum mean discrepancy (MMD) (Zhang et al. 2022b), which measures the number of successfully flipped class labels as well as minimal feature changes.

3.2 Explanation by simplification

The idea of Explanation by Simplification (ES) techniques is to introduce a surrogate model performing uncomplicated operations. The purpose is to allow the machine learning developer to formulate a mental model of the AI model's behavior. The surrogate model has to be interpretable and more importantly capture the performance of the black-box model with high fidelity. The latter property should be given a higher priority since there is little use for interpreting a low-fidelity solution. ML techniques which apply linear operations and rule

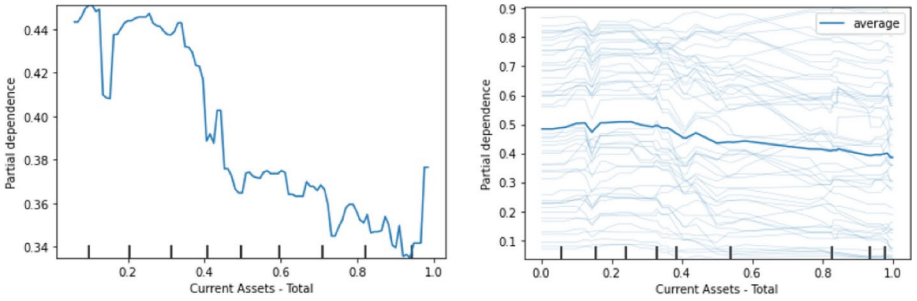

Fig. 7 [left] shows a PDP on averaged marginal effects of total assets on the probability of statement restatement and [right] displays ICE, which considers instance-level relationship. Both show a negative relationship (Zhang et al. 2022)

extraction are applicable as surrogate models in place of uninterpretable neural networks. These include decisions tree (DT) with limited depth, linear/logistic regression, K-Nearest Neighbors (KNN), and generalized linear models (GLM).

Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al. 2016): is perhaps one of the most popular explanation techniques across various use cases, including finance. LIME is a model-agnostic method that is used to provide insight as to why a certain prediction was made and can be constituted as an outcome explanation technique. Since LIME is a local-based technique, it only has to approximate the data points within a defined neighborhood, achieving a much more realistic goal instead of capturing an interpretable representation of the entire dataset. On a high level, LIME can be implemented as follows (see Fig 8):

1. The target instance to be explained is denoted as $x \in \mathbb{R}^d$. Uniformly sample *n* random subsets of nonzero elements of *x* to form local training points, $z \in z_1, z_2, ..., z_n$, where $z_i \in \mathbb{R}^d$ for $1 \leq i \leq n$.
2. Derive labels $f(z_i)$ for each point using the black-box model *f*. The surrogate model, *g* is then trained on the derived dataset, $\{z, f(z)\} \in Z^n$.
3. Choose a transparent surrogate model, *g* and train it on the dataset, Z^n via Eq. 1.
4. Interpret the outputs of the transparent model on the target instance, *g*(*x*).

LIME minimizes the following loss function to optimize for both fidelity of the local model as well as minimal complexity.

$$\text{explanation}(x) = \underset{g \in G}{\text{argmin}}[\, L(f, g, \pi_x) + \Omega(g)] \tag{1}$$

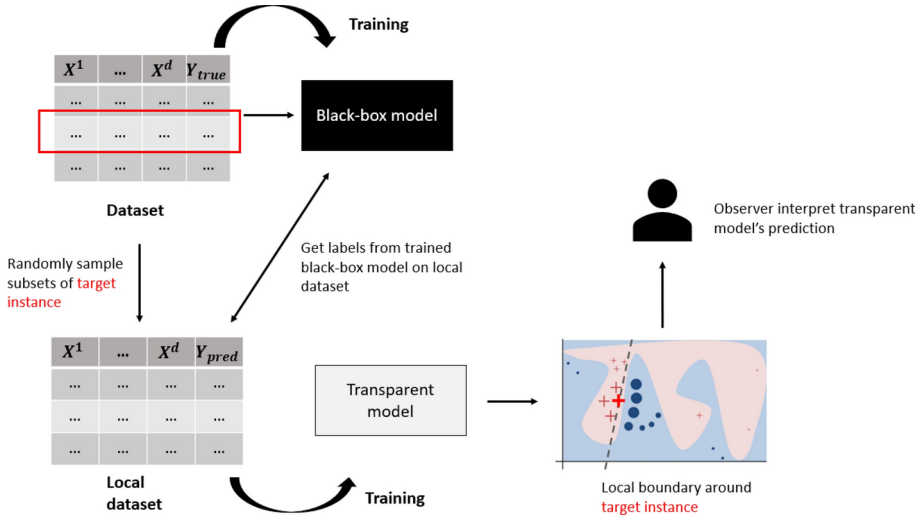

Fig. 8 LIME process: Predictions of black-box model are uninterpretable. The local instance in the red box is the target to be explained. Subsets of nonzero elements of the target instance are uniformly drawn to form a local dataset on which the surrogate transparent model is trained on. The prediction from the linear transparent model can then be interpreted by the user (Ribeiro et al. 2016)

L represents the loss function of the surrogate model g on the labels f, weighted by proximity π_x. Ω represents the complexity or number of features in the surrogate model. G is the set of all locally fitted models, where each explanation is produced by an individual local model. The authors additionally propose a sparse selection of features, named Submodular Pick LIME (SP-LIME), to present the observer with a global view, based on an allocated budget of maximal features to focus on. The method delivers diverse representation by omitting redundancy. Misheva et al. (2021); Serengil et al. (2022) use LIME on top of tree ensembles to identify the contributions of individual features pushing towards predicting a specific borrower as defaulting or successfully paying off the loan. Such explanations can be useful in preventing social bias by discovering any socially discriminative features on which the model may be focused, thereby instilling trust in the model's usability.

Yan et al. (2019) extend LIME towards financial regulators requiring commercial banks to adhere to a set of financial factors, where they propose a method named LIMER (R stands for Regtech). The authors of LIMER argue that high acceptance of financial solutions can be achieved if such factors are integrated into the explainability design of the AI model. Collaris et al. (2018) implement model simplification by extracting logical rules from a random forest and select the top most relevant rules. The decision rules are extracted from a local dataset, derived similarly to LIME without weighting the proximity of each drawn sample. Maree and Omlin (2022b) train a recurrent neural network (RNN) to classify customer spending into five categories. An interpretable linear regression model was subsequently trained to predict the nodes formed by the RNN model. The authors then perform inverse regression which provides a mapping from output space to state space where the features responsible for categorizing customer spending can be identified.

3.3 Feature relevance

Feature relevance (FR) techniques account for the majority of the proposed explanation methodology we reviewed. FR techniques revolve around computing a relevance score for each feature, highlighting the respective contribution of the target feature either at a global or local scale. Rawal et al. (2023) aims to provide a novel perspective on causal explainability, creating a model which extracts quantitative causal knowledge and relationships from observational data via Average treatment effect (ATE) estimation to generate robust explanations through comparison and validation of the ranked causally relevant features with results from correlation-based feature relevance explanations. *SHapley Additive exPlanations (SHAP)* (Lundberg and Lee 2017), motivated by the fair distribution among players from game theory (Shapley et al. 1953) is a highly popular FR technique, which seeks to estimate the fair value of each feature in contributing towards the outcome, $f(x)$. The fair value, otherwise known as shapley values are determined, based on estimating the difference between the black-box function over feature subset S with and without the target feature, $f(x_{S\cup\{i\}})$ and $f(x_S)$ respectively. The difference is then averaged across all possible coalitions within the feature set F.

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{2}$$

The final outcome is intuitively derived as an aggregate over all non-zero shapely values. SHAP's popularity stems from three attractive properties: guaranteeing a complete approximation of the original model $f(x)$ through additive feature attribution (see Eq. 2), ensuring non-contributing features have no impact on model output and consistency of feature values tracking the outcome contribution. We notice a large subset of papers reviewed has utilized SHAP as an explanation approach, likely given its flexibility towards explaining the model at both local and global scales (see Fig. 9). Dikmen and Burns (2022) incorporate SHAP with additional credit knowledge for the layperson to assess the logic of XGBoost's decision in a peer-to-peer lending scenario. Müller et al. (2022) introduce RESHAPE, designed for unsupervised deep learning networks, which provide explanations at the attribute level. Such explanations can assist auditors in understanding why an accounting statement is flagged as anomalous. The authors evaluated RESHAPE against other variants of SHAP, based on metrics measuring fidelity, stability, and robustness.

Attributing to the recent frenzy in cryptocurrency which has led to a number of studies attempting to predict movements in the cryptocurrency market, Fior et al. (2022) propose an interactive dashboard providing multiple graphical tools using SHAP for financial experts. Babaei et al. (2022) apply SHAP to explain predictions, generated by the popular mean-variance Markowitz model (Markowitz 1952) which is an optimization model for establishing the optimal portfolio balancing between returns and risk. The generated explanation provides regulators a means of asserting compliance of algorithmic automized traders, otherwise known as robot-advisors, with established rules and regulations. Demajo et al. (2020) incorporate Global Interpretation via Recursive Partitioning (GIRP) with SHAP as a global interpretability technique. GIRP uses the importance values generated by SHAP to further extract meaning insights from tree models, and the method is compared against a boolean rule technique in a credit scoring use case. Bussmann et al. (2021) construct a tree-like visual explanation with TreeSHAP (Lundberg et al. 2018), specifically designed for ensemble trees with an improvement in computational efficiency. The produced structure allows users to visualize clusters of similar outcomes describing company default risk. Yasodhara et al. (2021) compare TreeSHAP against impurity metrics using information gain, on ensemble tree models for investment quality prediction.

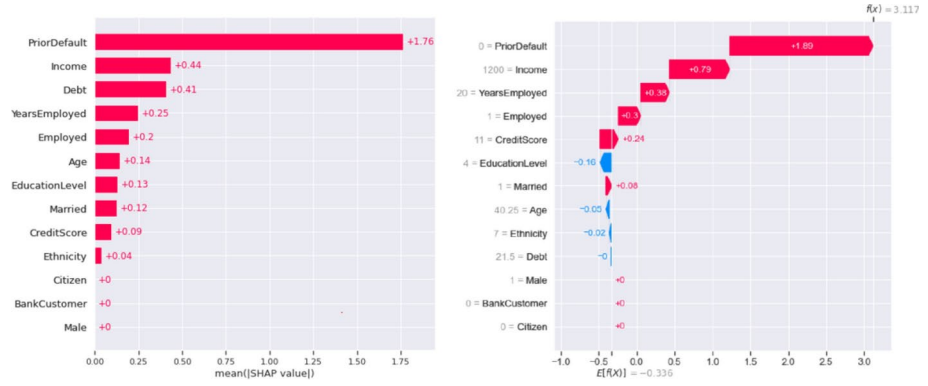

Fig. 9 [left] Values of feature importance at a global level for the ML model's decision in credit card approval. [right] An example of instance-level, $E(f(X))$ represents the model's base prediction if no features were considered, and $f(x)$ represents the final prediction after summing the contributing features (ϕ_i) (Rizinski et al. 2022)

Gramegna and Giudici (2020) identify relevant features leading to consumers' decision on purchasing insurance and further clusters them into least to most likely groups with Shapley values. Bussmann et al. (2020) similarly implement SHAP to explain XGBoost's classification of credit risk, while comparing it against an interpretable logistic regression model. Other studies include discovering the relationship between corporate social responsibility and financial performance (Lachuer and Jabeur 2022), customer satisfaction (Rallis et al. 2022), GDP growth rates (Park and Yang 2022), stock trading (Benhamou et al. 2021; Kumar et al. 2022), financial distress (Tran et al. 2022), market volatility forecast (Weng et al. 2022) and credit evaluation (Rizinski et al. 2022; Bueff et al. 2022; Fritz-Morgenthal et al. 2022).

Wand et al. (2022) perform K-means clustering on historical S&P 500 stock information to identify dominant sector correlations that describe the state of the market. This work applies Layer-wise Relevance Propagation (LRP) (Bach et al. 2015), after transforming the clustering classifier into a neural network since LRP is designed to work specifically with neural network architectures. Carta et al. (2022) prune unimportant technical indicators using different configurations of a permutation importance technique, before implementing decision tree techniques for stock market forecasting. The proposed technique was compared with LIME and demonstrated better reliability. Bracke et al. (2019) introduce a set of feature relevance techniques, Quantitative Input Influence (QII) (Datta et al. 2016) to compute interaction effects between influential features and additionally for each multi-class label. The authors additionally evaluated the ability of the XAI technique with five questions relating to each individual audience class. All of the XAI methods shown thus far are implemented in the post-modeling stage, while the work of Islam et al. (2019) is an example pertaining to pre-modeling where the identification of relevant features takes place before constructing the black-box model. This work explores the set of features relating to mortgage bankruptcy and performs feature mapping against a set of widely-used credit concepts. The utility of such an approach is confirmed through empirical evaluations.

As pointed out before, contrastive explanations are usually preferred. End-users subjected to an unfavorable AI model's decision would prefer a solution to the problem rather than a fact-based explanation which may present multiple possible reasons, giving little use to the explanation receiver. An explanation providing changes to be made such that the outcome can be reversed towards the favorable is referred to as a counterfactual explanation. Counterfactuals are derived by computing small changes to the input features continuously until the outcome is altered to the target class. Cho and Shin (2023) first identify significant features, attributing to bankruptcy through SHAP, and subsequently generates an optimal set of counterfactuals using Genetic Algorithm (GA). The loss optimized by GA composes of objectives describing desirable properties of a good counterfactual outcome, including minimizing the size of altered features and maximizing the feasibility of the outcome. Grath et al. (2018) additionally provide positive counterfactual explanations, describing the required changes to the current inputs that would instead reverse the loan approval to rejection. Such explanations can provide some form of safety margin for the user to be mindful of. Vivek et al. (2022) used various technique from DiCE De Bruin et al. (2009) to generate counterfactuals under five different experimental conditions. The experiment aims to identify and study the effects of the causal variables in the fraud detection of ATM transactions.

3.4 Explanation by example

Apart from techniques that identify feature relevance on varying scales or approximate with a surrogate model, another form of explanation exists by selecting representative samples to illustrate the model's behavior. Such techniques can be classified as Explanation by Example (EE). One such technique includes prototype-based explanations. Prototypes can be regarded as representatives of the entire dataset, chosen based on similarity and importance in the overall decision-making of the model. Demajo et al. (2020) implement protodash (Gurumoorthy et al. 2019), a gradient-based algorithm in a credit loan application to select top m prototypes, of which the top two are selected, with m being 6. The resulting outcome is a number of representative prototypes and each instance can be represented by either generated prototype in the clusters. In this case, the proportions of allocated instances were balanced between both prototypes. The number of prototypes is a hyperparameter to be fine-tuned. A higher value of m is frequently used where the complexity of the problem is a concern, albeit raises the risk of overfitting, while a lower value is used in simpler scenarios but incurs the risk of underfitting.

For example, in the credit loan dataset, the selection of two prototypes was considered too little by domain experts who instead prefer 3-4 as being sufficiently representative of the evaluated dataset. Davis et al. (2022) similarly extract representative instances using KNN and generates insights on out-of-sample instances by looking for similarities with the representative points. Additionally, the data points for computing the distance are instead replaced with Shapley values, taking into account the importance of input features. Representative samples are generally suitable if the user is interested in determining the types of patterns or behavior found in the dataset while being relatively fast and straightforward to implement.

3.5 Summary of numerical features

The above-mentioned approaches should be chosen according to the task at hand and the target audience. VE techniques, such as deconvolution and Grad-CAM, are less commonly used in the financial industry due to their limited applicability to networks other than CNNs. However, ALE, PDP, and ICE can be suitable approaches for financial analysts who might want to study the relationship between individual features and the model's outcome. ES is a straightforward approach that delegates the interpretability problem to a less complex surrogate model, though it incurs the additional cost of ensuring the faithfulness of the surrogate model. FR techniques allow users to observe each feature's contribution to the black-box prediction. Both global and local explanations serve different purposes for individual audiences. However, in situations where each feature equally contributes to the model outcome, such explanations might not be very helpful depending on the objective of the explanation. For example, a declined credit loan approval may have multiple features such as prior default, debt-to-income, and household capital contributing equivalently to the outcome. Such an explanation does not offer an obvious course of action for the applicant. EE techniques are particularly useful when the user wants a small set of representative samples to explain the model's outcome. This can provide a fast and straightforward explanation, but it has limited usefulness.

4 XAI for textual information

In this section, we review models operating with textual information. We note that the papers pertaining to this area represent the minority in the overall literature. When it comes to data preparation, textual data generally require additional pre-processing works such as stop word removal, stemming, lemmatizing, and tokenization. In terms of feature extraction, the semantics, and syntax structure around text is important and have to be learned to fully capture the information conveyed, unlike numerical features which are readily usable. Models suitable for training from textual data are also limited to a smaller subset of available techniques. Nevertheless, unstructured data such as text are in abundance. If properly processed, textual data can be used to derive informative signals such as market sentiment and emerging trends (Ma et al. 2023). Textual data are commonly classified under alternative information, which comes in a wide variety of sources including social media, online reviews, blog posts, and news headlines (Kolanovic and Krishnamachari 2017), in contrast to non-alternative information which refers to data commonly utilized for financial analysis. Fortunately, a wide variety of explanation techniques exist which are compatible with textual information. Conveniently, textual information is applicable for XAI techniques delivering explanation via text generation, which can be preferable for the layperson as natural language provides an easier form of interpretation as compared to statistical graphs.

4.1 Text explanation

Text explanation techniques provide clarity in the form of generating informative textual statements to assist in the understanding of the model's behavior. The generated text can either be re-generated text statements by using some form of generative model or replacing selected words in the original sentence. Srinivasan et al. (2019) utilize Generative Adversarial Networks (GAN) to produce text statements that seek to align with user-defined inputs. Specifically, the explanation can take two different objectives, either converting actionable text to educational text or vice versa. The actionable text briefs audiences on optimal actions to consider, based on real-world responses from human responses on multiple loan application scenarios while the latter seeks to educate the audience on reasons attributing to the response. The transfer of objectives from one to another is analogous to the implementation of style transfer on images, a popular application of GANs that translates the style of an instance to the target image while retaining the content. Figure 10 shows a snippet of an example, the proposed method can identify the semantics behind the statement and relay the relationship between consistency and time, while knowing if the current income is above or below the required threshold. Wang et al. (2023) presents a novel and straightforward method for generating high-quality text embedding using synthetic data with fewer than 1000 training steps. This approach contrasts sharply with existing methods, which typically require multi-stage intermediate pretraining involving billions of weakly-supervised text pairs, followed by fine-tuning with a handful of labeled datasets. Notably, their method eliminates the need for constructing complex training pipelines or depending on manually collected datasets, which are often limited by task diversity and language coverage.

Yang et al. (2020) generate plausible counterfactual text sentences with a transformer architecture, trained under contextual decomposition. The explanation technique, derived from Sampling and Contextual Decomposition (SCD) (Jin et al. 2019), performs different

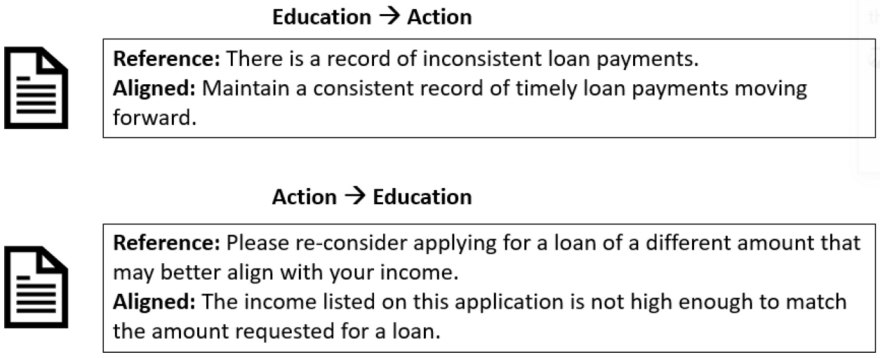

Fig. 10 [Top] Transfer of educating statement to actionable statements advising applicant on actions to take such that the subsequent loan application can be approved. [Bottom] Transfer of original statement highlighting actions to educating statement conveying the reason for rejected loan application (Srinivasan et al. 2019)

actions including inserting, removing, or replacing words that are representative of the context in the statement, based on the target objective. The high-level idea of counterfactual generation involves identifying the most relevant word and replacing it with an antonym from a reference dictionary and continues until the outcome is reversed. The proposed transformer outperforms even human experts in classifying financial articles on merger & acquisition event outcomes. Yuan and Zhang (2020) generate text explanations using a state-of-the-art natural language generation Transformer decoder, GPT-2 (Radford et al. 2019), while fulfilling soft constraints of including keywords. The proposed technique, soft-constrained dynamic beam allocation (SC-DBA) extracts keywords corresponding to various levels of predicted market volatility using a separate network on harvested news titles. The quantitative measurement is evaluated based on the fluency and utility of the explanation produced.

Koa et al. (2024) propose to perform financial adaptation on a much larger 13 billion parameter-sized LLM. The training includes both supervised fine-tuning (SFT) and PPO, an RL-based technique to align the model toward forecasting stock price movements and generating plausible textual explanations. The proposed framework consists of three stages: summarize, explain, and predict, (SEP). Through an extensive evaluation, the approach outperforms both DL and LLM models. The RL training was also shown to be useful in refining the explanations toward providing truthful rationales behind the outputted decision.

Du et al. (2024) proposed a contrastive learning framework to learn the nuanced differences between positive and negative samples for the stock price movement prediction of a target stock. The framework selects both positive and negative samples from historical data that present similar trends to the target stock over a seven-day trading period. The key difference is that the positive sample has the same future price movement direction as the target stock, while the negative sample diverges in the future movement direction. The authors also integrated attention mechanisms to highlight the differences in textual and numerical features between positive and negative samples. This approach allows domain experts and end users to better evaluate the prediction by comparing stocks with similar historical trends but different future movements. The attention mechanism further explains the feature relevance between the positive and negative samples and the target stock.

4.2 Visual explanation for text

Besides interpreting through text, users can understand through the form of visuals, which makes the use of attention a particularly attractive option. Attention was first introduced when it was used to consider correlations between words in a sentence in a parallel fashion and is a primary component in the Transformer architecture (Vaswani et al. 2017). Transformers are notably suitable for processing long sequences of text and through the use of attention. They are computationally efficient compared to RNN-based models. It so happens that, computing attention scores of each word serves as a natural form of interpretation, by allowing users to visualize how the network is capturing information from the input text (Han et al. 2022). Representative works in this area employ attention to highlight regions of text sentences that are deemed relevant for the output. Yang et al. (2018) utilize dual-level attention with Gated-Recurrent Units (GRU) (Chung et al. 2014), processing both inter-day and intra-day embedding of news titles relating to S&P 500 companies. The attention module assigns a relevance score to each news article and the authors additionally construct a knowledge graph conducting concept mapping between relevant entities as a visual explanation.

Corresponding to dual-level attention, Luo et al. (2018); Lin et al. (2021) propose a hierarchical attention model at both the word and sentence level and produced explanations in the form of a heatmap, highlighting relevant text. The proposed method, FISHQA was trained to detect loan arrears from financial text statements, similar to the compared baselines. The uniqueness of the proposed method lies in providing FISHQA with additional user queries. The model was able to highlight regions of the statement corresponding to the set of expert-defined concepts. This form of explanation allows users to verify if the model is focusing on the correct terms relating to the concept at hand (refer to Fig. 11).

Along the lines of hierarchical attention, Lin et al. (2021) introduce a quantitative measure to evaluate the precision and recall of captured against various lexicon dictionaries and expert annotated lists. The approach, analogous to the former study can be seen as an extrinsic process of ensuring the correctness of concept identification, by capturing words associated with financial risk. Deng et al. (2019) implement knowledge graphs to provide a visual linkage between event entities extracted from stock news articles. The approach offers users a visual understanding between the feature's relationship and the corresponding prediction.

Ito et al. (2020) introduce GINN, an interpretable neural network. The network is designed in a way that each layer represents different entities such as words and concepts at the node level. The approach identifies words attributing to the predicted sentiment labels, as well as the concepts it belong to.

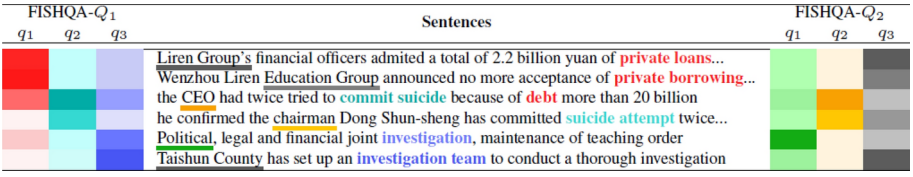

Fig. 11 FISQHA: hierarchical attention model, different colors relating to different financial concepts, grey - company, light brown - executives, red - financing, blue - litigation, teal - personnel (Luo et al. 2018)

4.3 Summary of textual information

TE techniques aim to augment existing input text or generate new text based on given inputs. Such explanations are commonly preferred since natural language is easily understood by humans if the explanation is concise and accurate. However, textual explanations may fail to capture the nuanced relationships between input features and model decisions. This can be especially detrimental and counterproductive to domain experts whose goal is to discover further improvements based on the provided explanations. Textual explanations might also require additional processing work to ensure fluency, coherence, and unambiguity. VE techniques in these works leverage the utility of attention to provide a glimpse into how the model is representing the input text, and the use of hierarchical attention allows for a more refined analysis. However, since attention captures the relationship between each word or sentence, such explanations might be overwhelming if the set of explainable features is too large. Audiences who are not well-versed in heatmaps or attention scores may have difficulty understanding the provided visuals.

5 XAI for hybrid information

The remaining studies implementing post-hoc explanation techniques utilize a combination of both textual and numerical/technical features. With respect to instance-level explanations, Bandi et al. (2021) combine the sentimental analysis of text with technical analysis of historical stock prices to train a random forest stock forecasting model, explained through LIME. The resulting explanation computes a set of relevant feature values and news wording corresponding to the respective outcome. Similarly, Gite et al. (2021) implement LIME with LSTM-CNN and accurately identify attentive words in consonant with the target sentiment. Liu et al. (2020) predict the possibility of litigation on financial firms from examining 10-K financial reports and numerical indicators concerning the firm's accounting knowledge. The authors additionally carry out an ablation study on the utility of hybrid information as opposed to individual and validated the initial approach. Correspondingly, the explanation served to regulators is framed as the identification of text leading to the suspicion of insider trading, with the help of an attention mechanism. Zhang et al. (2020) adopt the practice of shapley values and further integrate external knowledge regarding truth factors, namely Truth Default Theory (TDT) (Levine 2014) to detect information fraud. The explanation module incorporates both shapley values and TDT to generate a report highlighting numerical contributions of features as well as a *text explanation*.

A union of *explanation by simplification* and *feature relevance* was proposed by Cong et al. (2021); Ghosh and Sanyal (2021). Ghosh and Sanyal (2021) implement both LIME and SHAP, offering a global and local explanation of market fear prediction in the Indian financial market. Maree et al. (2020) use SHAP and identify textual information to be more important for classifying financial transactions and further perform clustering to identify top contributing keywords. Cong et al. (2021) interpret an RL-trained agent's behavior in algorithmic trading. The resulting explanation enables experts to focus on time-dependent variables alongside consideration of non-linearity effects, which are reduced to a small subset of initial variables. The learned policy is simplified via policy distillation, onto the space of linear regressions such that an interpretable Lasso regression model can be used as an inter-

pretable approximation. Subsequently, *k*-degree polynomial analysis is conducted to select salient features, with *k* acting as an additional flexibility for the developer to decide. Ong et al. (2023) utilize aspect-based sentiment analysis to study the relationship between stock price movement and top relevant aspects detected in tweets. The polarity of each aspect is derived from a SenticNet-based graph convolutional network (GCN) (Liang et al. 2022). The proposed method can be seen as analogous to the feature relevance technique, aimed at deriving top contributing aspects with polarity values. The proposed work focuses on the relationship between financial variables instead of making financial predictions. Such information can allow for further analysis, leveraging the relationship between the price movement of individual stocks and individual sentiment of popular terms detected in tweets.

Hybrid information combines the utility of both numerical and textual information, which can lead to better performance and an increase in the number of compatible explanation techniques. For example, *text generation* techniques can be used to generate natural language explanations for non-technical audiences, facilitating ease of understanding, while feature relevance approaches can be utilized to identify top contributing factors in the feature domain for technical experts. Models working with both numerical and textual information can also benefit from a performance point of view if such information can be processed without the risk of overfitting.

However, it may be difficult for models to seamlessly perform with hybrid information, as it ultimately depends on the task at hand and may require complex feature engineering. For instance, the utility of text information largely depends on the source and often requires a significant amount of preprocessing before the data can be useful. The combination of both text and numerical features may increase the complexity of the explanation and end up being counterproductive. Such issues limit the inclusion of textual information in use cases such as stock trading or market index predictions. Nonetheless, we note that leveraging hybrid information to provide explanations can be a promising approach if the aforementioned issues are addressed.

6 Explainability in transparent models

The remaining studies look at instigating explainability from inherently transparent models. These models are typically restricted to ML models performing linear operations or rule extraction. The medium of explanation in transparent models is by nature model-specific, in the sense that the same mode of explaining how a model functions to an audience likely cannot be reused by a different model. The usability of transparent models in complicated tasks is largely restricted due to their poor predictive strength. Nevertheless, transparent models still remain an attractive option if sufficient performance can be guaranteed.

Linear/Logistic Regression: Linear regression model is among the earliest ML models to be used for quantitative analysis. The prediction outcome can be easily derived as a weighted aggregate of input features. As such, the outcome can naturally be interpreted by inferring from the coefficients, W_i which serves as a quantitative measure of feature importance for the outcome. Attributing to the linearity assumption, the output y can be derived as such:

$$y = W_0 + W_1 x_1 + W_2 x_2 + ...W_n x_n + \epsilon \tag{3}$$

One can easily interpret the outcome as *"By increasing feature x_i by one unit, the output increases by W_i"*. On the other hand, the logistic regression model is interpreted in a slightly different manner, since the output is bounded between [0,1], a logistic function is used. Logistic regression looks at the probability ratio between both outputs: *"Increasing one unit of x_i is equivalent to increasing $\frac{P(y=1)}{P(y=0)}$ by $exp(W_i)$"* (Molnar 2020). Dumitrescu et al. (2022) address the trade-off between accuracy and interpretability by incorporating decision trees with logistic regression acting as the main operational backbone. The technique is coined as Penalised Logistic Tree Regression (PLTR). PLTR extracts binary variables from short-depth DTs, all the while establishing sparsity through a penalized lasso operation. The proposed model is able to account for non-linear effects in a credit-scoring dataset while retaining interpretability by observing the top-selected rules.

Decision Trees: Decision trees are one of the most commonly used techniques in machine learning problems due to their simplicity and easily understandable structure. Unlike linear/logistic regression, DT can approximate nonlinear relationships and yet remain interpretable via simple *if-else* logic. However, the transparency of tree models diminishes with increasing depth, and popular ensemble tree models such as XGBoost or gradient-boosting tree models completely eliminate any form of interpretability. The user can interpret decision trees by traversing through the root node and upon arrival at each leaf node. The outcome can simply be explained as *"if* x_1 *is* $> / <$ threshold$_1$ *AND* x_2 *is* $> / <$ threshold$_2$, $\cdots$ *, outputs Y*". Gramespacher and Posth (2021) employ a single DT and frames the loan approval task as one which maximizes profit for the lender firm. Carta et al. (2021) build a lexicon dictionary associated with stock price variation, extracted from a dataset comprising both news and historical stock prices. The combined effort provides users with two forms of explanation, observed in a sequential rule-based manner as well as words correlated with the predicted market direction.

Others: Adams and Hagras (2020) construct an interactive platform, Temenos XAI using fuzzy logic to make financial predictions. The authors demonstrated the efficacy and explainability in various downstream banking and trading scenarios. The usage of fuzzy-logic accounts for uncertainty, which is prevalent in the financial environment, and is especially useful for modeling imprecise information. The platform allows users to interpret the model on a global scale as well as at an instance-level, via observing the top contributing rules. Chen and Ye (2022) build on top of neural additive models (NAM) (Agarwal et al. 2021) and introduces a generalized form of NAM, GGNAMS which focuses on sparse nonlinear interactions. GGNAMS can be regarded as an intermediate between fully connected networks and logistic/linear regression with the intent being to retain linearity and minimize excessive interactions among features while maximizing accuracy. The additive components can then be interpreted similarly to LR.

Nazemi et al. (2022) similarly implement NAMs and Explainable Boosting Machine (Nori et al. 2019) to identify financial drivers leading to creditor recovery rates. Dumitrescu et al. (2022) propose a hybrid approach of combining decision trees with logistic regression, capturing nonlinear effects while retaining the transparency of the model's behavior. Sudjianto and Zhang (2021) advocate for designing inherently transparent models in the pre-modeling/modeling stages and suggested a qualitative template, describing properties of model interpretability. The intention of the template is to allow researchers to ensure model interpretability while designing the model architecture. As a proof of concept, this work designs an interpretable ReLU network while conforming to the proposed template, and

evaluates the network in a credit default classification task. The resulting network can be disentangled into a set of local linear models whose inherent transparency can be visualized by observing the local coefficients.

Transparent models have the advantage of being interpretable without requiring additional approaches to interpret the model or outcome. However, there exists a clear trade-off between desired performance and sufficient interpretability. Certain works have introduced approaches that combine different transparent models to achieve better performance while still retaining as much model transparency as possible. Transparent models remain a popular choice in the financial domain, as companies must undergo routine audits that require audited firms to provide accountability for their algorithmic services offered to end-users. Nonetheless, given the monotonic success of deep learning models, companies seeking to maintain their competitive edge must either improve on the existing transparent models or balance the performance-interpretability trade-off. A recommended approach would be to stick with transparent models if their performance proves sufficient and proceed with less interpretable models otherwise. One could also break down the task in a hierarchical manner, using interpretable models for lower-level tasks and better-performing models for more complicated tasks.

7 FinXAI and ethical goals

In Sects. 3-6, we have reviewed different FinXAI methods and summarized their technical strengths and weaknesses, based on representative papers over the past years. In this section, we analyze the contributions of these FinXAI techniques to the ethical goals that were set out in Sect. 2.3. We also discuss some of the goals lacking sufficient study in current FinXAI techniques. In this section, the goal of accessibility also encompasses the fact that developers and domain experts can easily access the decision-making mechanisms of complicated black-box models due to improved interpretability. This is slightly different from the narrative that we introduced in Sect. 2.3, which mainly focuses on accessibility for non-expert users. Such an extension can better explain the technical contributions of the reviewed works to XAI. As seen in Table 4, different explainable methods contribute to the ethical goals from different aspects. XAI for numerical features has proposed several methods to approach ethical goals, regarding trustworthiness, fairness, informativeness, accessibility, confidence, and causality.

For VE methods, Zhang et al. (2022a) is one example that advocates for trustworthiness by generating counterfactual explanations, while being informative by detecting important features that can alter the prediction. Counterfactual explanations reveal the slightest modifications that are necessary on the input data to achieve an alternative outcome. It helps to earn trustworthiness from target audiences because counterfactual explanations provide possible rescue measures for them to achieve their targets with minimum effort, e.g., proposing possible improvements to help borrowers pass the qualification review of credit agencies. Counterfactual explanations can also help to justify predictions besides factual explanations. Both merits of generating counterfactual explanations improve the trustworthiness of audiences. Many VE-based approaches improve informativeness by gaining insights into the decision-making mechanisms of models and revealing feature correlations (Achituve et al. 2019; Biecek et al. 2021; Chen et al. 2020; Crosato et al. 2021; Farzad 2019; Kumar et

al. 2017; Shi et al. 2021; Zhang et al. 2022; Zijiao et al. 2022), because visualization takes the advantages of demonstrating patterns and trends of data, e.g., model parameters and numerical features. VE can be also used to discover valuable features (Achituve et al. 2019; Crosato et al. 2021; Zhang et al. 2022). It is easy to communicate with both experts and non-domain experts by using graphical representations or visual images whenever available. Thus, VE also improves accessibility for broader audiences.

For ES methods, Collaris et al. (2018); Maree and Omlin (2022b); Yan et al. (2019) can explain why a certain prediction was made from outputs, which helps to improve the trustworthiness of AI predictions. Misheva et al. (2021); Serengil et al. (2022) use LIME to detect socially discriminative features to prevent social bias. ES is the only approach that was used for improving fairness in finance.

For FR methods, Babaei et al. (2022) use SHAP to improve the trustworthiness of algorithmic traders in crypto markets. Cho and Shin (2023); Grath et al. (2018) generate contrastive explanations to explain required changes for certain predictions. Bussmann et al. (2021) visualizes similar outcomes that describe the risk of a company default with SHAP, while most SHAP-based methods (Babaei et al. 2022; Benhamou et al. 2021; Bracke et al. 2019; Bueff et al. 2022; Bussmann et al. 2020; Carta et al. 2022; Demajo et al. 2020; Dikmen and Burns 2022; Fior et al. 2022; Fritz-Morgenthal et al. 2022; Gramegna and Giudici 2020; Islam et al. 2019; Kumar et al. 2022; Lachuer and Jabeur 2022; Park and Yang 2022; Müller et al. 2022; Rizinski et al. 2022; Tran et al. 2022; Vivek et al. 2022; Wand et al. 2022; Weng et al. 2022; Yasodhara et al. 2021) improve accessibility for technical audiences by discovering important features. Fior et al. (2022) improve usability by constructing interactive graphical tools upon SHAP, which likewise promotes accessibility. Vivek et al. (2022) is one of the rare works that study causal inference based on generated counterfactuals.

For EE methods, Davis et al. (2022) generate counterfactuals to explain the required changes, based on representative instances. The selected representatives of similar instances by EE methods (Davis et al. 2022; Demajo et al. 2020) can be used to select instances to represent a particular cluster in the output space. Similar instances aligned to such representatives can assure and improve the confidence of stakeholders.

XAI for textual information targets to improve trustworthiness, informativeness, accessibility, and causality. For TE methods, Yang et al. (2020); Srinivasan et al. (2019) similarly improves trustworthiness with counterfactual texts, with the latter providing alignment according to the user's prompt. The alignment from educational to actionable information enhances information flow, especially for individuals not familiar with the service interface. For VE techniques operating on text, attention weights are widely used for interpretation purposes. Such techniques enhance informativeness and accessibility by using the attention weights to understand regions of focus by the underlying model (Lin et al. 2021; Luo et al. 2018; Yang et al. 2018). On the other hand, Deng et al. (2019); Ito et al. (2020) improve the interpretability of graph neural networks in the financial domain.

XAI for hybrid information leverages both textual and numerical features to improve informativeness and accessibility. These works interpret the black-box model's behavior (Bandi et al. 2021; Cong et al. 2021; Ghosh and Sanyal 2021; Gite et al. 2021; Liu et al. 2020; Maree et al. 2020; Zhang et al. 2020), and provide textual evidence regarding predictions (Bandi et al. 2021; Cong et al. 2021; Ghosh and Sanyal 2021; Gite et al. 2021; Maree et al. 2020; Ong et al. 2023). Adams and Hagras (2020); Carta et al. (2021); Chen and Ye (2022); Dumitrescu et al. (2022); Nazemi et al. (2022); Gramespacher and Posth (2021);

Table 4 The contributions of current FinXAI techniques to ethical goals

Method		Trust	Fairness	Informative	Accessibility	Confidence	Causality
Numerical Features	VE	Zhang et al. (2022a)		Achituve et al. (2019); Biecek et al. (2021); Chen et al. (2020); Crosato et al. (2021); Farzad (2019); Kumar et al. (2017); Shi et al. (2021); Zhang et al. (2022); Zijiao et al. (2022)	Achituve et al. (2019); Crosato et al. (2021); Zhang et al. (2022)		
	ES	Collaris et al. (2018); Maree and Omlin (2022b); Yan et al. (2019)	Misheva et al. (2021); Serengil et al. (2022)				
	FR	Babaei et al. (2022); Cho and Shin (2023); Grath et al. (2018)		Bussmann et al. (2021)	Babaei et al. (2022); Benhamou et al. (2021); Bracke et al. (2019); Bueff et al. (2022); Bussmann et al. (2020); Carta et al. (2022); Demajo et al. (2020); Dikmen and Burns (2022); Fior et al. (2022); Fritz-Morgenthal et al. (2022); Gramegna and Giudici (2020); Islam et al. (2019); Kumar et al. (2022); Lachuer and Jabeur (2022); Müller et al. (2022); Park and Yang (2022); Rallis et al. (2022); Rizinski et al. (2022); Tran et al. (2022); Vivek et al. (2022); Wand et al. (2022); Weng et al. (2022); Yasodhara et al. (2021)		Vivek et al. (2022)
	EE	Davis et al. (2022)				Davis et al. (2022); Demajo et al. (2020)	
Textual Information	TE	Yang et al. (2020)		Srinivasan et al. (2019); Yuan and Zhang (2020)	Srinivasan et al. (2019); Yang et al. (2020)		Yuan and Zhang (2020)
	VE			Lin et al. (2021); Luo et al. (2018); Yang et al. (2018)	Deng et al. (2019); Ito et al. (2020); Lin et al. (2021)		

Table 4 (continued)

Method	Trust	Fairness	Informative	Accessibility	Confidence	Causality
Hybrid Information			Bandi et al. (2021); Cong et al. (2021); Ghosh and Sanyal (2021); Gite et al. (2021); Liu et al. (2020); Maree et al. (2020); Zhang et al. (2020)	Bandi et al. (2021); Cong et al. (2021); Ghosh and Sanyal (2021); Gite et al. (2021); Maree et al. (2020); Ong et al. (2023)		
Transparent Models				Adams and Hagras (2020); Carta et al. (2021); Chen and Ye (2022); Dumitrescu et al. (2022); Gramespacher and Posth (2021); Nazemi et al. (2022); Sudjianto and Zhang (2021)		

Privacy and *Transferability* are omitted due to insufficient works. VE, ES, FR, EE, and TE denote visual explanation, explanation by simplification, feature relevance, explanation by example, and text explanation, respectively

Sudjianto and Zhang (2021) implement transparent models, mitigating the need for post-hoc analysis, and the simplicity of such models improves upon the accessibility for non-expert users. Transparent models may be a suitable choice if the performance is satisfactory and the outcome has to be readily interpretable by non-technical stakeholders. Decision trees are one model which can be easily communicated to audiences without a technical background, given its easily understandable format.

From the above works, we can find that most of the XAI research lies in either studying the underlying model's behavior or identifying important features. Notably, the connotations of informativeness and accessibility goals are rich as seen in Table 4, no XAI technique can achieve all of the desired goals. It is therefore imperative that the XAI design process is tailored towards the desired goals of the target audience Fig. 3. Likewise, the format of presenting explanations is equally important. Non-technical audiences would very much prefer user-friendly visuals as compared to technical plots.

On the other hand, the ethical goal of preserving data-privacy has not been well studied in the works reviewed. Privacy-preserving techniques are a popular research direction, e.g., federated learning (Yang et al. 2019). It is a decentralized learning method that allows parties to collaboratively train a model within a local environment without sharing their data with each other. Generating synthetic data in place of actual data for training models can be one such approach. One example is LIME which generates a local dataset given a target instance, without requiring access to other data instances. This can help to minimize the amount of information being accessed outside of the accountable circle. The understanding of how various features lead to a certain output creates the opportunity of generating more synthetic data. This can be seen as a form of self-supervised learning with the purpose of preserving privacy. However, XAI techniques can also become a double-edged sword, attributing to privacy leakage instead. Such concerns are especially prevalent in techniques manipulating decision boundaries including SVM, K-nearest neighbors, and counterfactual explanations (Sokol and Flach 2019). For example, a counterfactual explanation on reserving a loan application might reveal a suite of sensitive information (location, 10-year income, marital status) to be modified, even though such information is meant to be anonymized. The leaked information can be accessed by third-party providers who may be part of the product design or malicious hackers.

A key challenge is managing the balance between the fidelity of the delivered explanation and the sensitive features altered. Data leakage goes against the privacy awareness goal of XAI and such events are not rare in the financial sector where there exists a constant supply of computerized bots looking to capitalize on these openings. The consequences often affect a large group of public stakeholders (Dellinger 2018), and the affected firm has to pay large fines and incur a loss of trust from their clients. In addition, overly-expressive explanations may allow external competitors to reverse-engineer the models and potentially replicate and improve upon them, thereby compromising the competitive edge a company holds.

XAI techniques that improve transferability are another less frequently studied area. In the field of general AI, transferable knowledge is usually acquired through transfer learning (Neyshabur et al. 2020), multi-task learning (Mao and Li 2021), meta-learning (He et al. 2023), and domain adaptation (Xie et al. 2022). However, the main carrier of these learning paradigms at present is usually deep neural networks. It is difficult to acquire explainability for a deep neural network by using these learning methods. In addition, knowledge forgetting also brings challenges to traditional neural network-based learning methods (He

et al. 2022). Thus, the old knowledge stored in the neural network is likely to be replaced by the learned new knowledge, if the old knowledge is not retained together with the new knowledge. In light of this, how can we leverage explainability and transferability, simultaneously? One possible direction is to utilize neural symbolic techniques. Neural symbolic AI has achieved significant impacts in natural language processing (NLP), e.g., sentiment analysis (Cambria et al. 2024; Zhang et al. 2024) and metaphor processing (Mao et al. 2023, 2024). It takes the merits of both neural networks and symbolic representations. For example, neural networks have strong generalization ability in learning feature representations. Symbolic reasoning enables human-understandable explanations of the system's decision-making process through transparency and interpretability. Since symbolic knowledge can be readily stored in a knowledge base permanently, it avoids the problem of knowledge forgetting in neural networks. A comprehensive and accurate knowledge base can weaken the fitting ability of the neural network. As a result, a more lightweight and transparent neural network can be used in a neural symbolic system. However, developing domain-specific knowledge for finance is costly. Besides, developing symbolic representations for numerical data is also challenging.

Finally, improving fairness, confidence, and causality is also important for ethical concerns. Whereas, the FinXAI research in these areas is very limited. As noted in Table 4, there are not many explanation methods that approach these goals, e.g., ES for fairness with numerical features; EE for confidence with numerical features; and FR and TE for causality with numerical and textual features, respectively. However, it is difficult for a one-fit-all explanation. Hence, we highlight the importance of an audience-centric XAI technique as a more realistic expectation.

8 Challenges and future directions

We exploit the knowledge and insights gained from the agglomeration of FinXAI research conducted thus far and put forward a list of challenges and directions we consider to be important for readers to consider. A few of these limitations have been similarly considered in previous works (Chen and Storchan 2021), which have presented seven major challenges encountered in the context of presenting explanations to stakeholders. Some of these limitations are evident from the reviewed XAI methodologies and we further elaborate on them and cater avenues for improvement.

8.1 Over-reliance

A means of interpreting the model can be helpful while transforming how users interact with data. However, it can cause users to over-rely on possibly inaccurate explanations. A survey was conducted to study how data scientists perceive explanations provided by different XAI tools and found out a large proportion tend to over-trust the explanations provided (Kaur et al. 2020), especially the ones which have received widespread usage. The visual explanations delivered by feature relevance techniques such as SHAP, tend to be absorbed at face value, which can cause researchers to not question their legitimacy. Concurrently, a data scientist who has spent an enormous amount of time designing the AI model may already have prior beliefs on the outcome or model and are more inclined to accept the explanation

if aligned with their initial beliefs (Hohman et al. 2019). Such an occurrence is commonly known as confirmation bias. Over-trusting these explanations can be especially damaging if conveyed to the layperson and can result in the spread of misinformation to a wider audience. It is crucial, therefore, to distinguish between the plausibility of an explanation and its faithfulness (Jacovi and Goldberg 2020). The consequences of this distinction can vary depending on the audience and their specific goals. For instance, a stock trader who mistakenly trusts a falsely attributed feature as the basis for the model's prediction could experience a loss of informativeness or erroneous causal reasoning from the explanation. Similarly, in a credit assessment scenario, an explanation might fail to identify gender as an influential feature, leading to the false assumption that gender is not a factor, even though the model might be disproportionately filtering applications based on gender.

8.2 Social aspects

As mentioned by Miller (2019), explanations are selective, the receiving users tend to only take a minor subset of the entire set of explanations, predominantly those that agree with their prior belief. This can sometimes cause the affected receiver to lose sight of the bigger picture and arrive at some misinterpreted conclusion. Kaur et al. (2020) notes this as a mismatch between the solution's conceptual purpose and the receiver's mental model. XAI tools that produce a feature ranking figure may overcloud the users with excessive information, thereby increasing their cognitive load and rendering the tool counterproductive. It is also observed that the amount of trust is correlated with the level of appreciation the receiver has in the explanation (Mohseni et al. 2021). Take for example the case of a rejected loan application, an under-appreciated explanation would just result in the applicant resubmitting the application to a different bank, without addressing the underlying root cause. Humans also tend to prefer contrastive explanations as opposed to visualizing a large number of probable causes, thus designing the explanation to be counterfactual can reduce under-appreciation and rejection of XAI tools. This thus contributes to the tricky and audience-centric nature of explainability. Future research on human-centric explanations can look to draw inspiration from social sciences and the study of human psychology (Mao et al. 2023, 2024) to bridge the gap between the two ends of the explanation chain.

8.3 Explanation evaluation

It is evident from Table 1, 2, 3, only a small subset of reviewed works attempt to provide some form of quantitative measurement of the proposed XAI technique. An even smaller number performs a comparison between multiple XAI techniques, possibly due to model incompatibility and differences between the explanatory structure of individual XAI techniques. Gurumoorthy et al. (2019) uses a variety of evaluation approaches, grounded on both statistical and human knowledge and involves experts and non-technical users, while admitting the limitation of ambiguity and inconsistency in human judgment. In the case of surrogate model explanations, the fidelity of the surrogate model can be used to measure the accuracy of the approximation. However, such an approach cannot be used for feature relevance tools like SHAP (Amparore et al. 2021). Jie et al. (2024) presents interpretability as a multi-faceted concept, with several traits, each corresponding to achieving a certain goal.

The authors assess the presence of these traits in natural language explanations by LLMs by designing specific tests, tailored towards each trait of interest.

A common basis for instilling interpretability in financial solutions goes beyond the social responsibilities of the provider firms but also concerns the need to comply with the rules and regulations laid out. Even so, the mismatch between each party's perception of explanation sufficiency extends beyond the model itself and includes commonly neglected variables such as accountable personnel, feedback process, and personnel training procedure (Kuiper et al. 2022). Hoffman et al. (2018) highlight that explanations can be seen as a dynamic interaction between the conveyed message and the receiver's thought process. The effectiveness can be measured via goodness and satisfaction in the form of feedback upon receiving the message. This further exemplifies the fact that what makes an explanation good is largely subjective and coming to a consensus on a suitable set of metrics is no trivial task. Among the set of reviewed works in this paper, there exist two forms of evaluation, either through statistical approaches (F1-score, accuracy, and t-test) or opinions of a human expert. The latter is defined as plausibility and should be made distinct from faithfulness, which reflects how the AI model reasons about its behavior. Jacovi and Goldberg (2020) state that the assessment of faithfulness should be independent of human judgment and a common ground can be established by evaluating XAI techniques with respect to a predefined set of goals, rather than on the basis of achieving universal satisfaction. Moreover, in financial contexts, different applications or audiences may prioritize certain aspects of an explanation over others (Agarwal et al. 2024).One criticized flaw in existing works on evaluating interpretation methods pertains to the distributional shift between training and test sets, as well as the infeasible requirement of retraining models. Turbé et al. (2023) creates a set of synthetic datasets with known discriminative features and additionally develops two new metrics which account for identifying top relevant time steps in terms of ranking and score. The proposed method takes into consideration the temporal elements in time-series analysis. We further note that it is imperative for future works on XAI evaluation criteria to precisely define the objective and target audience of the explanation.

8.4 Trade-off between performance and interpretability

It is often common to ponder *"shouldn't we deploy more transparent models if no interpretability enhancement work is required?"*, such an initiative is often plagued by the limited representativeness of transparent models. The trade-off between performance and interpretability is quite commonly a major cause of the dilemma in selecting between black-box models and inherently transparent models. Though there exist studies that have shown that black-box models performing more complex operations do not necessarily lead to better performance (Rudin 2019; Rudin and Radin 2019), it is often the case for unstructured information and noisy environments such as the financial markets. XAI tools explaining through a surrogate model have to face the burden of ensuring both the fidelity of the surrogate model and the effectiveness of the underlying AI model, all the while matching the required goals toward the receiving audience. In light of such a challenge, it highlights the necessity for a consensus metric to serve as a quantitative assessment. In general, it happens more often than not that the selected model at hand is more complex than required, resulting in additional explainability engineering. Moving forward, an efficient way of handling

the trade-off is to prioritize the usage of transparent models if the obtained performance is satisfactory and progress to a more complex model when necessary.

8.5 Better transparent models

We note that transparent models refer to models which exhibit inherent transparency without the need to apply post-hoc explainability techniques. However, caution should be taken in the assumption of such model (Jacovi and Goldberg 2020). The inherent transparency is dependent on the achievable explanation goals and the explanation receiver, while there is much doubt surrounding truly inherently transparent models (Serrano and Smith 2019). Nevertheless, there are numerous studies advocating for a greater need in adopting transparent models. A study by Lipton (2018) argues that transparent models are essential for promoting fairness in machine learning, as they allow for easier identification and mitigation of biases in the decision-making process. A team of researchers (Rudin and Radin 2019) participated in an explainable machine learning challenge and concluded that transparent models do not only sidestep the common issues of trust and misinterpretation but also exhibit the potential to match complex models in terms of performance on specific tasks. It can therefore be for the well-being of society that researchers prioritize the development of more sophisticated and robust transparent machine learning models that are able to balance the trade-off between model accuracy and interpretability.

8.6 Human-centric XAI

We note that, in order to effectively support human decision-making and ensure the interpretability of AI models, there is a growing need for human-centric XAI tools that prioritize user understandability and usability. Explanations are interactive and should be viewed as a bidirectional form of communication (Kaur et al. 2020), with the XAI tool explaining to the user and the user reciprocating back for clarity. Incorporating Human-Computer Interaction (HCI) principles into the design of XAI systems is beneficial in translating interpretability, as HCI principles embrace user-friendly interfaces that enhance human comprehension and engagement. One important aspect of this is the development of interactive systems that enable users to actively engage with and explore XAI models, allowing them to gain a deeper understanding of how the models work and how they can be understood. Several studies have highlighted the importance of incorporating HCI principles into the development of interactive XAI systems (Hohman et al. 2019). The results show that users had more trust when presented with virtual interactive explanations (Weitz et al. 2019). Some popular examples of interactive XAI toolkits include Microsoft AI widgets (2021) and What-if tool by Google (Wexler et al. 2019). Besides being an easily approachable and interpretable tool, interactive systems improve system usability and entice users to frequent the financial services provided, thus adding to the benefits of financial firms prioritizing the development of human-centric, interactive XAI tools.

8.7 Multimodal XAI

A less discussed avenue for improvement is in the incorporation of multimodal information, particularly natural language. Among the works reviewed, a large subset of processed input

data only involves numerical features, while the inclusion of textual information remains a minority. An underlying reason might be due to the redundancy of using textual data or the lack of substantial increase in performance while requiring additional pre-processing works due to the inclusion of such information. Nevertheless, there exist benefits from a transparency point-of-view in incorporating textual information. Danilevsky et al. (2020) highlighted that the separation of the underlying AI model from the explainability tool is less distinct since NLP models, particularly through the use of attention can produce both the prediction and explanation. Likewise, NLP-type explanations can be attractive for the layperson since they exhibit a natural feel which makes the whole process interactive and efficient Cambria et al. (2023). Ultimately, the incorporation of multimodal information entails more flexibility in crafting a good explanation and is supported by the abundance of textual information available. We believe the inclusion of NLP in XAI presents an exciting opportunity to enhance our understanding of financial models and further promote better transparency and trustworthiness in today's AI models.

9 Conclusion

Overall, explainability will continue to be a critical area of focus in FinTech as companies seek to build trust and confidence with consumers and regulators alike. To conclude our work, we have provided a comprehensive review of XAI tools in the financial domain (FinXAI), highlighting the significant progress made in recent years toward developing explainable AI models for financial applications. This includes both inherently transparent models and post-hoc explainability techniques, the former of which we advocate for more improvements to be made. We provided a framework that establishes the selection of appropriate FinXAI tools as a sequential decision-making process, placing great emphasis on the audience and iterative assessment of produced explanation. The reviewed works are categorized according to their respective characteristics for ease of access by interested readers. We also examine the contributions of current FinXAI to several ethical goals, e.g., trustworthiness, fairness, informativeness, accessibility, privacy, confidence, causality, and transparency.

Though there have been many great works done thus far, the review also reveals some limitations and challenges associated with FinXAI. This includes appropriate metrics to measure both the faithfulness and plausibility of explanations, as well as issues concerning the over-reliance on potentially misleading explanations. Future research should focus on addressing these challenges, as well as exploring new directions for FinXAI, including integrating NLP into explanation-generating techniques and a greater focus on inherently transparent models. Nevertheless, there is great potential for XAI techniques to enhance transparency, trust, and accountability in the financial domain. This underscores the importance of active research and development in this field.

References

Athey S, et al. (2018) The impact of machine learning on economics. The economics of artificial intelligence: an agenda, 507–547

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, et al. (2023) Gpt-4 technical report. Preprint at arXiv:2303.08774

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion 58:82–115

Adams J, Hagras H (2020) A type-2 fuzzy logic approach to explainable ai for regulatory compliance, fair customer outcomes and market stability in the global financial sector. In: 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8. IEEE

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

Alkaissi H, McFarlane SI (2023) Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus **15**(2)

Agarwal R, Melnick L, Frosst N, Zhang X, Lengerich B, Caruana R, Hinton GE (2021) Neural additive models: interpretable machine learning with neural nets. Adv Neural Inf Process Syst 34:4699–4711

Amparore E, Perotti A, Bajardi P (2021) To trust or not to trust an explanation: using leaf to evaluate local linear XAI methods. PeerJ Computer Science 7:479

Aghabozorgi S, Teh YW (2014) Stock market co-movement assessment using a three-phase clustering method. Expert Syst Appl 41(4):1301–1314

Agarwal C, Tanneru SH, Lakkaraju H (2024) Faithfulness vs. plausibility: on the (un) reliability of explanations from large language models. Preprint at arXiv:2402.04614

Apley DW, Zhu J (2020) Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Ser B Stat Methodol 82(4):1059–1086

Bahrammirzaee A (2010) A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. Neural Comput Appl 19(8):1165–1195

Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7):0130140

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

Biecek P, 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. Preprint at arXiv:2104.06735

Bracke P, Datta A, Jung C, Sen S (2019) Machine learning explainability in finance: an application to default risk analysis

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

Babaei G, Giudici P, Raffinetti E (2022) Explainable artificial intelligence for crypto asset allocation. Financ Res Lett 47:102941

Bandi H, Joshi S, Bhagat S, Ambawade D (2021) Integrated technical and sentiment analysis tool for market index movement prediction, comprehensible using XAI. In: 2021 International Conference on Communication Information and Computing Technology (ICCICT), pp. 1–8. IEEE

Benhamou E, Ohana J-J, Saltiel D, Guez B (2021) Explainable AI (XAI) models applied to planning in financial markets

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

Chen J-H, Chen SY-C, Tsai Y-C, Shur C-S (2020) Explainable deep convolutional candlestick learner. Preprint at arXiv:2001.02767

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794
Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. Preprint at arXiv:1412.3555
Crosato L, Liberati C, Repetto M (2021) Look who's talking: Interpretable machine learning for assessing Italian SMES credit default. Preprint at arXiv:2108.13914
Cambria E, Mao R, Chen M, Wang Z, Ho S-B (2023) Seven pillars for the future of Artificial Intelligence. IEEE Intell Syst 38(6):62–69
Cambria E, Malandri L, Mercorio F, Mezzanzanica M, Nobani N (2023) A survey on XAI and natural language explanations. Inform Process Manag 60(1):103111
Chen X-Q, Ma C-Q, Ren Y-S, Lei Y-T, Huynh NQA, Narayan S (2023) Explainable artificial intelligence in finance: a bibliometric review. Finance Research Letters, 104145
Carta S, Podda AS, Reforgiato Recupero D, Stanciu MM (2022) Explainable ai for financial forecasting. In: Machine Learning, Optimization, and Data Science: 7th International Conference, LOD 2021, Grasmere, UK, October 4–8, 2021, Revised Selected Papers, Part II, pp. 51–69. Springer
Chen J, Storchan V (2021) Seven challenges for harmonizing explainability requirements. Preprint at arXiv:2108.05390
Cho SH, Shin K-S (2023) Feature-weighted counterfactual-based explanation for bankruptcy prediction. Expert Syst Appl 216:119390
Cong LW, Tang K, Wang J, Zhang Y (2021) Alphaportfolio: direct construction through deep reinforcement learning and interpretable AI. Available at SSRN **3554486**
Collaris D, Vink LM, Wijk JJ (2018) Instance-level explanations for fraud detection: a case study. Preprint at arXiv:1806.07129
Chen D, Ye W (2022) Generalized gloves of neural additive models: Pursuing transparent and accurate machine learning models in finance. Preprint at arXiv:2209.10082
Cambria E, Zhang X, Mao R, Chen M, Kwok K (2024) SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing. In: Proceedings of international conference on human-computer interaction (HCII), Washington DC, USA
Dikmen M, Burns C (2022) The effects of domain knowledge on trust in explainable ai and task performance: a case of peer-to-peer lending. Int J Hum Comput Stud 162:102792
De Bruin KC, Dellink RB, Tol RS (2009) AD-DICE: an implementation of adaptation in the dice model. Clim Change 95:63–81
Dellinger, A.: Understanding The First American Financial Data Leak: How Did It Happen And What Does It Mean? https://www.forbes.com/sites/ajdellinger/2019/05/26/understanding-the-first-american-financial-data-leak-how-did-it-happen-and-what-does-it-mean/?sh=7716df86567f
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
Davis R, Lo AW, Mishra S, Nourian A, Singh M, Wu N, Zhang R (2022) Explainable machine learning models of consumer credit risk. Available at SSRN
Du K, Mao R, Xing F, Cambria E (2024) Explainable stock price movement prediction using contrastive learning. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM), Idaho, USA
Danilevsky M, Qian K, Aharonov R, Katsis Y, Kawas B, Sen P (2020) A survey of the state of explainable ai for natural language processing. Preprint at arXiv:2010.00711
Datta A, Sen S, Zick Y (2016) Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: 2016 IEEE Symposium on Security and Privacy (SP), pp. 598–617. IEEE
Demajo LM, Vella V, Dingli A (2020) Explainable ai for interpretable credit scoring. In: CS & IT Conference Proceedings, vol. 10. CS & IT Conference Proceedings
Du K, Xing F, Mao R, Cambria E (2024) Financial sentiment analysis: techniques and applications. ACM Comput Surv 56(9):1–42
Deng S, Zhang N, Zhang W, Chen J, Pan JZ, Chen H (2019) Knowledge-driven stock trend prediction and explanation via temporal convolutional network. In: Companion Proceedings of The 2019 World Wide Web Conference, pp. 678–685
Farzad T (2019) Determinants of Mortgage Loan Delinquency: Application of Interpretable Machine Learning
Fior J, Cagliero L, Garza P (2022) Leveraging explainable ai to support cryptocurrency investors. Future Internet 14(9):251
Fritz-Morgenthal S, Hein B, Papenbrock J (2022) Financial risk management and explainable, trustworthy, responsible AI. Front Artif Intell 5:5

Grath RM, Costabello L, Van CL, Sweeney P, Kamiab F, Shen Z, Lecue F (2018) Interpretable credit application predictions with counterfactual explanations. Preprint at arXiv:1811.05245
Gurumoorthy KS, Dhurandhar A, Cecchi G, Aggarwal C (2019) Efficient data representation by selecting prototypes with importance weights. In: 2019 IEEE International Conference on Data Mining (ICDM), pp. 260–269. IEEE
Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a "right to explanation". AI Mag 38(3):50–57
Gramegna A, Giudici P (2020) Why to buy insurance? An explainable artificial intelligence approach. Risks 8(4):137
Gite S, Khatavkar H, Kotecha K, Srivastava S, Maheshwari P, Pandey N (2021) Explainable stock prices prediction from financial news articles using sentiment analysis. Peer J Comput Sci 7:340
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
Gramespacher T, Posth J-A (2021) Employing explainable ai to optimize the return target function of a loan portfolio. Front Artif Intell 4:693022
Ghosh I, Sanyal MK (2021) Introspecting predictability of market fear in Indian context during covid-19 pandemic: an integrated approach of applied predictive modelling and explainable AI. Int J Inform Manag Data Insights 1(2):100039
Hohman F, Head A, Caruana R, DeLine R, Drucker SM (2019) Gamut: A design probe to understand how data scientists understand machine learning models. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13
HLEG, A.: Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
Han S, Mao R, Cambria E (2022) Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. In: Proceedings of the 29th International Conference on Computational Linguistics (COLING), pp. 94–104. International Committee on Computational Linguistics, Gyeongju, Republic of Korea
He K, Mao R, Gong T, Cambria E, Li C (2022) JCBIE: A joint continual learning neural network for biomedical information extraction. BMC Bioinform 23(549):1–20
He K, Mao R, Gong T, Li C, Cambria E (2023) Meta-based self-training and re-weighting for aspect-based sentiment analysis. IEEE Trans Affect Comput 14(3):1731–1742
Hoffman RR, Mueller ST, Klein G, Litman J (2018) Metrics for explainable ai: Challenges and prospects. Preprint at arXiv:1812.04608
Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, et al. (2023) A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. Preprint at arXiv:2311.05232
Islam SR, Eberle W, Bundy S, Ghafoor SK (2019) Infusing domain knowledge in ai-based " black box" models for better explainability with application in bankruptcy prediction. Preprint at arXiv:1905.11474
Ito T, Sakaji H, Izumi K, Tsubouchi K, Yamashita T (2020) Ginn: gradient interpretable neural networks for visualizing financial texts. Int J Data Sci Anal 9:431–445
Jacovi A, Goldberg Y (2020) Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? Preprint at arXiv:2004.03685
Jie YW, Satapathy R, Goh R, Cambria E (2024) How interpretable are reasoning explanations from prompting large language models? In: Findings of the Association for Computational Linguistics: NAACL 2024, pp 2148–2164
Jin X, Wei Z, Du J, Xue X, Ren X (2019) Towards hierarchical importance attribution: explaining compositional semantics for neural sequence models. Preprint at arXiv:1911.06194
Kolanovic M, Krishnamachari RT. Big data and AI strategies, machine learning and alternative data approach to investing (2017)
Koa KJ, Ma Y, Ng R, Chua T-S (2024) Learning to generate explainable stock predictions using self-reflective large language models. Preprint at arXiv:2402.03659
Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J (2020) Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp 1–14
Kumar D, Taylor GW, Wong A (2017) Opening the black box of financial ai with clear-trade: A class-enhanced attentive response approach for explaining and visualizing deep learning-driven stock market prediction. Preprint at arXiv:1709.01574
Kuiper O, Berg M, Burgt J, Leijnen S (2022) Exploring explainable ai in the financial sector: Perspectives of banks and supervisory authorities. In: Artificial Intelligence and Machine Learning: 33rd Benelux Conference on Artificial Intelligence, BNAIC/Benelearn 2021, Esch-sur-Alzette, Luxembourg, November 10–12, 2021, Revised Selected Papers 33, pp 105–119. Springer

Kumar S, Vishal M, Ravi V (2022) Explainable reinforcement learning on financial stock trading using shap. Preprint at arXiv:2208.08790
Luo L, Ao X, Pan F, Wang J, Zhao T, Yu N, He Q (2018) Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention. In: IJCAI, pp. 4244–4250
Li J, Bian Y, Wang G, Lei Y, Cheng D, Ding Z, Jiang C (2023) Cfgpt: Chinese financial assistant with large language model. Preprint at arXiv:2309.10654
Lundberg SM, Erion GG, Lee S-I (2018) Consistent individualized feature attribution for tree ensembles. Preprint at arXiv:1802.03888
Levine TR (2014) Truth-default theory (TDT) a theory of human deception and deception detection. J Lang Soc Psychol 33(4):378–392
Lipton P (1990) Contrastive explanation. Royal Inst Philos Suppl 27:247–266
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
Lachuer J, Jabeur SB (2022) Explainable artificial intelligence modeling for corporate social responsibility and financial performance. J Asset Manag 23(7):619–630
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Advances in neural information processing systems **30**
Lopez-Lira A, Tang Y (2023) Can chatgpt forecast stock price movements? return predictability and large language models. Preprint at arXiv:2304.07619
Liu R, Mai F, Shan Z, Wu Y (2020) Predicting shareholder litigation on insider trading from financial text: an interpretable deep learning approach. Inform Manag 57(8):103387
Li Y, Ma S, Wang X, Huang S, Jiang C, Zheng H-T, Xie P, Huang F, Jiang Y (2024) Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 18582–18590
Lin T-W, Sun R-Y, Chang H-L, Wang C-J, Tsai M-F (2021) Xrr: Explainable risk ranking for financial reports. In: Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part IV 21, pp 253–268. Springer
Liang B, Su H, Gui L, Cambria E, Xu R (2022) Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. Knowl-Based Syst 235:107643
Markowitz H (1952) Portfolio selection. J Finance 7(1):77–91
Mao R, Chen G, Zhang X, Guerin F, Cambria E (2024) GPTEval: A survey on assessments of ChatGPT and GPT-4. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp 7844–7866. ELRA and ICCL, Torino, Italia
Mao R, Du K, Ma Y, Zhu L, Cambria E (2023) Discovering the cognition behind language: Financial metaphor analysis with MetaPro. In: 2023 IEEE International Conference on Data Mining (ICDM), Shanghai, China, pp 1211–1216. IEEE
Mellon B. Why every financial institution should consider explainable AI. https://www.bnymellon.com/us/en/insights/all-insights/why-every-financial-institution-should-consider-explainable-ai.html
Mueller ST, Hoffman RR, Clancey W, Emrey A, Klein G (2019) Explanation in human-AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. Preprint at arXiv:1902.01876
Mao R, He K, Ong CB, Liu Q, Cambria E (2024) MetaPro 2.0: Computational metaphor processing on the effectiveness of anomalous language modeling. In: Findings of the Association for Computational Linguistics: ACL, pp 9891–9908. Association for Computational Linguistics, Bangkok, Thailand
Microsoft Responsible AI Toolbox. https://github.com/microsoft/responsible-ai-toolbox
Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38
Mao R, Li X (2021) Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. Proceed AAAI Conf Artif Intell 35(15):13534–13542
Mao R, Li X, He K, Ge M, Cambria E (2023) MetaPro Online: A computational metaphor processing online system. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), vol. 3, pp 127–135. Association for Computational Linguistics, Toronto, Canada
Ma Y, Mao R, Lin Q, Wu P, Cambria E (2023) Multi-source aggregated classification for stock price movement prediction. Inform Fusion 91:515–528
Ma Y, Mao R, Lin Q, Wu P, Cambria E (2024) Quantitative stock portfolio optimization by multi-task learning risk and return. Inform Fusion 104:102165
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

Maree C, Omlin CW (2022) Can interpretable reinforcement learning manage prosperity your way? AI 3(2), 526–537
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
Misheva BH, Osterrieder J, Hirsa A, Kulkarni O, Lin SF (2021) Explainable ai in credit risk management. Preprint at arXiv:2103.00949
Molnar C (2020) Interpretable Machine Learning. Lulu.com
Mroczkowska A. What Is a Fintech Application?, Definition and Insights for Business Owners. https://www.thedroidsonroids.com/blog/what-is-a-fintech-application-definition-and-insights-for-business-owners/
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
Mohseni S, Yang F, Pentyala S, Du M, Liu Y, Lupfer N, Hu X, Ji S, Ragan E (2021) Machine learning explanations to prevent overtrust in fake news detection. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, pp 421–431
Mao R, Zhang T, Liu Q, Hussain A, Cambria E (2024) Unveiling diplomatic narratives: Analyzing United Nations Security Council debates through metaphorical cognition. In: Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci), vol. 46. Rotterdam, the Netherlands, pp 1709–1716
Mohseni S, Zarei N, Ragan ED (2021) A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transact Interact Intell Syst (TiiS) 11(3–4):1–45
Nori H, Jenkins S, Koch P, Caruana R (2019) Interpretml: A unified framework for machine learning interpretability. Preprint at arXiv:1909.09223
Nazemi A, Rauch J, Fabozzi FJ (2019) Interpretable machine learning for creditor recovery rates. Available at SSRN 4190345
Neyshabur B, Sedghi H, Zhang C (2020) What is being transferred in transfer learning? Adv Neural Inf Process Syst 33:512–523
Ong K, Mao R, Satapathy R, Cambria E, Sulaeman J, Mengaldo G, et al. (2024) Explainable natural language processing for corporate sustainability analysis. Preprint at arXiv:2407.17487
Singapore, M.A.: Veritas Initiative addresses implementation challenges in the responsible use of artificial intelligence and data analytics. https://www.mas.gov.sg/news/media-releases/2021/veritas-initiative-addresses-implementation-challenges
Ong K, Heever W, Satapathy R, Cambria E, Mengaldo G (2023) Finxabsa: explainable finance through aspect-based sentiment analysis. In: 2023 IEEE International Conference on Data Mining Workshops (ICDMW), pp 773–782. IEEE
Park S, Yang J-S (2022) Interpretable deep learning LSTM model for intelligent economic decision-making. Knowl-Based Syst 248:108907
Rallis I, Markoulidakis Y, Georgoulas I, Kopsiaftis G, Kaselimi M, Doulamis N, Doulamis A (2022) Interpretation of net promoter score attributes using explainable AI. In: Proceedings of the 15th international conference on pervasive technologies related to assistive environments, pp 113–117
Rawal A, McCoy J, Raglin A, Rawat DB (2023) A quantitative comparison of causality and feature relevance via explainable ai (xai) for robust, and trustworthy artificial reasoning systems. In: International conference on human-computer interaction, pp 274–285. Springer
Rojat T, Puget R, Filliat D, Del Ser J, Gelin R, Díaz-Rodríguez N (2021) Explainable artificial intelligence (XAI) on timeseries data: a survey. Preprint at arXiv:2104.00950
Rizinski M, Peshov H, Mishev K, Chitkushev LT, Vodenska I, Trajanov D (2022) Ethically responsible machine learning in fintech. IEEE Access 10:97531–97554
Rudin C, Radin J (2019) Why are we using black box models in ai when we don't need to? A lesson from an explainable AI competition. Harvard Data Sci Rev 1(2):10–1162
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
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
Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
Shapley LS, et al. (1953) A value for n-person games
Sahakyan M, Aung Z, Rahwan T (2021) Explainable artificial intelligence for tabular data: a survey. IEEE Access 9:135392–135422

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626

Srinivasan R, Chander A, Pezeshkpour P (2019) Generating user-friendly explanations for loan denials using Gans. Preprint at arXiv:1906.10244

Sokol K, Flach PA (2019) Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. SafeAI@ AAAI

Serengil SI, Imece S, Tosun UG, Buyukbas EB, Koroglu B (2022) A comparative study of machine learning approaches for non performing loan prediction with explain ability. Int J Mach Learn Comput **12**(5)

Shi S, Li J, Li G, Pan P, Liu K (2021) Xpm: An explainable deep reinforcement learning framework for portfolio management. In: Proceedings of the 30th ACM international conference on information & knowledge management, pp. 1661–1670

Serrano S, Smith NA (2019) Is attention interpretable? Preprint at arXiv:1906.03731

Sudjianto A, Zhang A (2021) Designing inherently interpretable machine learning models. Preprint at arXiv:2111.01743

Tomsett R, Braines D, Harborne D, Preece A, Chakraborty S (2018) Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. Preprint at arXiv:1806.07552

Turbé H, Bjelogrlic M, Lovis C, Mengaldo G (2023) Evaluation of post-HOC interpretability methods in time-series classification. Nat Mach Intell. https://doi.org/10.1038/s42256-023-00620-w

Team C. Finance overview: personal,business and government. https://corporatefinanceinstitute.com/resources/wealth-management/finance-industry-overview/

Tran KL, Le HA, Nguyen TH, Nguyen DT (2022) Explainable machine learning for financial distress prediction: evidence from Vietnam. Data 7(11):160

Berg M, Kuiper O (2020) Xai in the financial sector: a conceptual framework for explainable AI (XAI). https://www.hu.nl/-/media/hu/documenten/onderzoek/projecten/

VINCENT J. Google 'fixed' Its Racist Algorithm by Removing Gorillas from Its Image-labeling Tech. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

Vivek Y, Ravi V, Mane AA, Naidu LR (2022) Explainable artificial intelligence and causal inference based atm fraud detection. Preprint at arXiv:2211.10595

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inform Process Syst **30**

Weber P, Carl KV, Hinz O (2023) Applications of explainable artificial intelligence in finance-a systematic review of finance, information systems, and computer science literature. Manag Rev Quart 1–41

Wand T, Heßler M, Kamps O (2022) Identifying dominant industrial sectors in market states of the s &p 500 financial data. Preprint at arXiv:2208.14106

Wu S, Irsoy O, Lu S, Dabravolski V, Dredze M, Gehrmann S, Kambadur P, Rosenberg D, Mann G (2023) Bloomberggpt: a large language model for finance. Preprint at arXiv:2303.17564

Wexler J, Pushkarna M, Bolukbasi T, Wattenberg M, Viégas F, Wilson J (2019) The what-if tool: interactive probing of machine learning models. IEEE Trans Visual Comput Graphics 26(1):56–65

Weitz K, Schiller D, Schlagowski R, Huber T, André E (2019) " do you trust me?" increasing user-trust by integrating virtual agents in explainable ai interaction design. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, pp 7–9

Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F (2023) Improving text embeddings with large language models. Preprint at arXiv:2401.00368

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

Xie Q, Han W, Lai Y, Peng M, Huang J (2023) The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges. Preprint at arXiv:2304.05351

Xie Q, Han W, Zhang X, Lai Y, Peng M, Lopez-Lira A, Huang J (2024) Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. Adv Neural Inform Process Syst **36**

Xie B, Yuan L, Li S, Liu CH, Cheng X, Wang G (2022) Active learning for domain adaptation: An energy-based approach. In: Proceedings of the AAAI conference on artificial intelligence, vol. 36, pp 8708–8716

Yasodhara A, Asgarian A, Huang D, Sobhan'i P (2021) On the trustworthiness of tree ensemble explainability methods. In: Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings 5, pp 293–308. Springer

Yang L, Kenny E, Ng TLJ, Yang Y, Smyth B, Dong R (2020) Generating plausible counterfactual explanations for deep transformers in financial text classification. In: Proceedings of the 28th International Conference on Computational Linguistics, pp 6150–6160

Yan H, Lin S et al (2019) New trend in fintech: Research on artificial intelligence model interpretability in financial fields. Open J Appl Sci 9(10):761
Yang Q, Liu Y, Chen T, Tong Y (2019) Federated machine learning: concept and applications. ACM Transact Intell Syst Technol (TIST) 10(2):1–19
Yang H, Liu X-Y, Wang CD (2023) Fingpt: Open-source financial large language models. Preprint at arXiv:2306.06031
Yuan J, Zhang Z (2020) Connecting the dots: forecasting and explaining short-term market volatility. In: Proceedings of the First ACM International Conference on AI in Finance, pp 1–8
Yeong Zee Kin TWR. Lee Wan Sie: How Singapore is developing trustworthy AI. https://www.weforum.org/agenda/2023/01/how-singapore-is-demonstrating-trustworthy-ai-davos2023/
Yang L, Zhang Z, Xiong S, Wei L, Ng J, Xu L, Dong R (2018) Explainable text-driven neural network for stock prediction. In: 2018 5th IEEE international conference on cloud computing and intelligence systems (CCIS), pp 441–445. IEEE
Zhang W, Barr B, Paisley J (2022) An interpretable deep classifier for counterfactual generation. In: Proceedings of the Third ACM International Conference on AI in Finance, pp 36–43
Zhang W, Barr B, Paisley J (2022) Understanding counterfactual generation using maximum mean discrepancy. In: Proceedings of the Third ACM International Conference on AI in Finance, pp 44–52
Zhang CA, Cho S, Vasarhelyi M (2022) Explainable artificial intelligence (XAI) in auditing. Int J Account Inf Syst 46:100572
Zhang X, Du Q, Zhang Z (2020) An explainable machine learning framework for fake financial news detection. In: 2020 International Conference on Information Systems-Making Digital Inclusive: Blending the Local and the Global, ICIS 2020. Association for Information Systems
Zhang X, Mao R, Cambria E (2024) SenticVec: Toward robust and human-centric neurosymbolic sentiment analysis. In: Findings of the Association for Computational Linguistics: ACL, pp 4851–4863. Association for Computational Linguistics, Bangkok, Thailand
Zijiao Z, Wu C, Qu S, Chen X (2022) An explainable artificial intelligence approach for financial distress prediction. Inform Process Manag 59(4):102988
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