Artificial Intelligence Review (2025) 58:232
https://doi.org/10.1007/s10462-025-11215-9

Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions

Farhina Sardar Khan[1] · Syed Shahid Mazhar[2] · Kashif Mazhar[3] · Dhoha A. AlSaleh[4] · Amir Mazhar[2]

Accepted: 29 March 2025 / Published online: 3 May 2025
© The Author(s) 2025

Abstract

The increasing integration of Artificial Intelligence (AI) and Machine Learning (ML)—algorithms that enable computers to identify patterns from data—in financial applications has significantly improved predictive capabilities in areas such as credit scoring, fraud detection, portfolio management, and risk assessment. Despite these advancements, the opaque, "black box" nature of many AI and ML models raises critical concerns related to transparency, trust, and regulatory compliance. Explainable Artificial Intelligence (XAI) aims to address these issues by providing interpretable and transparent decision-making processes. This study systematically reviews Model-Agnostic Explainable AI techniques, which can be applied across different types of ML models in finance, to evaluate their effectiveness, scalability, and practical applicability. Through analysis of 150 peer-reviewed studies, the paper identifies key challenges, such as balancing interpretability with predictive accuracy, managing computational complexity, and meeting regulatory requirements. The review highlights emerging trends toward hybrid models that combine powerful ML algorithms with interpretability techniques, real-time explanations suitable for dynamic financial markets, and XAI frameworks explicitly designed to align with regulatory standards. The study concludes by outlining specific future research directions, including the development of computationally efficient explainability methods, regulatory-compliant frameworks, and ethical AI solutions to ensure transparent and accountable financial decision-making.

Keywords Artificial intelligence · Machine learning · Explainable AI · Finance · Transparency · Regulatory compliance

Extended author information available on the last page of the article

1 Introduction

1.1 Background and motivation

In the past two decades, AI has advanced rapidly and is now applied across various sectors and activities, including and not limited to finance (Bahoo et al. 2024), business management and marketing (Verma et al. 2021; Gil et al. 2020; Raisch and Krakowski 2021; Thakur et al. 2023), healthcare (Saraswat et al. 2022; AlSaleh 2019; Shaheen 2021) and engineering (Ozkaya 2020; Barenkamp et al. 2020; Ebid 2021). The first two decades of the twenty-first century have witnessed unparalleled technological advancements, propelled by the development of state-of-the-art digitally supported technologies and applications in AI (Weber et al. 2024). AI is a field of computer science that focuses on creating intelligent machines that can perform cognitive tasks typically associated with human abilities, such as reasoning, learning, decision-making, and speech recognition (Eluwole and Akande 2022; Bahoo et al. 2024). Different features of AI have played a major role in various fields, such as finance, engineering, and medical sciences. AI systems must ensure the safety and security of citizens, act as a safeguard for the well-being of society (Stahl 2021). Therefore, Fig. 1 highlights the key aspects of various AI applications.

The most notable advancement and proliferation of AI-related technologies have occurred recently, driven by the availability of large unstructured datasets, a surge in computing power, and increased venture capital funding for innovative technological projects (Ernst et al. 2019). The implementation of AI is poised to have significant implications for adopters and society at large, potentially boosting global GDP. A study by PricewaterhouseCoopers (PwC) in 2017 suggested that GDP could rise significantly by up to 14% by 2030. Furthermore, companies that integrate AI-enabled solutions and technologies often report improved performance (Roy et al. 2020). ML is the primary technology that drives AI. ML methods empower machines to perform intricate tasks, such as facial recognition,

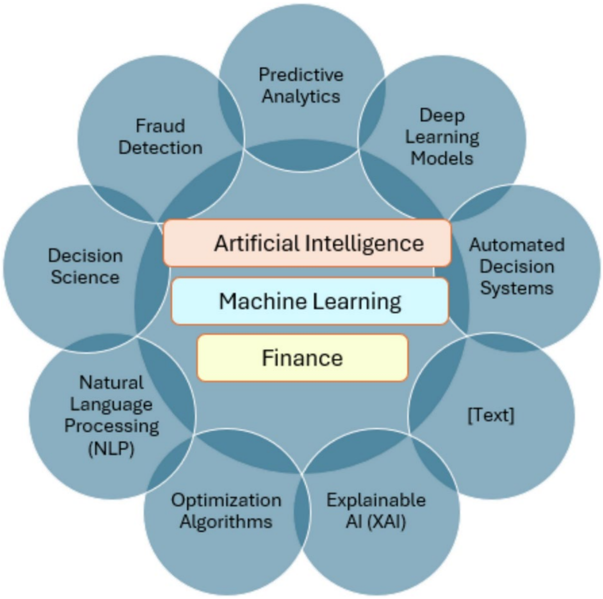


Fig. 1 Key features of AI across multiple domains, highlighting its applications in finance, healthcare, and decision-making systems

speech understanding, and message responses (Bonissone 2015). Given the capabilities of ML technology, its potential applicability in other domains has been questioned (Hoang and Wiegratz 2023). The finance sector is continually evolving, actively embracing and adapting to emerging technological opportunities such as AI and data analytics, which significantly influence personal and professional lives globally (Gimpel et al. 2018). AI has progressed significantly in the last decade, driven by substantial funding and the ambition of AI experts to transition narrow AI into artificial general intelligence capable of seamlessly performing tasks that humans typically do, potentially passing the Turing test in all routine activities (Ali et al. 2023a, b). AI has witnessed extensive adoption across various domains of finance in recent years for important financial applications, including multi-language financial sentiment analysis (Ardekani et al. 2024), forecasting and prediction of inflation in emerging economies (Mirza et al. 2024), management of trading and portfolios (Zhang et al. 2020), financial modelling of risks (Mashrur et al. 2020), volatility index prediction (Gunnarsson et al. 2024), financial text mining problems (Gupta et al. 2020; Pagliaro et al. 2021), credit risk assessment problems using neural networks (NNs; Bhattacharjee et al. 2017), financial advisory and customer services (Shah et al. 2020), Large Language Models (LLMs; Li et al. 2023), classification and prediction, as well as in image processing, computer vision and audio-visual recognition (Jalal et al. 2022; Rupapara et al. 2021) and determining the voluntary disclosure using the eXtreme gradient boost (XGBoosT) algorithm (Lu and Lin 2024). Although DL was instigated in computer science, its applications have been extended to diverse fields including neuroscience, physics, medicine, astronomy, and operations management (Rupapara et al. 2022; Rashid et al. 2013). The impressive success of DL as a data-processing method has garnered substantial attention from researchers. In recent years, with the rapid expansion of Fintech, DL has been increasingly adopted in the financial and investment sectors (Huang et al. 2020). Various ML and DL models have been extensively applied in the financial domain such as Support Vector Machines (SVM; Kim 2003), Xgboost (Zolotareva 2021), Long Short-Term Memory (LSTM) networks (Sezer et al. 2017), Convolutional Neural Networks (CNN; Sezer and Ozbayoglu 2018), and transformers (Wen et al. 2022), which have been extensively used for profit and loss estimation, price forecasting, portfolio selection (Jiang et al. 2024), automatic trading, and portfolio optimization with over 40 research publications dedicated to this topic (Ozbayoglu et al. 2020). The authors of (Roy et al. 2018) developed a DL-based solution for financial fraud detection by leveraging user history and real-time transaction data. Similar approaches have been employed by researchers in credit scoring tasks (Luo et al. 2017; West 2000) and the prediction of bankruptcy or default (Chen 2011). DL models provide efficient insights from large datasets quickly, benefiting finance with timely and accurate decision making. Study (Kim 2020) examined knowledge imbalances, unethical behaviour, agency relationships, and strategies to address the principal-agent issue using DL algorithms. LLMs extend AI's reach of AI, tackling previously impossible tasks and broadening AI applications (Li et al. 2023) in finance, as shown in Fig. 2.

1.2 Objectives of the study

The objectives of this study are:

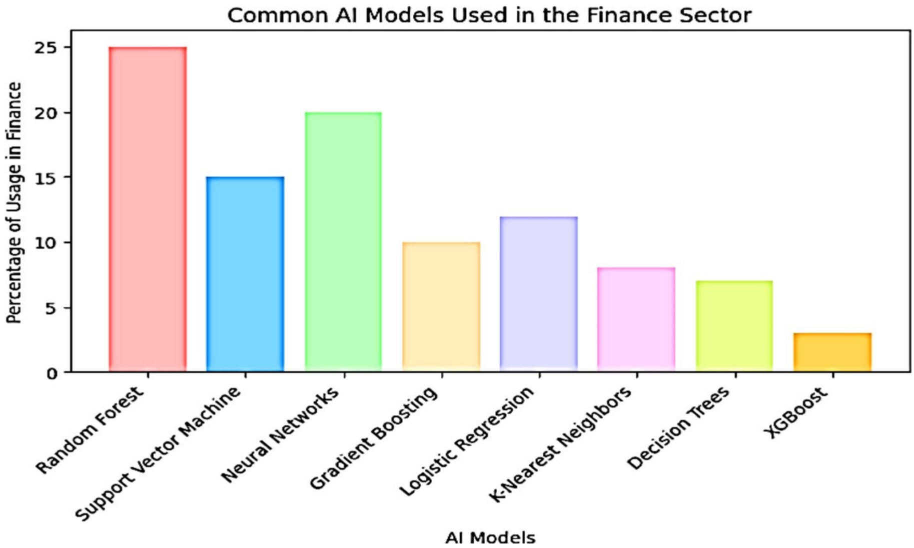


Fig. 2 A comparative overview of commonly used AI models in finance, including ML, DL, and XAI, illustrating their respective roles in financial decision-making

1.2.1 Systematic literature review (SLR)

To perform a comprehensive review of existing literature on Explainable Artificial Intelligence (XAI) in finance, particularly focusing on Model-Agnostic (MA) explanations.

1.2.2 Rigorous documentation

To meticulously document 150 selected studies using stringent filtering criteria in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

1.2.3 Analysis of XAI techniques

To explore and analyze prevalent Model-Agnostic (MA-XAI) techniques in finance, such as SHAP, LIME, Counterfactual Explanations, and Partial Dependence Plots (PDPs), highlighting their applications and effectiveness.

1.2.4 Exploration of datasets and performance metrics

To investigate commonly used financial datasets and examine performance metrics utilized in evaluating the effectiveness of XAI methods within financial research contexts.

1.2.5 Criteria for selecting MA-XAI methods

To discuss and detail the criteria guiding the selection and application of MA-XAI methods specifically within financial applications.

1.2.6 Identification of limitations and advantages

To outline the limitations and advantages associated with the implementation of MA-XAI techniques in the finance sector.

1.2.7 Future research directions

To propose future research directions emphasizing hybrid XAI methods, domain-specific customization, and enhancing real-time interpretability, facilitating the practical adoption of XAI solutions in financial decision-making contexts.

1.3 Terminologies in XAI

1.3.1 Explainability

The process of clarifying or uncovering the decision-making processes of models allows users to see the mathematical connections between the inputs and outputs. This pertains to the ability to comprehend why AI models make specific decisions. The ability to automatically interpret and explain the inner workings of an AI system in human terms is known as explainability. An explainable method provides a summary of the reasons behind the decisions made by an AI model. Additionally, "post-hoc explainability" refers to the methods or algorithms used to explain the decisions made by AI models after they have been made (Adadi and Berrada 2018; Arrieta et al. 2020; Das and Rad 2020; Bruckert et al. 2020; Schwalbe and Finzel 2023; Shams Khoozani et al. 2024; Li et al. 2022; Viswan et al. 2024; Raees et al. 2024). According to (Yang et al. 2022) explainability refers to a category of systems designed to provide insight into how an AI system makes decisions and predictions. XAI delves into the rationale behind the decision-making process, highlights the strengths and weaknesses of the system, and offers a preview of the system's future behavior.

1.3.2 Transparency

Transparency refers to the ability to comprehend and explain the decisions and reasoning of an AI system. As AI systems become increasingly complex and impactful across various fields, the need for transparency is rising to ensure accountability, fairness, and trustworthiness (Letrache and Ramdani 2023). This is achieved through an intrinsic method that produces a human-readable explanation of the model's decisions. Transparency is crucial for evaluating the quality of a model's decisions and protecting it against adversarial attacks (Li et al. 2022; Dosilovic et al. 2018; Larsson and Heintz 2020; Bogina et al. 2022).

1.3.3 Fairness

Owing to the inherent biases in certain datasets and algorithms, AI systems can unfairly discriminate against specific groups of people. In this context fairness means that a model can make impartial decisions without showing favouritism towards any population represented in the input data distribution (Das and Rad 2020). Biases related to factors such as birth location, socioeconomic status, and skills should not influence AI models (Mehrabi et al. 2022; Bogina et al. 2022). Throughout the development and deployment of AI systems, it is crucial to implement specialized methods for gathering and integrating user feedback (Calders et al. 2021; Lyu et al. 2020).

1.3.4 Interpretability

The ability to understand and explain the decisions or behaviors of AI models and systems in a manner that is meaningful and understandable to humans. It aims to provide insights into the internal workings and reasoning of AI systems, allowing users to trust, validate, and comprehend their outputs (Ali et al. 2023a, b). AI systems that explain the internals of an AI model in a manner that humans can comprehend are known as model intrinsic techniques (Adadi and Berrada 2018; Li et al. 2022; Das and Rad 2020; Carvalho et al. 2019; Cabitza et al. 2019; Lipton 2018; Lundberg and Lee 2017; Montavon et al. 2018; Saleem et al. 2022; Hassija et al. 2024).

1.3.5 Correctability

Correctability refers to the ability of a human actor to modify an AI system to ensure accurate decision-making (Schwalbe and Finzel 2023; Kulesza et al. 2015).

1.3.6 Comprehensibility

Similar to interpretability, comprehensibility involves both local and global justifications and functional understanding. Moreover, an understandable AI meets the criteria for effective interaction (Bruckert et al. 2020; Schmid and Finzel 2020). Interpretable presentation and intervention are viewed as crucial components for thorough comprehension and as prerequisites for comprehensibility (Schwalbe and Finzel 2023; Gleicher 2016).

1.3.7 Responsible XAI

Establishing trust and transparency is crucial for ensuring a model's reliability; however, to ensure responsibility, societal values, morals, and ethical considerations must also be considered. Therefore, Transparency, Responsibility, Accountability (Das and Rad 2020; Bogina et al. 2022; Smith 2021), Fairness, and Ethics (Bogina et al. 2022; Smith 2021; Lepri et al. 2021) are the fundamental principles underpinning Responsible AI.

1.3.8 Explainable artificial intelligence (XAI)

A collection of techniques and approaches designed to empower human users to comprehend, trust, and oversee AI outputs and decisions. Its objective is to enhance the transparency and comprehensibility of AI systems' decision-making processes, addressing the opaque nature often associated with sophisticated AI models (Viswan et al. 2024; Arrieta et al. 2020; Mavrepis et al. 2024; Longo et al. 2024; Weber et al. 2024; Martins et al. 2024; Madapatha and Fernando 2024; Clement et al. 2023; Buijsman 2022; Mavrepis et al. 2024; Zhou et al. 2023; Nizam and Zafar 2023; Borys et al. 2023; Kenny et al. 2021; Ali et al. 2023a, b; Nazir et al. 2023).

1.4 Organization of the paper

The primary aim of this survey is to present a comprehensive overview of recent developments in Model-Agnostic Explainable Artificial Intelligence (MA-XAI) techniques within the financial sector. By conducting a quantitative analysis, this study identifies the most frequently utilized MA-XAI methods in finance. The paper is structured as follows: Sect. 2 discusses the recent studies on XAI in finance. Section 3 provides the overview and applications of AI in Finance. Section 4 discusses about the limitations of AI and the emergence of XAI. Section 5 discusses the systematic literature review (SLR) approach. Section 6 presents the taxonomy of Explainable AI methods. Section 7 discusses in detail about the model-agnostic XAI (MA-XAI) methods in finance. Section 8 discusses about the quantitative analysis and research findings. Section 9 highlights the limitations and challenges in implementing MA-XAI methods in finance. Section 10 discusses the significance and impact of this survey. Section 11 outlines future research directions and provides an overall discussion of findings. Finally, Sect. 11 discusses about the discussions and future directions. Section 12 concludes the survey paper, as shown in Fig. 3.

2 Recent studies on XAI in finance (related works and comparative analysis)

Although extensive research has been conducted on AI applications in finance, studies focusing on XAI in finance in prominent international journals and conferences remain relatively limited. Key research areas include evaluating AI's trustworthiness in systemic risk

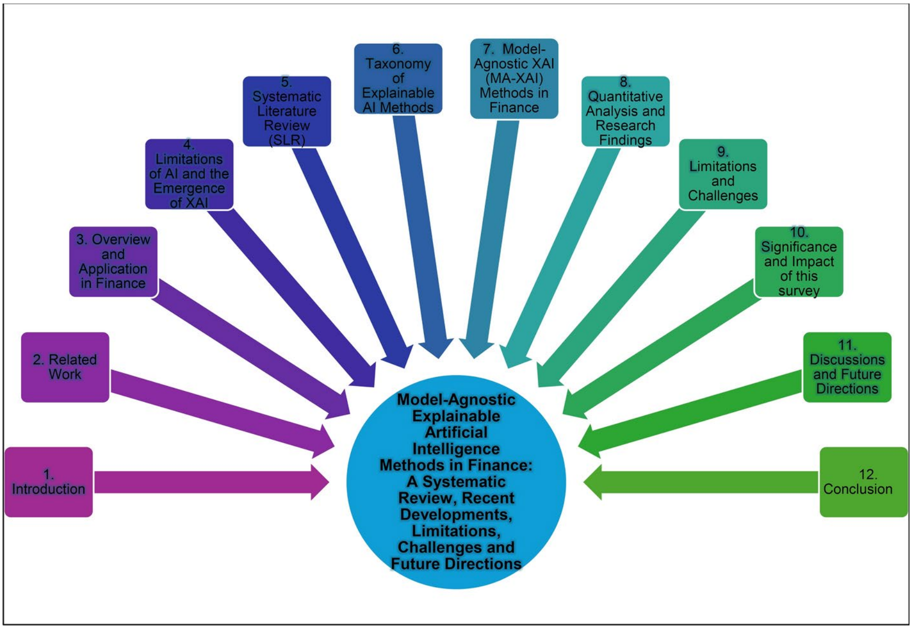


Fig. 3 Structural breakdown of the survey paper, outlining key sections and the logical progression of discussions on XAI in finance

assessment (Daníelsson et al. 2022), integrating DL with XAI for anti-money laundering frameworks (Kute et al. 2021), and designing smart markets that enhance human decision-making in complex trading environments (Bichler et al. 2010). Additionally, XAI plays a significant role in banking and financial services (Burgt 2020), particularly in credit scoring and risk management (Demajo et al. 2020; Biecek et al. 2021; Misheva et al. 2021). For instance, XAI has been leveraged to understand why policyholders purchase or discontinue non-life insurance coverage, enabling more precise policyholder segmentation and providing valuable insights into consumer behaviour (Gramegna and Giudici 2020). The application of XAI fosters trust among consumers and employees while ensuring accountability in AI-driven financial models (Rai et al. 2019; Martin 2017; Elliott et al. 2021).

Several studies have explored different methodologies for achieving explainability in finance-related AI models. For example, (Moore 1987) utilized the Classification and Regression Trees (CART) technique to introduce explainability through a hierarchical, transparent structure where decisions are made at internal nodes based on predefined conditions. Angelov et al. (2021) provided a historical overview of XAI, categorizing various methods and highlighting key applications in domains such as fraud detection and criminal justice. This study also emphasizes the relationship between DL and neuroscience and discusses future directions for bridging the gap between interpretability and model complexity. A systematic review conducted in (Islam et al. 2022) identified key application domains for XAI, by analysing 137 papers, including three in the financial sector. Moreover, (Malhi et al. 2020) combined LIME and Shapley values to enhance the interpretability of AI models, while (Mazhar and Dwivedi 2024) applied LIME to understand convolutional neural networks (CNNs) in social media sentiment classification. The use of XAI in financial market behaviour analysis was explored in (Benhamou et al. 2021; Ohana et al. 2021), where ML-based XAI models were employed to evaluate market dynamics and model performance. Furthermore, (Carta et al. 2022) examined how automatic feature selection in ML can improve financial forecasting, utilizing XAI-driven strategies to predict next-day stock returns.

The complexity and opacity of advanced AI models in finance necessitate the use of robust XAI techniques to enhance their transparency. Rane et al. (2023) evaluated various explainability methods, including rule-based systems, model-agnostic (MA) approaches, and interpretable ML models, to provide clear explanations for financial decisions. To aid future research, (Černevičienė and Kabašinskas 2022) classified multi-criteria decision-making methods to develop AI systems that are both explainable and interpretable for financial decision making. To enhance user trust, (Hanif 2021) proposed an interactive digital dashboard that visualizes XAI results and improves the interpretability for data scientists. Addressing concerns about AI's "black box" nature in financial assessments, (Meena and Mishra 2023) outlined future research directions on risk evaluation, transparency, and regulatory compliance in banking.

The application of XAI in financial distress prediction was investigated by (Zhang et al. 2022), who utilized SHAP, partial dependence plots, and counterfactual explanations to generate both local and global explanations for black-box models. Similarly, (Bhowmik et al. 2022) introduced a fraud detection methodology that leveraged nonlinear embedded clustering to address dataset imbalances, followed by a Deep Belief Network (DBN) for transaction analysis. This approach, which incorporates XAI, achieved an accuracy of 94% with a 70:30 training-validation split. The role of XAI in risk management for fintech applications was explored in (Bussmann et al. 2020), where Shapley values were used to interpret AI predictions for peer-to-peer lending. Çelik et al. (2023) proposed an XAI-driven approach, using LIME to assess prediction reliability, preventing erroneous decision-making in stock market forecasting using the KOSPI dataset. Additionally, (Freeborough and van Zyl 2022) evaluated the transferability of XAI methods for financial time-series prediction, applying techniques such as ablation, permutation, and integrated gradients to recurrent neural networks (RNNs), long short-term memory (LSTM), and gated recurrent unit (GRU) models trained on the S&P 500 data. The study found that GRU was the most effective in retaining long-term dependencies, whereas LSTM provided finer granularity by filtering out less relevant inputs.

Further analysis of XAI techniques in the fintech domain was conducted in (Gawantka et al. 2024), where methods such as LIME, SHAP, Contextual Importance and Utility (CIU), and Integrated Gradients (IG) were compared based on their similarities in model explanations. Meanwhile, (Ghosh and Dragan 2023) proposed hybrid predictive frameworks by combining Empirical Mode Decomposition (EEMD) with LSTM and Facebook's Prophet Algorithm, utilizing permutation feature importance and LIME to uncover financial stress patterns. In the banking sector, (Huang et al. 2024) employed ML and XAI to examine the complexity and opacity of financial models and identified significant correlations between firms and industries. Finally, (David et al. 2021) explored how different sources of advice (human vs. AI-based) and the presence of local and global explanation labels influence consumers' trust and willingness to adopt AI-driven financial consulting.

Recent studies highlight its growing role in investment strategies, where SHAP-based feature attribution improves risk-return trade-offs (Yan and Li 2024), and hybrid XAI models enhance asset allocation and risk mitigation (Han and Li 2023). In credit risk assessment, SHAP and LIME have been used to enhance loan approval transparency and fairness (Nallakaruppan et al. 2024), whereas DL-based credit-scoring models integrate XAI techniques to reduce bias (Schmitt and Cummins 2023). For fraud detection, SHAP-enhanced ML models can improve regulatory compliance and enhance financial transparency (Thanathamathee et al. 2024). Additionally, LLMs are being explored in financial risk analysis (Tao et al. 2024) where explainability techniques such as LIME and counterfactual explanations enhance interpretability (Zhao et al. 2024a). In financial forecasting, XAI methods such as permutation feature importance and integrated gradients improve the interpretability of models for stock market prediction (Kumar et al. 2024). Future research is focusing on hybrid XAI frameworks that integrate rule-based explanations with DL architectures to enhance both interpretability and accuracy (Saw et al. 2025). These advancements highlight XAI's growing significance in ensuring transparency, regulatory compliance, and model reliability in financial AI systems.

2.1 Comparison with other work

2.1.1 Peer-to-peer lending

Babaei et al. (2023) investigated explainable fintech lending, particularly focusing on peer-to-peer lending platforms. They emphasized local interpretability and the importance of SHAP and LIME for explaining credit decisions. While their work provides an in-depth focus on a specific financial area, our study broadens this perspective by systematically reviewing MA-XAI techniques in various financial applications beyond lending, such as portfolio management, risk assessment, and trading, thus offering a more holistic analysis of XAI's role in finance.

2.1.2 Crypto asset management

Babaei et al. (2022) provided insights into XAI for crypto asset allocation, utilizing methods like SHAP to enhance transparency in investment decisions. While their analysis contributes significantly to asset management, particularly crypto assets, our systematic review extends their findings by including broader financial applications such as fraud detection, credit scoring, and algorithmic trading. Furthermore, our review evaluates a wider range of MA-XAI methods, offering comparative insights into their scalability and interpretability across different financial scenarios.

2.1.3 Cyber risk management

Calzarossa et al. (2025) addressed explainability robustness in ensemble machine learning methods specifically for cyber risk management. They critically assessed ensemble-based explanations' robustness, emphasizing the reliability and consistency of explainability methods. Our paper complements their findings by highlighting broader limitations of MA-XAI methods related to scalability, interpretability, and computational efficiency across diverse financial datasets and contexts. We further propose hybrid solutions and optimizations to address these concerns, extending their discussion into a broader financial framework beyond cybersecurity.

2.1.4 Financial time-series prediction

Giudici et al. (2024) explored explainable AI methods tailored specifically for financial time-series predictions, highlighting the challenges related to temporal dynamics and the limitations of existing interpretability methods like SHAP and LIME. While their paper extensively analyzed time-series contexts, our review synthesizes these insights and integrates additional financial applications and MA-XAI methods. We further discuss global interpretability and ethical considerations, providing a more comprehensive and interdisciplinary understanding of XAI's potential and limitations in finance.

2.1.5 Connection with SAFE AI literature

Our systematic review also aligns with recent efforts to establish SAFE (Sustainable, Accountable, Fair, Explainable) machine learning practices in finance, as presented by Babaei et al. (2025). Their proposal of a "Rank graduation box" emphasizes safety and fairness metrics for AI-driven financial decisions. We extend these discussions by reviewing multiple MA-XAI methodologies that enhance transparency and regulatory compliance. By integrating ethical AI practices, fairness-aware techniques, and computational optimizations, our work explicitly contributes to the ongoing efforts toward SAFE AI frameworks in finance.

XAI techniques can be classified into model-specific (MS) and model-agnostic (MA) approaches. MS methods focus on interpretability within specific AI architectures (Fontes et al. 2024; Ahmed et al. 2022; Schwalbe and Finzel 2023), whereas MA methods provide broader applicability across various ML models (Owens et al. 2022; Gianfagna and Di Cecco 2021; Ribeiro et al. 2016a). Figure 11 provides an overview of the different XAI methods and their corresponding AI categories. The classification of XAI techniques in financial applications was further examined in (Černevičienė and Kabašinskas 2024) where articles were grouped based on the financial tasks they addressed, variations in XAI methodologies, and their implementation in different domains. Model agnosticism in XAI refers to techniques that can be applied across diverse ML models without being constrained by a particular architecture (Letrache and Ramdani 2023; Martins et al. 2024; Ribeiro et al. 2016a) making them highly versatile and widely applicable in financial analysis.

3 Artificial intelligence in finance: overview and applications

3.1 AI and ML: definitions and context

Artificial Intelligence (AI) encompasses the development of computational systems capable of performing tasks typically requiring human intelligence, including reasoning, decision-making, learning, and problem-solving (Bahoo et al. 2024; Jain et al. 2024). Within AI, Machine Learning (ML) specifically refers to algorithms that improve automatically through experience and data exposure, enabling systems to identify patterns and make data-driven predictions or decisions without explicit programming. In the financial sector, the integration of AI and ML has significantly transformed areas such as credit risk assessment, fraud detection, stock market prediction, and investment strategy formulation. AI's capability to analyze vast datasets rapidly and accurately has facilitated predictive analytics and informed decision-making, driving efficiency and precision within financial services (Varadarajan and Priya 2024; Eluwole and Akande 2022; Mishra et al. 2024; Jain et al. 2024; Bahoo et al. 2024).

3.2 AI applications in financial decision-making

The integration of AI/ML techniques into the financial sector has significantly enhanced various financial tasks. AI has revolutionized industries by automating complex tasks, enhancing decision making, and improving efficiency (Rahim and Chishti 2024). In finance, AI powers credit scoring, fraud detection, portfolio management and stock market prediction. Its ability to process large datasets, identify patterns, and generate predictive insights has transformed financial services, enabling faster, more accurate, and transparent operations.

3.2.1 AI and the stock market

AI has transformed the stock market by enabling real-time data analysis, predictive modeling and automated trading. ML algorithms can be used to forecast stock prices, detect market trends, and optimize investment strategies. Figure 3 shows the general outlook for the impact of news and social media on the stock market, and the experimental results indicate that the highest prediction accuracies of 80.53 and 75.16% are obtained using social media and financial news, respectively (Khan et al. 2022a, b, c). This has increased trading efficiency, reduced human error, and enhanced investors decision-making. Dixon et al. (2017) investigated that deep neural networks (DNNs) demonstrated strong predictive power with 68% accuracy. Zhang et al. (2021) shows that long short-term memory (LSTM) networks surpass traditional ANNs in accuracy and efficiency, especially when incorporating online investor attention metrics such as Internet search volume. Ozbayoglu et al. (2020) used an LSTM model for stock price forecasting and trading signals, achieving 91.5% accuracy, which surpassed traditional moving average strategies. Wang et al. (2021) used a sequence-to-sequence model to predict market trends with 85% accuracy, enhancing trading algorithms and enabling real-time dynamic trading strategies. Huang (2018) designed a deep reinforcement learning model for trading, achieving 92% precision and higher cumulative returns than conventional strategies, enabling adaptive and autonomous trading agents.

3.2.2 AI in fraud detection

DL has revolutionized fraud detection by identifying complex patterns in large transaction datasets. Models such as CNNs, RNNs (Recurrent Neural Networks), and autoencoders excel at detecting nonlinear and temporal fraud patterns in real time (Mienye and Sun 2023). Payment processors such as PayPal and Visa use these models to enhance detection accuracy and reduce false positives (Din et al. 2021). Jurgovsky et al. (2018) used LSTM networks for credit card fraud detection, achieving an F1-score of 0.93, surpassing traditional models such as RF and logistic regression (F1-score 0.85). Gandhar et al. (2024) developed a DL model for detecting financial transaction anomalies, effectively reducing false positives to minimize disruptions to legitimate transactions. Talukder et al. (2024) proposed an Integrated Multistage Ensemble Machine Learning (IMEML) model using classifiers such as EIC, EBC, and EMC, combined with data balancing techniques such as IHT+EMC, CC, and RUS. On a credit card dataset of 284,807 transactions, our model achieved an accuracy, precision, recall, F1-score, and AUC of 99.94%, 99.91%, 99.14%, 99.52%, and 100%, respectively. Studies such as “Fraud detection in publicly traded US firms using Beetle Antennae Search” and “Fraud detection in capital markets: A novel ML approach” (Khan et al. 2022a, b, c) present optimization-driven and ML-based fraud detection mechanisms, emphasizing their importance for financial security. Given the regulatory sensitivity of fraud detection, integrating XAI techniques into fraud detection models is crucial for ensuring accountability and compliance. Explainability techniques such as SHAP, LIME, and Counterfactual Explanations can enhance fraud detection models by identifying key transaction features associated with fraudulent behavior while ensuring that AI-driven anomaly detection systems align with compliance and forensic accounting requirements (Kapale et al. 2024). Future research should explore MA-XAI frameworks tailored for financial fraud detection, ensuring interpretability, regulatory alignment, and fairness in fraud risk modeling.

3.2.3 AI and portfolio management

AI enhances portfolio management by automating asset allocation, risk assessment, and investment strategy optimization. It analyzes historical data and market trends using ML models to predict the performance of assets. This enables more efficient data-driven decision-making to maximize returns and minimize risk. Soleymani and Vasighi (2022); Zhao et al. (2018) used a clustering approach combined with value-at-risk (VaR) analysis to enhance asset-allocation strategies. highlight that the asymmetric copula method for estimating return dependencies enhances the portfolio optimization process. Most studies indicate that AI-based prediction models significantly enhance the portfolio selection process by accurately forecasting the stock returns. Ye et al. (2020) developed a reinforcement learning model for portfolio management that adapts to market changes by learning from historical data, enabling dynamic investment strategies. Jiang and Liang (2016) used a GAN-based model for cryptocurrency portfolio optimization, outperforming traditional methods. GANs generate synthetic market scenarios thereby enabling strategy testing under various market conditions, which is essential for volatile assets such as cryptocurrencies. Shi et al. (2021) developed a DL framework that customizes investment strategies based on individual preferences and risk tolerance, integrates reinforcement learning for real-time asset allocation optimization, and showcases DL's potential for personalized investment solutions. Recent studies have explored Beetle Antennae Search (BAS)-based portfolio optimization techniques, including Quantum BAS, Non-linear Activated BAS, and Quadratic Interpolated BAS, which effectively address non-convex constraints, transaction costs, and tax-aware asset allocation (Khan et al. 2022a, b, c). Works such as "Optimal portfolio management for engineering problems using nonconvex cardinality constraints" (Khan et al. 2020) and "Time-varying mean–variance portfolio selection under transaction costs" (Katsikis et al. 2021) highlight the role of intelligent search algorithms in optimizing financial portfolios under real-world constraints. Integrating model-agnostic explainability techniques into these metaheuristic-driven optimization models can provide insights into portfolio rebalancing decisions, risk exposure, and factor-based investment strategies. Additionally, neural network-based portfolio management techniques, including recurrent neural networks (RNNs) and decomposition-based neural dynamics approaches, have emerged as powerful tools for optimizing risk-return trade-offs in high-frequency trading and asset allocations. Studies such as "Neural Networks for Portfolio Analysis in High-Frequency Trading" (Cao et al. 2024) and "Artificial Neural Dynamics for Portfolio Allocation" (Cao et al. 2025) introduce data-driven methods for adaptive portfolio optimization, where explainability is essential for understanding how AI-generated allocations align with investors' risk profiles.

3.2.4 AI and performance, risk, default valuation

AI enhances performance, risk assessment, and default valuation in finance by analyzing large datasets for accurate predictions. ML models assess credit risk, forecast defaults and optimize investment portfolios. This enables better decision-making, reduces uncertainty, and supports more resilient financial strategies than the traditional methods. Jones et al. (2017) and Gepp et al. (2010) assess corporate default probabilities, while (Popa et al. 2021) predict business performance using a composite financial index. These studies confirm that AI-powered classifiers are highly accurate and interpretable, outperforming traditional linear models. Feldman and Gross (2005); Episcopos et al. (1998) studied mortgage and loan default prediction. A study on the Malaysian and Islamic banking sectors using NN models finds that factors such as negative cost structure, cultural aspects, and regulatory barriers contribute to inefficiency, whereas U.S. banks are more resilient, healthier, and better regulated (Papadimitriou et al. 2022).

3.2.5 AI and credit risk assessment in the banking sector

AI is revolutionizing credit risk assessment in banks by using ML to predict loan defaults and evaluate borrower risk. It analyzes extensive data to improve credit scoring and decision making. This leads to better risk management, reduced default, and enhanced lending efficiency. The first substream focuses on predicting bank failures, with ML and ANNs outperforming traditional statistical methods, although they lack transparency (Le and Viviani 2018). To address this, Durango-Gutiérrez et al. (2021) combined logistic regression with AI models such as MLP, offering better insights into explanatory variables. AI-based models have significantly enhanced financial decision-support systems (FDSSs). This approach is crucial for preventing future global financial crises (Abedin et al. 2019). Shi et al. (2022) reviewed 76 key studies from the past eight years on credit risk using statistical, ML, and DL techniques, proposed a classification method for ML-based credit risk models, ranked their performance, and discussed challenges such as data imbalance, model transparency, and limited DL use (Lahmiri 2016; Khandani et al. 2010). The second substream compares classic and advanced consumer credit risk models. Supervised learning tools, such as SVM, RF, and decision trees, can predict credit card delinquency up to 12 months in advance. Abedin et al. (2019) proposed an LVQ neural network, improving accuracy with categorical variables and offering 6–25% cost savings over logit-based methods.

The last group focuses on intelligent credit scoring models, with ML systems such as Adaboost and RF providing the best forecasts for credit rating changes. These models are robust to outliers, missing values, and overfitting, and require minimal data intervention (Jones et al. 2015). Xu et al. (2019) combined data mining and ML to build an advanced model that selects key predictors and eliminates noisy variables. Xiao et al. (2024) proposed a DNN for credit scoring, achieving 20% higher predictive accuracy than FICO scores with an AUC of 0.92, and capturing nonlinear variable interactions for better credit assessment. Figure 3 shows the author reviewed DL model applications across seven Finance & Banking domains focusing on feasibility through data preprocessing, inputs, and evaluation criteria. The authors also identified the optimal DL models for each domain (Huang et al. 2020).

3.2.6 AI in foreign exchange management

AI in foreign exchange management optimizes trading strategies, forecasts currency fluctuations, and automates decision-making. ML algorithms analyze data to predict market movements and execute trades. AI models, such as neural networks (NNs) and reinforcement learning, improve accuracy, reduce errors, and enhance risk management in forex trading. Cost-effective trading in Forex requires accurate exchange- rate forecasts (Galeshchuk and Mukherjee 2017). The HONN model outperforms traditional NNs in forecasting the EUR/USD pair using ECB data (Dunis et al. 2013). However, (Galeshchuk and Mukherjee 2017) found these methods ineffective for predicting forex rate changes and instead used DNNs to forecast EUR/USD, GBP/USD, and JPY/USD, outperforming time-series models such as ARIMA. Overall, AI-based models such as NARX provide better prediction performance than statistical models (Amelot et al. 2021).

3.2.7 Investor sentiment analysis using AI

Applies ML and NLP (Natutal language processing (NLP) to analyze financial news, social media, and reports, identifying positive, negative, or neutral sentiment. This helps predict stock movements, asset prices, and market volatility. AI uncovers insights from unstructured data, enabling informed investment decisions and effective risk management. Investor sentiment is crucial for stock prediction, with sentiment analysis using NLP and data mining on platforms such as StockTwits and Yahoo Finance. It is used to forecast asset price direction, stock liquidity, and intraday returns (Yin et al. 2022). Sentiment is positively correlated with stock liquidity, especially in slow markets, and affects stock returns, particularly around major events, such as earnings announcements (Houlihan and Creamer 2021; Heston and Sinha 2017).

3.2.8 Financial document analysis and information extraction

This method uses techniques such as NLP, Optical Character Recognition, and DL models (e.g., RNNs, CNNs, Transformers) to automate the extraction of key data from financial texts, improving efficiency, accuracy, and scalability in financial analysis, fraud detection, and compliance. Memon et al. (2020) conducted an extensive literature review with OCR to analyze scanned financial documents and convert images into text for information extraction. This integration helps automate compliance and reporting, thereby reducing errors. Yang et al. (2020) developed FinBERT, a model fine-tuned on financial texts for better sentiment analysis and risk assessment. Montariol et al. (2024) proposed a multitask BERT model for extracting features from financial reports, improving task performance and generalization. Moirangthem and Lee (2021) used GRUs with a hierarchical structure for financial text classification, enhancing accuracy by focusing on relevant content.

3.2.9 Large language models (LLMs) in finance

The increasing adoption of LLMs in financial AI has introduced novel applications in automated financial analysis, regulatory reporting, sentiment analysis, and decision support systems. Works such as "Empowering Financial Futures: Large Language Models in the Modern Financial Landscape" (Cao et al. 2024) illustrate the growing role of LLMs in financial intelligence, leveraging vast textual datasets for market trend analysis and automated financial advisory services. However, the integration of LLMs into financial decision-making introduces new challenges related to the explainability, bias detection, and interpretability of generated financial insights (Zhao et al. 2024b). Given the opaque nature of LLM-based decision models, MA-XAI techniques can be instrumental in enhancing their trustworthiness by providing transparent explanations of AI-generated financial insights. Integrating AI-driven techniques into finance has significantly enhanced decision-making in areas such as portfolio optimization, risk management, and fraud detection. The use of optimization algorithms and neural networks has improved predictive accuracy, but the lack of transparency remains a challenge. Incorporating model-agnostic explainability techniques such as SHAP, LIME, and counterfactual explanations can provide deeper insights into these models. In portfolio optimization, explainability helps investors understand AI-driven asset selection and risk-return trade-offs. In risk management, interpretable AI aids in credit scoring, stress testing, and regulatory compliance by offering clear justifications for risk assessments. Similarly, in financial anomaly detection, explainability techniques enhance fraud detection by identifying key contributing factors in suspicious transactions. Expanding the survey to include these aspects would not only provide a more comprehensive view but also increase its practical relevance for financial analysts and policymakers. To enhance the impact and relevance of this survey, future research should also focus on how XAI enhances AI-driven portfolio optimization by ensuring interpretability in asset selection and rebalancing, while in trading systems, it clarifies risk-return tradeoffs. For fraud detection, MA-XAI improves the transparency of anomaly identification and transaction monitoring. In financial LLM applications, XAI ensures transparency in sentiment analysis, risk assessment, and compliance monitoring.

4 Limitations of AI and the emergence of XAI

4.1 Limitations of black-box AI models

The use of AI models is limited by several factors. The foremost among these is the lack of transparency in the internal workings of the network, which makes it difficult to understand how the model reaches its conclusions (Cremer 2021; Sarker 2021). These models are considered black-box models because they lack the ability to provide understandable explanations for the predictions they generate, leading to ambiguity in decision making (Garg et al. 2021; Rai 2020). NNs show impressive results but operate as black boxes (van der Velden et al. 2022), because of their inability to offer clear, justifiable explanations for the predictions they produce which is commonly known as interpretable DL or XAI (Adadi and Berrada 2018; Murdoch et al. 2019). They mimic human behaviour but update weights and biases through gradient descent, lacking full understanding, which limits the control and explanation of their operations (Ali et al. 2023a, b).

Figure 4 shows the working of the general typical AI model and XAI model. Such black-box models frequently result in ambiguous situations, prompting questions like “Why did you classify this as class X instead of class Y?”, “When will you succeed or fail?”, “How can incorrect feature selection be corrected?”, “Which dominant feature are you focusing on to train the model?”, “Can I trust the prediction you provided?” and similar studies (Yang et al. 2022).

4.2 Explainable AI: concepts and importance

Explainable Artificial Intelligence (XAI) refers to methodologies and techniques aimed at making machine learning and AI models understandable and transparent to humans (Kalasampath et al. 2025). Unlike traditional “black box” models, XAI provides clear explanations regarding the rationale behind model predictions or decisions. It achieves this by revealing feature contributions, decision logic, and causal relationships within complex algorithms. XAI has emerged to address concerns about AI algorithm transparency, offering tools and frameworks to help humans understand AI model operations, which is particularly crucial in fields such as finance, medical science and defence where transparency is critical for patient safety (Weber et al. 2024; Ali et al. 2023a, b; Clement et al. 2023; Mavrepis et

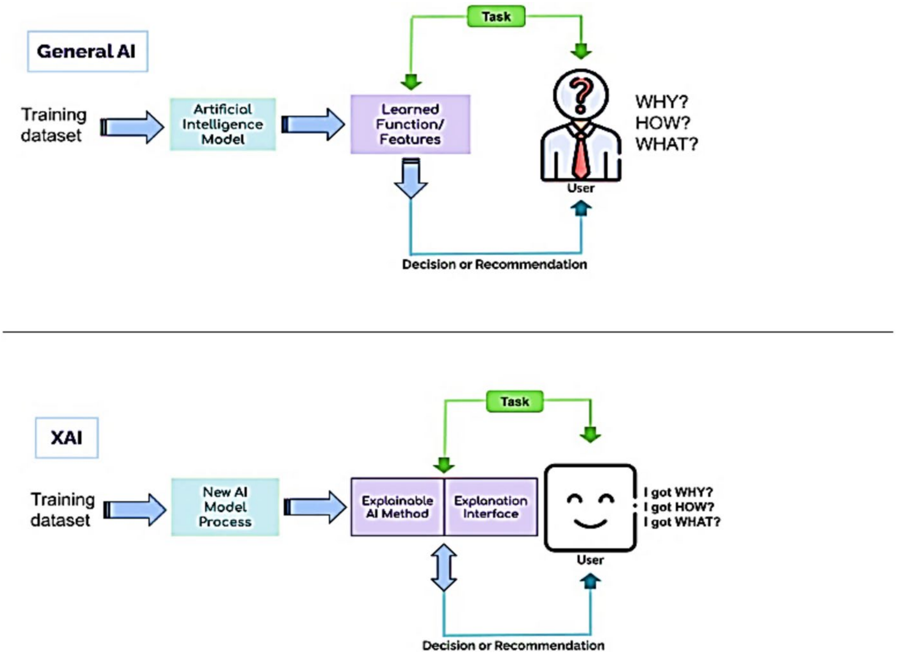


Fig. 4 Comparison between traditional AI (black-box) models XAI models, emphasizing the need for interpretability in high-stakes applications such as finance and healthcare

al. 2024; Nizam and Zafar 2023; Kenny et al. 2021; Yeo et al. 2023; Holzinger et al. 2022; Lamberti 2023). Figure 5 presents the distribution of XAI applications in finance.

(Lundberg and Lee 2017) described explainability as the "interpretable approximation of the original complex [AI] model". XAI encompasses methods that empower stakeholders (Tomsett et al. 2018) to gain a deeper understanding of AI algorithms and their decision-making processes. An AI system is deemed explainable if its task model is intrinsically interpretable (where the AI system serves as its own task model) or if a non-interpretable task model is accompanied by an interpretable and accurate explanation (where the AI system integrates a post-hoc explanation; Markus et al. 2021). XAI methods can mitigate the challenges related to adoption and implementation, allowing regulated industries, such as finance, to fully leverage the potential of automation.

4.3 Challenges of financial AI models and the need for XAI

While AI has significantly transformed financial decision-making by improving risk assessment, fraud detection, and predictive modelling, its increasing complexity raises critical concerns regarding trust, accountability, and regulatory compliance. The inherent opacity of complex AI models, such as deep learning algorithms, limits their interpretability, raises regulatory compliance issues, and undermines stakeholder trust. Financial regulators, institutions, and customers demand transparency to ensure fairness, accountability, and regulatory compliance, making Explainable AI (XAI) essential (Kalasampath et al. 2025).

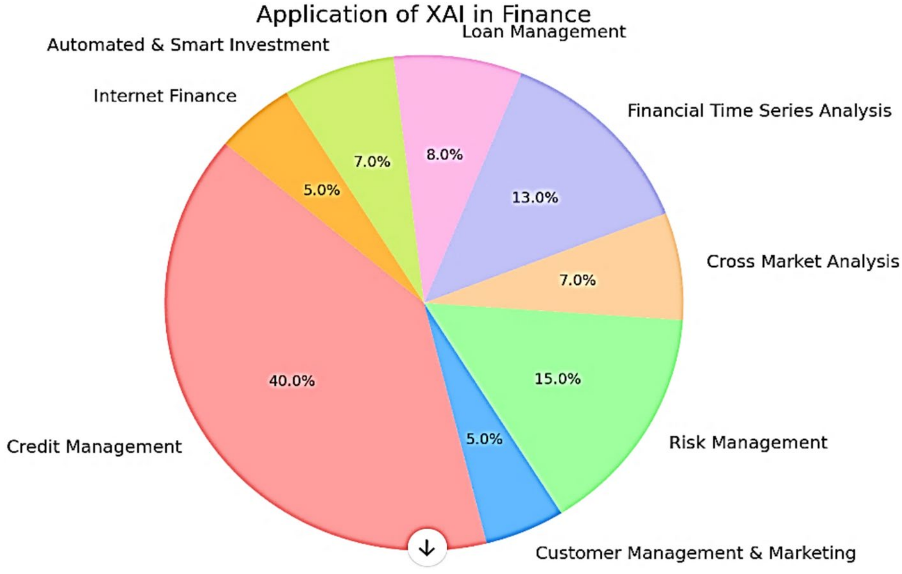


Fig.5 Percentage-wise distribution of XAI techniques across different financial applications, including credit scoring, fraud detection, and risk management

Many financial AI models function as black boxes, making it difficult for stakeholders, including regulators, investors, and consumers, to understand and validate decisions. This opacity introduces risks such as biased lending decisions, market manipulation, and regulatory non-compliance, necessitating the use of XAI techniques to enhance transparency. Certain AI architectures are more challenging to interpret than others, requiring advanced XAI techniques to ensure their reliability in financial applications.

4.3.1 Long short-term memory (LSTM) networks

LSTMs are extensively used in time-series forecasting for stock price prediction, credit risk modelling, and volatility analysis. Their reliance on hidden states and long-term dependencies makes decision interpretation difficult, particularly in financial contexts where explainability is crucial. Techniques such as Layer-wise Relevance Propagation (LRP) and attention-based visualization can help highlight which past time steps contribute most to the model's predictions, improving interpretability (Park and Yang 2022).

4.3.2 Generative adversarial networks (GANs)

GANs are increasingly being applied to fraud detection, synthetic financial data generation, and anomaly detection. Their adversarial training framework makes them inherently difficult to explain, as decisions emerge from a competitive learning process between the generator and discriminator. Shapley values (SHAP) and Integrated Gradients can help uncover feature importance, allowing stakeholders to detect biases in synthetic data and ensure fairness in AI-driven financial systems (Choi and Kim 2024).

4.3.3 Transformers (e.g., BERT, GPT-based models)

Transformers are widely used in NLP-based financial analytics, credit scoring, document classification, and sentiment analysis. Their self-attention mechanism enables powerful contextual learning but creates highly nonlinear feature interactions, making it difficult to determine the factors that influence predictions. Explainable Attention mechanisms, SHAP, and Feature Importance Analysis can help identify the most influential words or phrases that affect financial model decisions (Govindaraj et al. 2023).

4.3.4 The goal of XAI in bridging the gap between AI and human understanding

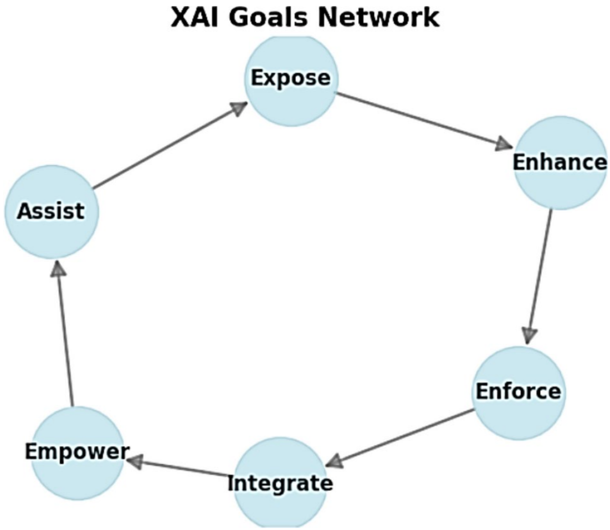
The key objective of XAI is to create models that are interpretable by humans, which is particularly crucial in sensitive fields such as banking, healthcare, and defence. Domain experts need these models to solve problems more effectively and receive outputs that they can understand and trust. It benefits not only specialists by providing meaningful outputs but also developers, as any incorrect output prompts system investigation and improvement. AI methods facilitate (i) the assessment of existing knowledge, (ii) the progress of knowledge, and (iii) the development of new hypotheses and theories (Rieg et al. 2020). XAI also aims to achieve enhanced justification, control, improvement, and discovery (Adadi and Berrada 2018). The following points summarize the benefits of making black-box systems more transparent (Guidotti et al. 2019a, b), as shown in Fig. 6.

- This will enable individuals to tackle the adverse effects of automated decision-making.
- This will aid individuals in making more informed decisions.
- It can detect and safeguard against security vulnerabilities.
- Align algorithms with human values.
- Raise industry standards for developing AI-powered products, thereby boosting consumer and business confidence.
- Enforce the Right of Explanation Policy.

4.4 Trade-off between performance accuracy and explainability

A trade-off often exists between model accuracy and associated explainability (Herm et al. 2023). Balancing model accuracy and explainability is a persistent challenge in AI. Simple models, such as linear regression and decision trees, are easy to interpret but may sacrifice predictive power. In contrast, complex models, such as CNNs, excel in accuracy but are less transparent in their decision-making processes (Jung et al. 2021). This trade-off is crucial, especially in healthcare, where both precision and explainability are vital for patient trust and safety, as illustrated in the Fig. 7. Advances in post hoc interpretability are critical for bridging this gap and ensuring the accuracy and understandability of AI models that are accurate and understandable across various applications (Bauer et al. 2021). The ideal solution should have both high explainability and performance (Yang et al. 2022; Viswan et al. 2024; Love et al. 2023; Swathi and Challa 2023; Raees et al. 2024).

Fig. 6 Goals of XAI



5 Systematic literature review (SLR) approach

In this segment of the analysis (Fig. 4), the guidelines for systematic reviews and meta-analyses outlined by the pertinent authorities were strictly adhered to (Kitchenham and Charters 2007; Kitchenham 2007). Figures 8 and 9 illustrates the number of articles selected per year and published country-wise, where India has published the highest number of articles in this domain, followed by the United States and Germany.

5.1 Search strategy and initial screening

- A comprehensive search was conducted using domain-specific keywords such as “Explainable AI in Finance,” “XAI for Credit Scoring,” “Interpretable AI in Banking,” “XAI in Financial Risk Management,” and “Financial Market Predictions with XAI.”
- To ensure a rigorous and transparent selection process, we employed a multistage filtering approach to retrieve relevant studies from IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, Web of Science, and Google Scholar. Our methodology was designed to systematically identify high-quality research on XAI in financial applications, ensuring both comprehensiveness and methodological rigor, as shown in Fig. 10.
- Boolean operators (AND/OR) were used to refine the search results and ensure interdisciplinary coverage.
- The initial search yielded 1,115 articles published between 2010 and July 2024.

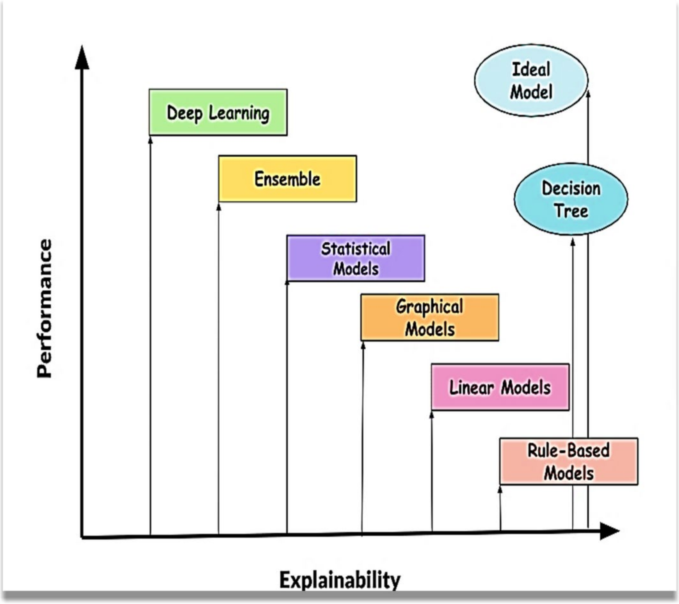


Fig. 7 Visualization of the trade-off between model explainability and performance accuracy, demonstrating the balance between interpretability and predictive power in AI models

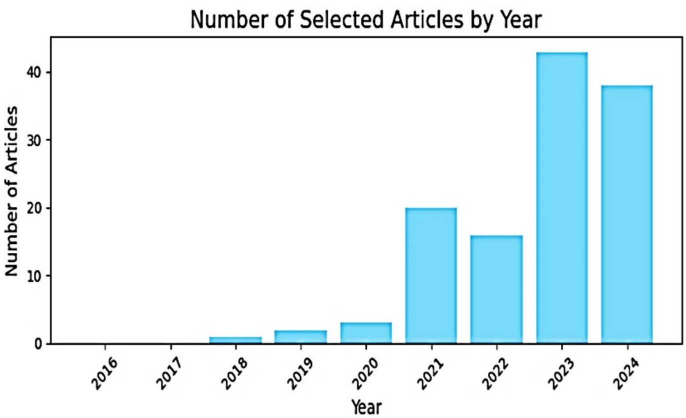


Fig. 8 Number of articles published year-wise

5.2 Automated filtering and duplicate removal

- Duplicate entries and records flagged as ineligible by automation tools were removed, along with studies marked as irrelevant based on metadata analysis.
- After filtering, 370 articles remained, eliminating 795 non-relevant studies from the dataset.

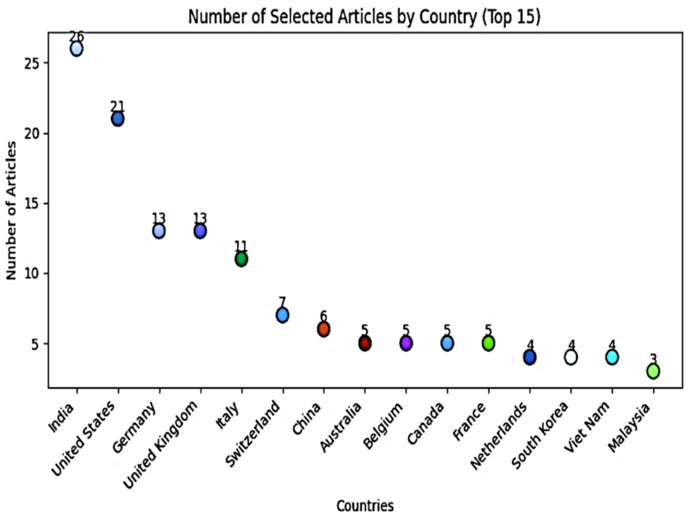


Fig. 9 Geographical distribution of research publications on XAI in finance, highlighting the leading contributors in this domain

5.3 Title and abstract review

- A secondary screening phase was conducted to evaluate each paper's relevance by reviewing the titles and abstracts.
- Studies that did not explicitly focus on XAI in financial applications, lacked explainability methodologies, or addressed non-financial AI use cases were excluded from the review.
- As a result, 130 additional papers were removed, leaving 240 articles for an in-depth evaluation.

5.4 Full-text analysis and final selection

The remaining studies underwent a comprehensive full-text review, in which we assessed the following:

- Empirical validation and real-world applications are discussed.
- Relevance to XAI and financial decision-making.
- The contribution to explainability and model transparency.
- Publications in high-impact journals or top-tier conferences.

Based on these criteria, 150 high-quality studies were selected for inclusion in the final dataset.

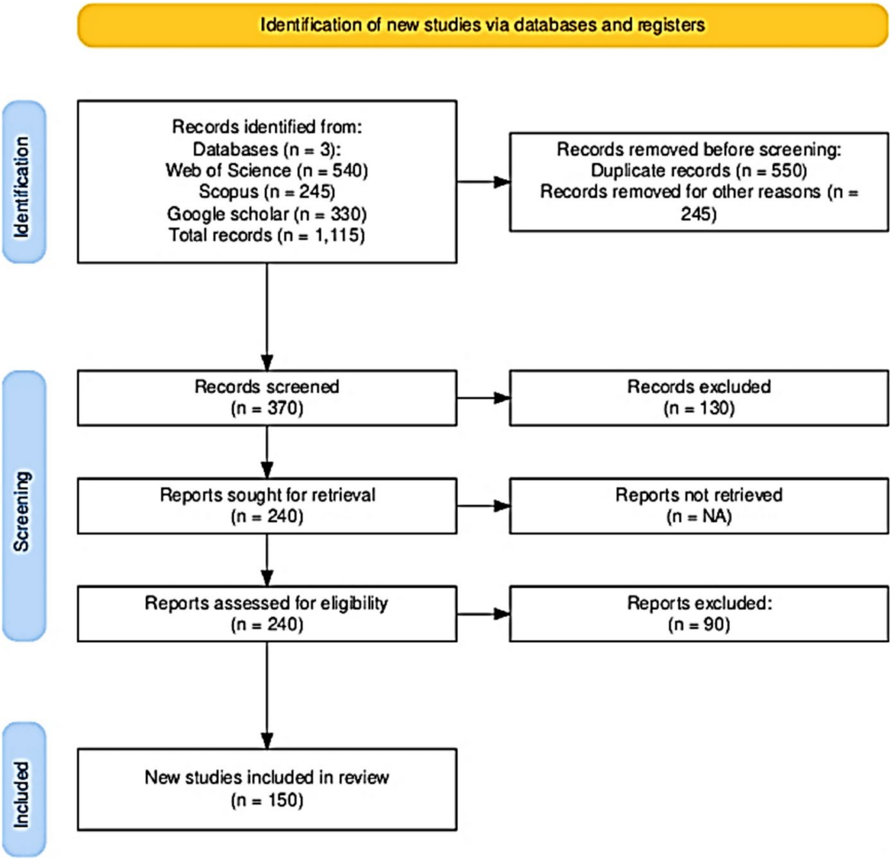


Fig. 10 Systematic literature review (SLR) methodology following the PRISMA framework, detailing the selection process for research articles included in this study

6 Research questions (RQ)

The primary objective of this study was to identify advanced technologies, algorithms, evaluation methodologies, and datasets related to XAI in the finance sector. To perform a comprehensive systematic mapping review, the main research question was divided into several specific inquiries, as detailed in Table 1. These questions aimed to offer a detailed framework for the study, facilitating a clear understanding of its organization and focus.

7 Taxonomy of explainable AI methods

In this section, we provide a concise overview of the XAI techniques used in AI for financial domain analysis. Detailed comprehensive surveys dedicated exclusively to XAI are presented in (Adadi and Berrada 2018; Murdoch et al. 2019). We differentiate XAI techniques using three criteria: MS versus MA, global versus local (scope of the explanation),

Table 1 Research questions

RQ#	Research questions
RQ1	Which MA-XAI techniques or methods are frequently investigated/applied by researchers in the context of the financial domain?
RQ2	Which XAI framework has been widely used by the researchers in studying the financial datasets while ap-plying MA-XAI techniques?
RQ3	Which AI/ML/DL algorithms have researchers princi-pally employed in the analysis of financial datasets when applying MA-XAI methods?
RQ4	Which datasets are most commonly and widely used in research that focuses on MA-XAI methods for analysis??
RQ5	What are the different performance metrics examined in the research context to MA-XAI methods specifically concerning financial domain?

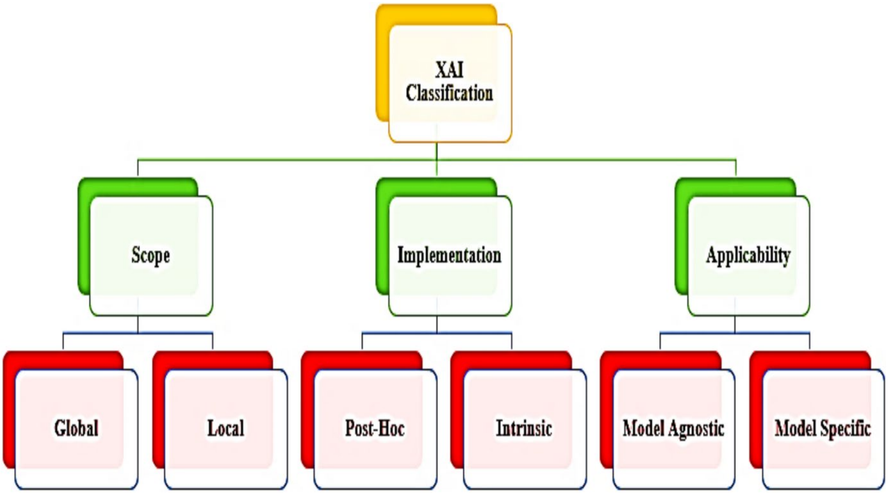


Fig. 11 Taxonomy of XAI methods categorizing different approaches based on their applicability in financial AI

and model-based versus post hoc. This framework, adapted from (Adadi and Berrada 2018; Murdoch et al. 2019), is depicted in the Fig. 11. The following sections explain these criteria.

7.1 Model-specific vs. model-agnostic methods

7.1.1 Model-specific (MS) explanation

MS explanation methods are tailored to classes of models, such as specific types of NNs. This limitation can restrict the choice of NNs, possibly excluding better-fitting NNs. Model-based explanations are inherently MS (Adadi and Berrada 2018), but not all MS explanations are model-based. For instance, some post hoc saliency mapping techniques are specific to certain CNNs but are not considered model-based explanations (Murdoch et al. 2019).

7.1.2 Model-agnostic explanation

MA explanation does not depend on the type of neural network and operates solely on its input and output. By altering the input, users can observe changes in the output, revealing which regions influence the outcome.

Evidence of MS and MA methods can be found in the literature (Olden et al. 2004; Olah et al. 2017; Zeiler and Fergus 2014; Siami et al. 2021; Neumann et al. 2019; Adadi and Berrada 2018; Islam et al. 2022; Linardatos et al. 2020; Sahakyan et al. 2021; Lin et al. 2021; Speith 2022; Molnar et al. 2023).

7.2 Scope of explanation

7.2.1 Global explanation

Global explanation or dataset-level explanation reveals the overall relationships learned by the neural network. It can provide feature importance scores across the entire dataset, such as indicating how much high blood pressure increases the risk of cardiac events. It also includes visualizing the learned filters to show which features the network extracts and their relevance to the task (Olden et al. 2004; Olah et al. 2017; Zeiler and Fergus 2014). The authors in (Siami et al. 2021; Neumann et al. 2019; Kwak et al. 2021; Kašćelan et al. 2016; Jain et al. 2019; Guelman 2012; Devriendt et al. 2021; Kwak et al. 2021; Carfora et al. 2019; Baecke and Bocca 2017; Xiao and Benbasat 2007; Jeong et al. 2018; Gramegna and Giudici 2020; Karamizadeh and Zolfagharifar 2016) used the global explanation concept of XAI in the analysis of their AI model used for the prediction or recommendation.

7.2.2 Local explanation

Local explanation focuses on a single input. For instance, in assessing cardiac risk, it explains why blood pressure is significant for an individual's risk, unlike the global explanation, which covers the entire dataset. Another example is a saliency map highlighting a brain tumor on an MRI, showing which part of the image influenced the 'tumor' classification for that specific person. Local interpretability methods, such as LIME, enhance explainability by identifying relevant features and their importance for a subset of data, aiding the understanding of individual instances (Mazhar and Dwivedi 2024). This category is widely recognized in the literature and is frequently used as a primary classification for XAI methods (Adadi and Berrada 2018; Islam et al. 2022; Linardatos et al. 2020; Hu et al. 2021; Molnar et al. 2023; Alshamsi 2014; Morik et al. 2002; Lariviere and Vandenpoel 2005; Sheehan et al. 2017; Tillmanns et al. 2017; Wang 2020; Xiao and Benbasat 2007; Bian et al. 2018; Bonisone et al. 2002; Boodhun and Jayabalan 2018; Christmann 2004; David 2015; Gan 2013; Gan and Huang 2017; Gan and Valdez 2017; Gweon et al. 2020; Jiang et al. 2019; Kumar et al. 2010).

7.3 Stage of explanation

7.3.1 Intrinsic

Intrinsic models are inherently interpretable because of their simple and transparent structure. Their decision-making process can be understood directly from their design without additional explanation tools, such as decision trees, linear regression, and rule-based systems.

7.3.2 Post-hoc explanation

Post-hoc explanations are methods applied after training a model has been trained to provide insights into its decision-making process. These methods are not part of the model's initial design but are used to interpret and explain its predictions. Methods that provide post-hoc explanations include the inspection of learned features, feature importance, and feature interaction. Examples include LIME, SHAP, and saliency maps.

Unlike post-hoc methods, ante-hoc techniques, such as Decision Trees and CART (Moore 1987), are inherently explainable owing to their clear structure, with internal nodes split by specific conditions. Although they can become complex, the most relevant decisions are visible at the top levels. This introduces the "Stage" category, distinguishing methods used post-prediction (post-hoc) from those that are intrinsically explainable (ante-hoc), supported by evidence in (Adadi and Berrada 2018; Islam et al. 2022; Vilone and Longo 2020; Linardatos et al. 2020; Minh et al. 2022; Lin et al. 2021; Speith 2022; Arrieta et al. 2020; Sevim et al. 2016; Neumann et al. 2019; Smith et al. 2000; Baudry and Robert 2019; Bermúdez et al. 2008; Cao and Zhang 2019; Lin et al. 2021; Cheng et al. 2020; Sun et al. 2019; Viaene et al. 2004, 2002; Li et al. 2018; Matloob et al. 2020; Smyth and Jørgensen 2002).

8 Model-agnostic XAI (MA-XAI) methods in finance

MA-XAI methods, as discussed in Table 3, are techniques used to explain the predictions of any ML model, regardless of its architecture. In finance, where decision-making is heavily regulated and explanations are crucial for transparency and trust, MA-XAI methods play a key role in interpreting complex model outputs. The criteria for choosing the MA methods are discussed in Tables 2, 3.

8.1 Feature interaction and importance

Feature interaction and importance are critical concepts XAI that help us understand how features contribute individually and jointly to model predictions. Feature importance measures the contribution of each feature to the predictive performance of a model. Permutation Feature Importance, Mean Decrease in Impurity, SHAP are some feature importance methods used to explain a model. Feature interaction examines how two or more features work together to influence the model predictions. Some commonly used methods are PDPs, Individual Conditional Expectation (ICE) Plots, SHAP Interaction Values and Accumulated Local Effects (ALE) plots. The authors in (Ghosh and Dragan 2023; Bussmann et al. 2021;

Table 3 MA-XAI methods in finance

Authors	XAI technique	Model agnostic	Local	Global	Post-hoc	In-trinsic
Malhi et al. (2020); Zhang et al. (2022); Bussmann et al. (2020); Gawantka et al. (2024); Mandeep et al. (2022); Ullah et al. (2021); Dastile and Celik (2021); Tyagi (2022); Redelmeier et al. (2020); Chromik (2021); Watson (2022); Kim and Woo (2021); Bussmann et al. (2021); Maree et al. (2020); Sohail et al. (2021); Hastie et al. (2009); Friedman (2001); Ji (2021)	SHAP	Yes	Yes	Yes	Yes	No
Malhi et al. (2020); Mazhar and Dwivedi (2024); Çelik et al. (2023); Gawantka et al. (2024); Ghosh and Dragan (2023); Mandeep et al. (2022); Ullah et al. (2021); Wu and Wang (2021); Dastile and Celik (2021); De et al. (2020); Tian and Liu (2020); Alblooshi et al. (2024)	LIME	Yes	Yes	Yes	Yes	No
Zhang et al. (2022); Friedman (2001)	PDPs	Yes	No	Yes	Yes	No
Goldstein et al. (2015)	ICE Plots	Yes	Yes	No	Yes	No
Okoli (2023)	ALE Plots	Yes	No	Yes	Yes	No
Zhang et al. (2022); Hashemi and Fathi (2020); Dastile et al. (2022); Hastie et al. (2009); Zhang et al. (2022); Watson (2022); Mutlu et al. (2022); White and Garcez (2019); Guidotti et al. (2019a, b); Guidotti (2024)	Counterfactuals	Yes	Yes	Yes	Yes	No
La Gatta et al. (2021b)	PASTLE	Yes	Yes	No	Yes	No
La Gatta et al. (2021a)	CASTLE	Yes	Yes	No	Yes	No
Ribeiro et al. (2018)	Anchors	Yes	Yes	Yes	Yes	No
Tian and Liu (2020)	MANE	Yes	No	Yes	Yes	No
Gkolemis et al. (2022)	DALE	Yes	No	Yes	Yes	No
Watson (2022)	Rational Shapley Values	Yes	Yes	No	Yes	No
De et al. (2020)	TREPAN	Yes	Yes	No	Yes	No

Bove et al. 2021; Viaene et al. 2005; Tao et al. 2012; Sohail et al. 2021; Smith et al. 2000; Biddle et al. 2018; Tillmanns et al. 2017; Shah and Guez 2009; Khodairy and Abosamra 2021; Chang and Lai 2021) used feature interaction and importance methods in XAI to address the problems in their research as illustrated in Fig. 12.

Table 2 Criteria for the selection of the MA methods in finance

Criteria	Model-agnostic
1. What? (What does the method for explain?)	This criterion addresses whether the method provides explanations at a local level (for individual predictions) or global level (for the entire model)
2. Examples (Popular Methods)	LIME SHAP Counterfactuals Feature Importance
3. Mechanism (How does it work?)	The underlying approach used to generate explanations **Examples: Perturbation-based, Surrogate Models, Gradient-based and Feature Importance**
4. Applicability (Where can it be applied)?	Whether the method is applicable to any model type (MA) or is limited to specific types of models Examples: MA and MS
5. Explainability (What kind of insights does it provide?)	The nature of the explanation generated, such as feature importance, feature interactions, or counterfactuals Examples: Feature Importance, Feature Interaction and Counterfactuals
6. Type (Local vs. Global)	Whether the method provides insights into individual predictions or the overall model behaviour Examples: Local and Global
7. Ease of Use (How easy is it to implement?)	The complexity involved in using the method, including implementation difficulty and interpretability of results Examples: Easy, Moderate and Complex

8.1.1 Shapley additive explanations (SHAP)

SHAP is a unified approach for interpreting ML models. It is based on cooperative game theory, particularly the concept of Shapley values, which provides a fair distribution of payoffs among players. In the context of ML, SHAP values explain the contribution of each feature to the model's prediction. Originally from cooperative game theory, Shapley values assign a value to each player (feature) based on their contribution to the total payout (i.e., the prediction). In ML, this means quantifying the contribution of each feature to the final prediction. SHAP was introduced by Lundberg and Lee (Lundberg and Lee 2017). Authors in (Malhi et al. 2020; Zhang et al. 2022; Bussmann et al. 2020, 2021; Gawantka et al. 2024; Mandeep et al. 2022; Ullah et al. 2021; Dastile and Celik 2021; Tyagi 2022; Redelmeier et al. 2020; Chromik 2021; Watson 2022; Kim and Woo 2021; Maree et al. 2020; Hastie et al. 2009; Friedman 2001; Ji 2021). SHAP Interaction Values are an extension of the SHAP method to capture and quantify the interactions between features. They provide insights into not only individual feature contributions but also how pairs of features interact to influence the model's predictions. An overview of the SHAP interaction values and their applications in explainability is shown in Fig. 13.

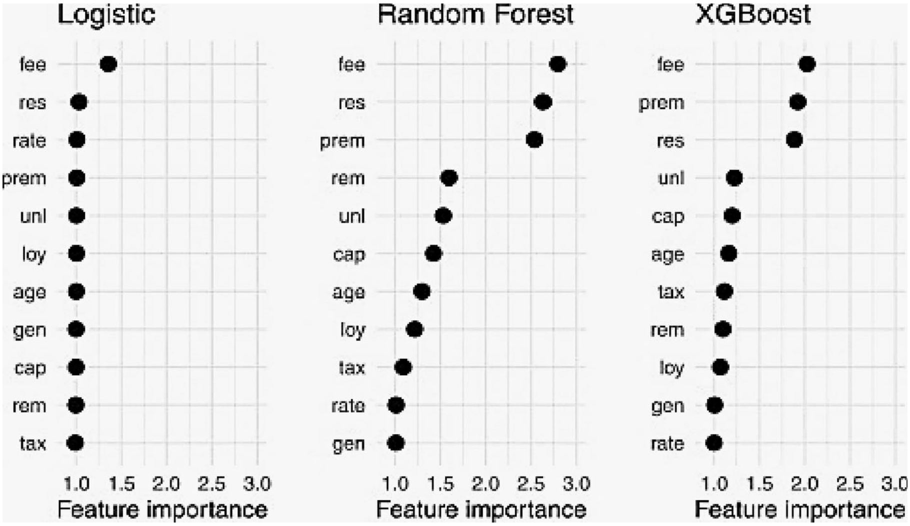


Fig. 12 Feature importance comparison for three ML models, evaluated based on cross-entropy loss. The plot highlights the relative influence of individual features on the model predictions, demonstrating how key financial variables impact the classification outcomes. Higher feature importance values indicate stronger predictive contributions, aiding model interpretability and explainability in AI-driven financial applications (Bermúdez et al. 2023)

8.1.2 Partial dependence plots (PDPs)

PDPs (Friedman 2001) are a popular method for explainability, showing how one feature influences another and helping to explain the target feature. This visual representation clarifies these relationships. PDPs can be applied to any predictive model and offer global explanations in (Zhang et al. 2022) as shown in the Fig. 14.

8.1.3 Individual conditional expectation (ICE) plots

ICE (Goldstein et al. 2015) plots are a valuable tool in XAI for visualizing the effect of a single feature on the predicted outcome of a model across individual instances. Unlike PDPs, which show the average effect of a feature, ICE plots provide a more granular view by displaying how each instance's prediction changes when a feature is varied, as shown in the Fig. 15.

8.1.4 Accumulated local effects (ALE) plots

ALE (Okoli 2023) plots are a powerful tool in XAI for interpreting complex ML models. ALE plots address some limitations of PDPs by considering the local distribution of features, thereby providing unbiased and more accurate insights, especially in the presence of feature interactions, as shown in the Fig. 16.

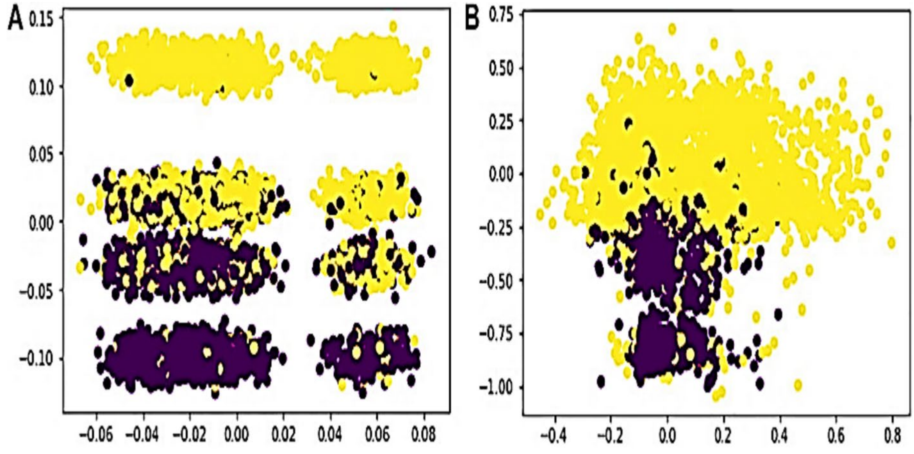


Fig. 13 Visualization of SHAP and LIME feature explanations using spectral clustering, demonstrating model interpretability differences in AI-based financial applications (Gramegna and Giudici 2021)

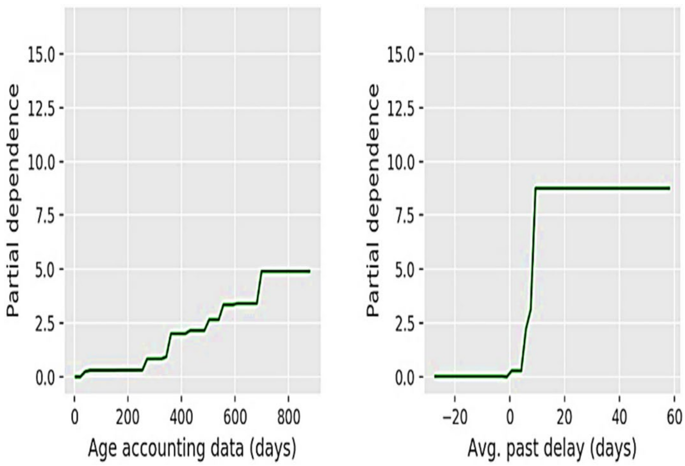


Fig. 14 Example of a PDP illustrating the relationship between input features and model predictions, providing global interpretability in AI-driven financial models (Sigrist and Hirnschall 2019)

8.1.5 Counterfactual

Counterfactual explanations (Hashemi and Fathi 2020; Dastile et al. 2022; Guidotti 2024) are a powerful method in the field of XAI that provides insights by showing how changing certain features can alter a model's prediction. These explanations are particularly useful for understanding model behavior and answering "what-if" scenarios. They offer a way to make AI systems more transparent and interpretable, especially in high-stakes applications, such as finance, healthcare, and criminal justice. The authors in (Hastie et al. 2009; Zhang et al. 2022; Watson 2022; Mutlu et al. 2022; White and Garcez 2019; Guidotti et al. 2019a, b; Pawelczyk et al. 2019) applied this method for their problem-solving.

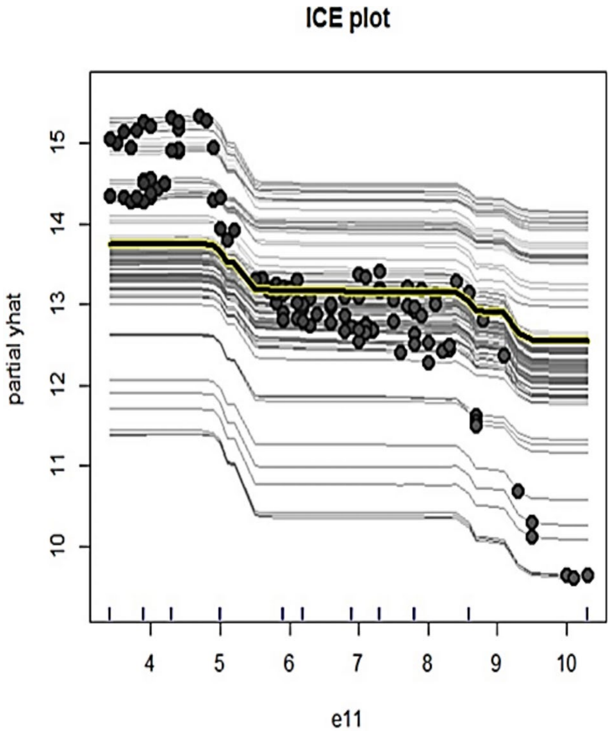


Fig. 15 ICE plot illustrating how a single feature influences model predictions at an individual instance level. Unlike PDP, ICE plots reveal heterogeneous feature effects by displaying multiple conditional response curves, making them particularly useful for detecting interactions and nonlinear relationships in AI-driven financial models (Fernández 2020)

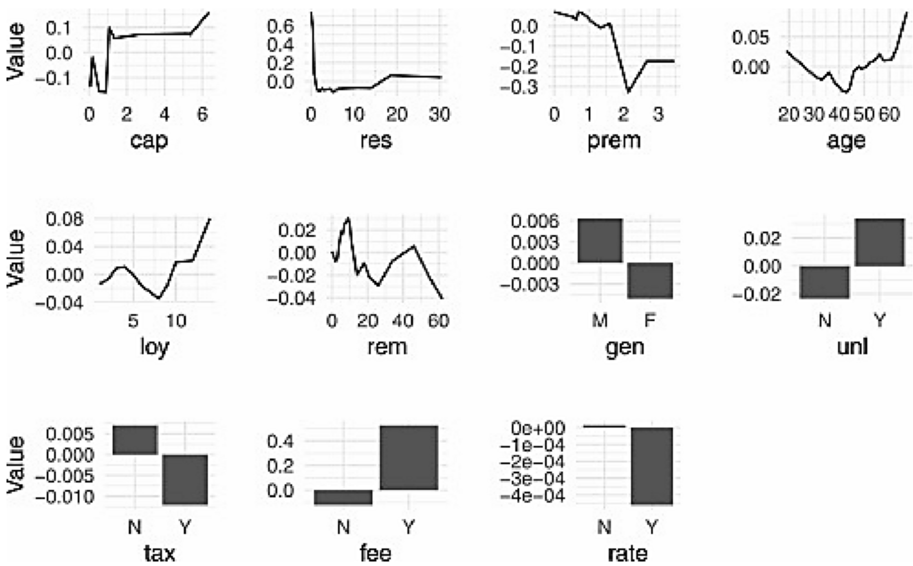


Fig. 16 Accumulated local effects (ALE) plot showing how individual features influence model predictions while considering feature interactions, improving fairness and transparency in financial AI (Bermúdez et al. 2023)

8.1.6 PASTLE

PASTLE (Partial Dependency and Accumulated Local Effects; La Gatta et al. 2021b) is a hybrid method that combines the strengths of PDPs and ALE plots to provide a comprehensive and nuanced view of the feature effects in ML models. PASTLE aims to leverage the global interpretability of PDPs and the local accuracy of ALE plots, ensuring that users can understand both the overall and local behaviors of their models.

8.1.7 CASTLE

CASTLE (Conditional Accumulated SHAP and Local Effects; La Gatta et al. 2021a) is an advanced method that combines the strengths of SHAP values and ALE to provide comprehensive model explanations. CASTLE aims to offer both global and local interpretability, addressing the limitations of individual methods and providing a more nuanced understanding of complex ML models.

8.1.8 Anchors

Anchors (Ribeiro et al. 2018) is a method developed to provide high-precision, human-interpretable explanations for ML. It aims to produce explanations that are easy to understand and closely tied to the decision-making process of the model. Anchors are specific conditions or rules that guarantee a certain prediction with high precision when met. These conditions serve as "anchors" for the prediction, ensuring that similar instances receive the same output.

8.1.9 MANE

Model-Agnostic Neural Explanations (MANE; Tian and Liu 2020) aim to provide interpretations for any ML model using NNs. The core idea is to create explanations that are MA, meaning they can be applied regardless of the underlying ML model, whether it is a DNNs, decision tree, or support vector machine.

8.1.10 DALE

Differential Accumulated Local Effects (DALE) focus on providing explanations for ML models by examining how changes in input features affect predictions. It extends the ALE concept to compare the effects of feature changes between different groups or contexts, such as comparing predictions between different classes or demographic groups.

8.1.11 Rational Shapley values

Rational Shapley Values (RSV; Watson 2022) are a refinement of the traditional Shapley values used in cooperative game theory and XAI. They aim to address certain limitations of Shapley values, particularly in scenarios where interactions between features (or players in game theory terms) are significant.

8.1.12 TREPAN

TREPAN (Decision Tree Induction based on TREPANning; De et al. 2020) is an algorithm designed to build decision trees that prioritize interpretability. It was developed to address some of the limitations of traditional decision tree algorithms, such as ID3 and C4.5, focusing specifically on producing compact and understandable trees.

8.2 Local interpretable model agnostic explanation (LIME)

LIME is a technique used to explain the predictions of ML models. It is particularly useful for understanding complex, black-box models by locally approximating them with interpretable models (Ribeiro et al. 2016b). LIME focuses on explaining individual predictions rather than the model. It creates an interpretable model that approximates the black-box model in the vicinity of the prediction being elucidated. To generate explanations, LIME perturbs the input data and observes how the predictions are changed. By sampling points around the instance being explained, LIME can build a local dataset that reflects the behavior of the black-box model in that region, as shown in Fig. 17. The authors in (Malhi et al. 2020; Mazhar and Dwivedi 2024; Çelik et al. 2023; Gawantka et al. 2024; Ghosh and Dragan 2023; Mandeep et al. 2022; Ullah et al. 2021; Wu and Wang 2021; Dastile and Celik

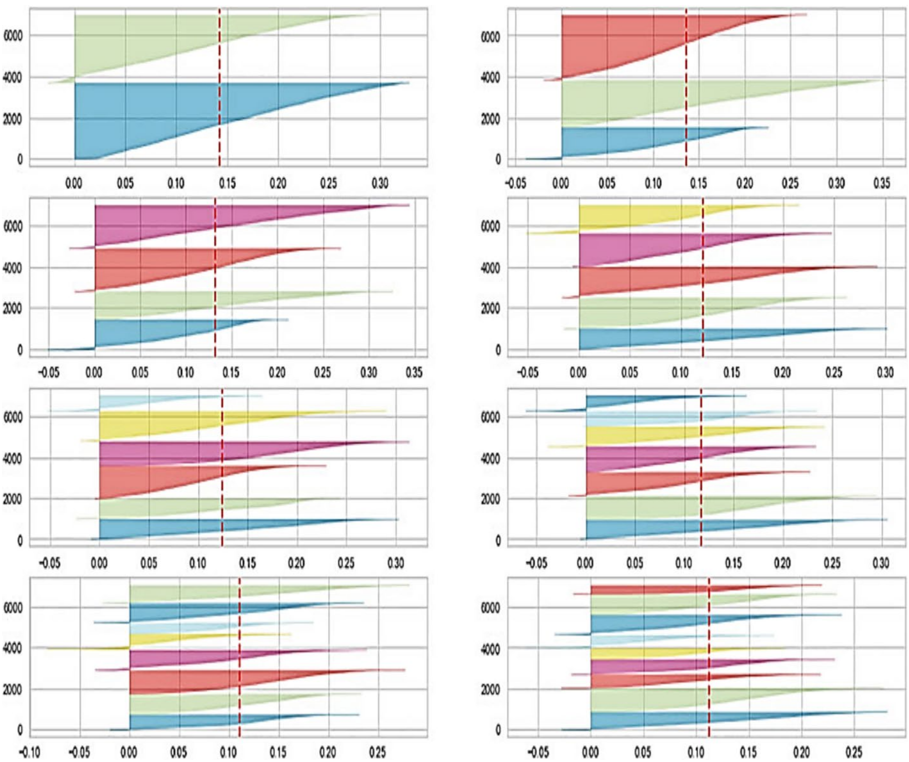


Fig. 17 Silhouette analysis of LIME-based data clustering, evaluating cluster cohesion and separation for model interpretability (Gramegna and Giudici 2021)

2021; De et al. 2020; Tian and Liu 2020; Tyagi 2022; Ji 2021; Alblooshi et al. 2024) used LIME in their studies to explain the model decision.

8.3 Attention mechanism

The attention mechanism allows a model to focus on specific parts of the input data when making predictions rather than processing the entire input at once. This is particularly useful for tasks in which different parts of the input data have varying levels of importance. Attention mechanisms have been widely used in models such as transformers, which are the backbone of many state-of-the-art NLP models, such as BERT and GPT. The authors of (Delong and Wüthrich 2020; Deprez et al. 2017; Zhang and Kong 2020) used this method for their model descriptions.

8.4 Dimensionality reduction

Dimensionality reduction plays a significant role in XAI by simplifying complex datasets and models, making them more interpretable and easier to understand than before. The authors of (Huang and Meng 2019; Cao and Zhang 2019; Wang and Xu 2018; Behera et al. 2016) used this method to explain their models.

8.5 Knowledge distillation and rule extraction

Knowledge distillation is a technique in which a "teacher" model (typically a large, complex model) transfers its knowledge to a "student" model (a smaller, simpler model). The goal is to retain most of the teacher model's performance while benefiting from the simplicity and interpretability of the student model. Rule extraction aims to derive human-readable rules from complex ML models. These rules help in understanding the decision-making process of the model, rendering it more interpretable and transparent. The authors in (Pathak et al. 2005; Kose et al. 2015; Duval and Pigeon 2019; Bermúdez et al. 2008; Kašćelan et al. 2016; Gweon et al. 2020) used this method to explain their model decision.

9 Quantitative analysis and research findings

We performed a quantitative analysis to investigate the studies reviewed. This involved collecting data on multiple aspects, such as the distribution of pioneering research among various XAI methods in finance. In addition, we provided detailed answers to the research questions presented in Table 1.

9.1 RQ1

> Which XAI framework has been widely used by researchers to study the financial domain while applying XAI techniques?

Among the various XAI frameworks, MA Explanations have been widely used by researchers to study the financial domain while applying XAI techniques. These frameworks are popular because of their ability to provide clear and interpretable explanations for complex ML models, making them suitable for financial applications where transparency and accountability are crucial. The ratio of MA methods to MS methods used in the financial domain is not universally fixed and can vary depending on the specific context and applications. However, MA methods tend to be more widely adopted because of their versatility and broad applicability across different models, as shown in the Fig. 18.

9.2 RQ2

> Which MA-XAI techniques or methods are frequently employed by researchers in the financial analysis domain?

As shown in Table 4 and Fig. 19, LIME and SHAP have been widely used in the finance domain because they can be applied to any ML model, which accounts for 52% of the total MA methods used in this study. This flexibility is crucial in finance, where various types of models (e.g., decision trees, NNs, and ensemble methods) are used for different applications. The finance industry requires high levels of interpretability owing to regulatory requirements and the need for stakeholders to understand and trust the decision-making process. LIME and SHAP provide clear, human-understandable explanations for complex models, making them suitable for regulated environments. LIME and SHAP are effective in detecting bias and ensuring fairness in model predictions. This is particularly important in finance, where biased decisions can lead to significant financial and reputational risk.

9.3 RQ3

> Which AI/ML algorithms have researchers predominantly employed in the investigation of financial datasets when applying MA-XAI methods?

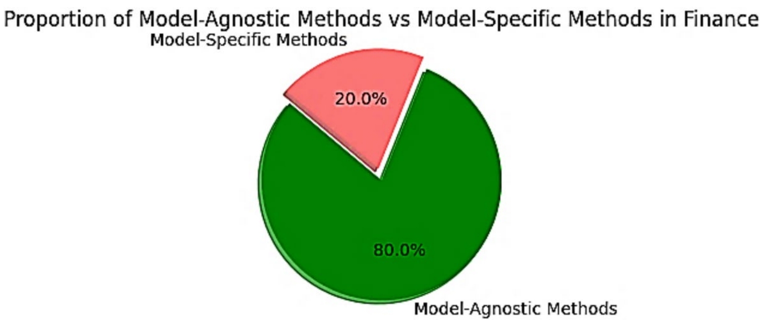


Fig. 18 Comparison of model-specific vs. model-agnostic explainability methods in financial AI, highlighting their usage distribution and applicability across different financial tasks

Table 4 Author-wise MA-XAI publications in finance

Authors	XAI technique	Count
Malhi et al. (2020); Zhang et al. (2022); Bussmann et al. (2020); Gawantka et al. (2024); Mandeep et al. (2022); Ullah et al. (2021); Dastile and Celik (2021); Tyagi (2022); Redelmeier et al. (2020); Chromik (2021); Watson (2022); Kim and Woo (2021); Bussmann et al. (2021); Maree et al. (2020); Sohail et al. (2021); Hastie et al. (2009); Friedman (2001); Ji (2021)	SHAP	18
Malhi et al. (2020); Mazhar and Dwivedi (2024); Çelik et al. (2023); Gawantka et al. (2024); Ghosh and Dragan (2023); Mandeep et al. (2022); Ullah et al. (2021); Wu and Wang (2021); Dastile and Celik (2021); De et al. (2020); Tian and Liu (2020); Tyagi (2022); Ji (2021); Alblooshi et al. (2024)	LIME	14
Zhang et al. (2022); Friedman (2001)	PDPs	2
Goldstein et al. (2015)	ICE Plots	1
Okoli (2023)	ALE Plots	1
Zhang et al. (2022); Hashemi and Fathi (2020); Dastile et al. (2022); Hastie et al. (2009); Zhang et al. (2022); Watson (2022); Mutlu et al. (2022); White and Garcez (2019); Guidotti et al. (2019a, b); Pawelczyk et al. (2019)	Counterfactuals	10
La Gatta et al. (2021b)	PASTLE	1
La Gatta et al. (2021a)	CASTLE	1
Ribeiro et al. (2018)	Anchors	1
Tian and Liu (2020)	MANE	2
Gkolemis et al. (2022)	DALE	1
Watson (2022)	Rational Shapley values	1
De et al. (2020)	TREPAN	1
Pathak et al. (2005); Kose et al. (2015); Duval and Pigeon (2019); Bermúdez et al. (2008); Kašćelan et al. (2016); Gweon et al. (2020)	Teacher-student model	6
Huang and Meng (2019); Cao and Zhang (2019); Wang and Xu (2018); Behera et al. (2016)	Dimensionality reduction	4
Delong and Wüthrich (2020); Deprez et al. (2017); Zhang and Kong (2020)	Attention mechanism	3

Researchers have predominantly employed a variety of AI/ML algorithms to investigate financial datasets when applying MA-XAI methods. These algorithms range from traditional ML models to more complex DL models. Some of the commonly used models, as shown in Table 5, are:

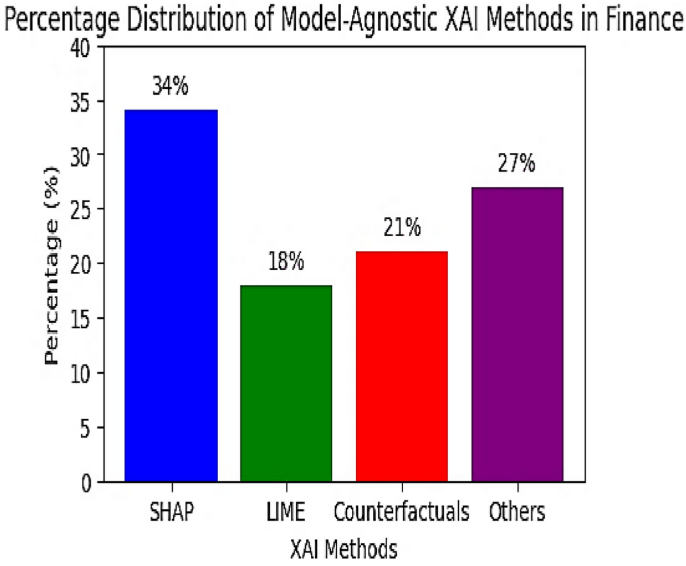


Fig. 19 Percentage distribution of model-agnostic (MA) explainability methods, highlighting their usage across financial AI applications

9.3.1 Random forest (RF)

Reason for use: RF is widely used in finance because of its robustness and ability to handle complex financial data with many features. They are highly interpretable when combined with MA methods, such as SHAP or LIME.

Applications: Credit scoring, fraud detection, loan default prediction, and risk management.

9.3.2 Gradient boosting machines (GBM)/XGBoost/LightGBM

Reason for use: These algorithms are popular because of their high predictive accuracy, especially in financial datasets, where nonlinear relationships and interactions between features are common. XAI methods, such as SHAP, are particularly useful for explaining these black-box models.

Applications: Stock price prediction, credit risk modelling, investment analysis, and customer churn prediction.

9.3.3 Logistic regression (LR)

Reason for use: While inherently interpretable, logistic regression is often paired with MA methods to analyze residuals or interactions between variables. It remains a popular baseline model in finance for tasks, such as binary classification.

Applications: Bankruptcy prediction, credit scoring, fraud detection, and customer segmentation.

Table 5 List of AI models used by the researchers

AI algorithms	Authors	Count
ANN (GAM, GLM, CANN, SOFM, DNN)	Maree et al. (2020); Viaene et al. (2005); Smith et al. (2000); Shah and Guez (2009); Chang and Lai (2021); Delong and Wüthrich (2020); Huang and Meng (2019); Cao and Zhang (2019)	8
Logistic Regression (LR), Bayesian LR	Kašćelan et al. (2016); Bermúdez et al. (2008); Biddle et al. (2018); Huang and Meng (2019); Behera et al. (2016)	5
LSTM	Khodairy and Abosamra (2021)	1
PCA	Viaene et al. (2005); Tillmanns et al. (2017); Cao and Zhang (2019)	3
Naïve Baiyes, Bayesian Approach	Zhang and Kong (2020)	1
Decision Tree Classifier	Maree et al. (2020); Smith et al. (2000)	2
General ML Model	Bove et al. (2021); Smith et al. (2000)	2
Boosting (XGB, Regression Tree, Light GBM)	Gweon et al. (2020); Bussmann et al. (2021); Alblooshi et al. (2024); Smith et al. (2000); Biddle et al. (2018); Deprez et al. (2017); Huang and Meng (2019)	7
Bagging (RF)	Gweon et al. (2020); Ji (2021); Tillmanns et al. (2017); Huang and Meng (2019)	4
SVM, SVM Regression, Dual fuzzy SVM	Kašćelan et al. (2016); Huang and Meng (2019); Tao et al. (2012); Wang and Xu (2018)	4
Regression, Poisson Regression	Delong and Wüthrich (2020); Huang and Meng (2019)	2
Genetic Algorithm (Clustering)	Smith et al. (2000)	1
Decision Support System (Clustering)	Kose et al. (2015)	1
Fuzzy Logic	Pathak et al. (2005)	1
Dimensionality Reduction	Wang and Xu (2018)	1
DL Model	Ji (2021); Wang and Xu (2018)	2

9.3.4 DL models

Reason for use: DL models, particularly feedforward neural networks (NNs) and recurrent neural networks (RNNs), are increasingly used for their ability to handle large, complex financial datasets. MA-XAI methods, such as LIME, SHAP, and counterfactual explanations, are essential for explaining the predictions of these models.

Applications: Algorithmic trading, stock market forecasting, portfolio management, and time series analysis (RNN, LSTM).

9.3.5 Support vector machines (SVM)

Reason for use: SVMs are powerful for high-dimensional financial data but are black-box in nature. MA methods help interpret decisions, particularly in classification tasks.

Applications: Fraud detection, anomaly detection, and credit risk modelling.

9.3.6 k-Nearest neighbors (k-NN)

Reason for use: k-NN is a non-parametric algorithm used in various financial applications, particularly for clustering and classification tasks. Despite its simplicity, its decisions can benefit from XAI methods to explain why certain predictions are made.

Applications: Customer segmentation, fraud detection, and portfolio optimization.

9.3.7 Decision trees (DT)

Reason for use: Decision trees are relatively interpretable, they are often used as base models for more complex ensemble methods (e.g., RF and GBMs). MA-XAI methods, such as SHAP, can further clarify feature importance and interactions.

Applications: Credit scoring, risk analysis, and asset valuation.

9.3.8 k-Means clustering

Reason for use: Clustering techniques such as k-means, while simple, are used for segmentation and exploratory data analysis in finance. MA methods, such as SHAP, can be used to explain the clustering results.

Applications: Customer segmentation, market segmentation, and investment strategy groupings.

9.3.9 Autoencoders

Reason for use: Autoencoders are used for dimensionality reduction and anomaly detection in financial data sets. MA methods, such as SHAP or feature attribution, can help interpret compressed representations and explain anomalies.

*Applications***:** Fraud detection and anomaly detection in trading data.

9.3.10 Time series models (ARIMA, LSTM, GRU)

Reason for use: These models are popular for predicting financial time-series data, such as stock prices, exchange rates, and market trends. Although these models are complex, XAI methods such as SHAP and LIME can be used to explain their outputs.

Applications: Stock price forecasting, interest rate prediction, and financial market trend analysis.

9.4 RQ4

> Which datasets are primarily utilized in research focusing on MA-XAI methods for the analysis of financial datasets?

In research focusing on MA-XAI methods for the analysis of financial datasets, several publicly available datasets have been widely used. These datasets were chosen for their relevance to financial modelling tasks, such as credit scoring, fraud detection, stock market prediction, and risk assessment. Below are some of the most utilized datasets, as shown in Table 6:

9.4.1 Common applications of XAI in finance using these datasets

- Credit scoring and loan approval: Datasets such as the Home Credit Default Risk, German Credit, and UCI Credit Card are extensively used to develop credit-scoring models, with XAI methods applied to explain loan approval decisions and highlight important features.
- Fraud detection: Kaggle Fraud Detection Dataset is widely used to train models to detect fraudulent transactions, with SHAP, LIME, and other MA-XAI methods helping to interpret why certain transactions are flagged as fraud.
- Stock price prediction: Kaggle Stock Market Datasets and S&P 500 data are used to predict stock market movements. XAI techniques, such as SHAP and PDPs, are applied to interpret the relationship between technical indicators, news, and market prices.
- Customer behavior and marketing: The Bank Marketing Dataset is used for customer conversion and retention models, where XAI methods help explain which marketing efforts lead to successful customer engagement.
- Risk management and anomaly detection: Datasets such as FICO, LendingClub, and Fraud Detection are used in risk assessment and anomaly detection models, with XAI providing insights into the factors driving predictions.

These datasets provide a robust foundation for applying MA-XAI methods in financial research, offering real-world financial scenarios that can be analyzed using various ML models and explained using advanced interpretability techniques.

9.5 RQ5

> In XAI research in the finance field, specifically concerning MA-XAI methods, what are the distinctive performance metrics used to justify the results?

In XAI research within the finance field, especially when focusing on MA-XAI methods, performance metrics are typically divided into two categories: model performance metrics (to evaluate the accuracy and effectiveness of ML models) and explainability metrics (to assess the quality of the explanations). Both are crucial for justifying the results, as accuracy

Table 6 Financial dataset description

Dataset	Description	Common use	Link
1. UCI Credit Card Dataset (Default of Credit Card Clients Dataset)	This dataset contains information about credit card clients in Taiwan, including demographic factors, credit data, payment history, and whether they defaulted on their payments	It is widely used in credit risk modelling and classification tasks. MA-XAI methods such as SHAP and LIME are often applied to explain predictions of default risk	UCI Machine Learning Repository—Credit Card Dataset (UCI Machine Learning Repository)
2. FICO Explainable Machine Learning Challenge Dataset	This dataset consists of anonymized data for credit risk scoring, used for the FICO XAI challenge. The data includes various financial features about individuals and whether they defaulted on loans	This dataset is specifically geared towards explainable AI applications in credit scoring, making it a go-to choice for testing XAI methods like SHAP and counterfactual explanations	FICO XAI Challenge Dataset (Explainable Machine Learning Challenge (fico.com)
3. Home Credit Default Risk Dataset (Kaggle)	This dataset contains a large set of features about customers applying for loans at a home credit institution. It includes financial, demographic, and transactional data	Researchers use this dataset to develop credit scoring models and apply XAI methods to interpret model predictions regarding loan default risk	Kaggle—Home Credit Default Risk (Home Credit Default Risk| Kaggle)
4. Lending Club Loan Data	Lending Club, a peer-to-peer lending platform, has released its loan data, which includes information on loan applicants, loan terms, repayment status, and defaults	This dataset is often used to study credit risk, default prediction, and loan approval decisions, with XAI methods applied to explain which factors lead to a loan being approved or rejected	LendingClub Loan Data (All Lending Club loan data (kaggle.com)
5. Kaggle Fraud Detection Dataset	This dataset contains a highly unbalanced dataset of financial transactions labelled as fraudulent or non-fraudulent. It includes features related to transactions such as amount, timestamp, and anonymized variables	XAI methods like SHAP and LIME are used to interpret ML models applied to fraud detection, making it a commonly used dataset in financial fraud analysis	Kaggle—Credit Card Fraud Detection (Credit Card Fraud Detection (kaggle.com)
6. German Credit Dataset	This dataset consists of 1,000 loan applicants with 20 features, indicating whether the applicant poses a good or bad credit risk	It is frequently used in credit scoring studies and applied in MA-XAI research to explain predictions of creditworthiness	UCI German Credit Dataset (UCI Machine Learning Repository)
7. Kaggle Stock Market Datasets	Kaggle hosts various datasets related to stock market prices, such as daily historical prices, technical indicators, and financial news	These datasets are used in stock price prediction models where researchers apply XAI methods like SHAP or PDPs to interpret feature importance and market trends	Kaggle Stock Market Datasets (NIFTY-50 Stock Market Data (2000—2021; kaggle.com)
8. S&P 500 Stock Data	This dataset contains daily historical prices of the S&P 500 stock index, including opening, closing, and adjusted prices over several years	It is employed in predictive modelling for stock prices and is often paired with XAI techniques to explain stock price movements based on market indicators	Kaggle—S&P 500 Stock Data (S&P 500 stock data (kaggle.com)
9. Financial News Datasets	These datasets include large volumes of financial news articles, used to study the impact of sentiment on stock prices, market trends, and investment decisions	Text-based financial models like sentiment analysis are combined with XAI methods (like LIME or SHAP) to explain how certain news events influence stock prices or investment decisions	Kaggle Financial News Data (Sentiment Analysis for Financial News (kaggle.com)

Table 6 (continued)

Dataset	Description	Common use	Link
10. Yahoo Finance Historical Market Data	Yahoo Finance provides historical data for various stock indices, companies, and commodities. This dataset can include stock prices, volume, and other relevant financial information	Frequently used in forecasting models for market analysis, and XAI methods are applied to explain market movements and stock price fluctuations	Yahoo Finance Data (Yahoo Finance—Stock Market Live, Quotes, Business & Finance News)
11. Bank Marketing Dataset (UCI)	This dataset contains marketing data for a Portuguese bank, including details of customer interactions, offers, and whether the customer subscribed to a term deposit	Used for customer behaviour modelling, with XAI methods explaining predictions related to customer conversion and marketing effectiveness	UCI Bank Marketing Dataset (Bank Marketing—UCI Machine Learning Repository)

alone is insufficient in finance, where interpretability, transparency, and trust are key. The following distinctive metrics were used:

9.5.1 Model performance metrics

These metrics evaluate the predictive power of the underlying ML models used in the financial datasets. They are necessary to ensure that the model is robust and reliable before focusing on explanations. These metrics evaluate the predictive power of the underlying ML models used in the financial datasets. They are necessary to ensure that the model is robust and reliable before focusing on the explanations, as shown in Fig. 12 and Table 7.

Accuracy/precision/recall/F1-score *Use case*: These metrics are standard for classification tasks, such as predicting credit defaults, fraud detection, or customer churn. They measured the model's ability to correctly predict the target classes (Liu 2024; Onasoga and Hwidi 2024).

Relevance in finance: High accuracy ensures that the model is reliable in predicting outcomes such as loan approvals or fraud detection; however, explainability is required to justify such decisions.

Area under the curve-receiver operating characteristic (AUC-ROC) *Use case*: Often used in binary classification problems, such as credit risk modelling or fraud detection, to measure the trade-off between the true positive rate and the false positive rate.

Relevance in finance: This metric is critical in financial risk management, where it is important to balance missed frauds with false alarms.

Log loss/cross-entropy loss *Use case*: This is a measure of the classification performance based on probabilistic outputs. This is particularly useful in probabilistic credit risk models.

Relevance in finance: As many financial models output probabilities (e.g., the probability of loan default), a lower log loss indicates better probabilistic predictions.

Table 7 Performance metrics used by the researchers

Authors	Evaluation metrics	Count
Behera et al. (2016); Tillmanns et al. (2017); Zhang and Kong (2020); Biddle et al. (2018); Kašćelan et al. (2016); Smith et al. (2000); Bermúdez et al. (2008); Cao and Zhang (2019); Kose et al. (2015); Wang and Xu (2018); Maree et al. (2020); Zhang et al. (2022); Dastile and Celik (2021)	Accuracy	13
Chang and Lai (2021); Kašćelan et al. (2016)	Precision & Recall	2
Khodairy and Abosamra (2021)	F1-Score	1
Shah and Guez (2009); Gweon et al. (2020); Mandeep et al. (2022)	Mean Squared Error (MSE)	3
Huang and Meng (2019); Duval and Pigeon (2019); Mandeep et al. (2022)	Root Mean Squared Error (RMSE)	3
Bove et al. (2021)	Standard deviation (SD)	1
Deprez et al. (2017)	Poisson Distribution	1
Huang and Meng (2019)	P-Value	1
Gweon et al. (2020)	Percentage Error	1
Gweon et al. (2020); Duval and Pigeon (2019)	Mean Absolute Error (MAE)	2
Pathak et al. (2005)	Root Sum Square (RSS)	1
Tao et al. (2012); Bussmann et al. (2021); Ullah et al. (2021)	Confusion Matrix	3
Bussmann et al. (2021); Park et al. (2021)	Receiver Operating Characteristics (ROC)	2

Mean absolute error (MAE)/mean squared error (MSE) *Use case*: Commonly used in regression tasks, such as stock price prediction, interest rate prediction, portfolio returns, financial technology, and financial capability (Nourallah et al. 2024).

Relevance in finance: These metrics quantify the error in predicted financial values (e.g., stock prices), with lower errors being desirable in financial forecasting models.

9.6 Explainability metrics

MA-XAI methods aim to interpret and explain the model predictions. The effectiveness of these explanations was measured using the following metrics:

9.6.1 Fidelity (or approximation accuracy)

Use case: Fidelity measures how well a simpler interpretable model (used by methods such as LIME) approximates the behavior of the original complex model.

Relevance in finance: Ensuring that the surrogate model closely approximates the original model is critical for explaining decisions such as loan approvals or stock predictions, especially in regulated environments.

9.6.1.1 Consistency (stability) of explanations *Use case*: Measures the stability or consistency of the explanations when small changes are made to the input data.

Relevance in finance: Stability is crucial in financial applications, such as credit scoring and risk modelling. Inconsistent explanations could erode trust, especially when similar customers receive different rationales for decisions such as loan approvals or interest rates.

9.6.1.2 Sparsity *Use case*: Measures the conciseness of the explanation, typically by counting the number of features used in the explanation.

Relevance in finance: Financial practitioners prefer sparse explanations because simpler explanations are easier to interpret and justify to stakeholders (e.g., regulators and customers).

To enhance the discussion on performance metrics in XAI, it is important to analyze not only the key evaluation criteria but also the reasons why certain XAI methods outperform others in financial applications. The performance of XAI methods is typically assessed using metrics such as fidelity, consistency, stability, comprehensibility, robustness, computational efficiency and human interpretability.

One of the primary factors influencing the superiority of certain XAI methods over others is their fidelity to the original model, that is, how well the explanation method represents the true decision boundary of the AI model. SHAP provides highly faithful, globally, and locally consistent feature attributions, making it a preferred choice for financial decision-making, where transparency and accountability are critical. In contrast, LIME, while computationally efficient, may suffer from stability issues, as different perturbations can yield slightly different explanations for the same instance, making it less reliable in high-stakes financial applications such as risk management.

Furthermore, computational efficiency plays a significant role in selecting XAI techniques. Although SHAP provides high-fidelity explanations, it is computationally expensive, particularly for DL models with large datasets. Methods such as Integrated Gradients and Feature Importance-based methods offer a more efficient alternative, but they may lack the depth of explanation provided by SHAP. Future research should focus on developing scalable, real-time XAI solutions that optimize both accuracy and computational feasibility, particularly in the context of high-frequency financial transactions.

Additionally, the domain-specific relevance of an XAI method significantly influences its performance in a specific domain. For example, Counterfactual Explanations are more suitable for credit scoring and regulatory compliance, where decision-makers need to understand what minimal changes would result in a different outcome. In contrast, PDPs and ALE provide more meaningful insights into stock market forecasting by visualizing feature interactions and global model behaviour.

To advance XAI in financial applications, future research should explore hybrid XAI frameworks that combine multiple interpretability techniques to enhance both explanation reliability and computational efficiency. Additionally, more benchmarking studies are needed to systematically compare XAI methods across different financial datasets and tasks to provide standardized performance evaluations. By addressing these aspects, XAI can become more robust, scalable, and aligned with the needs of the financial industry.

10 Limitations and challenges in implementing MA-XAI methods in finance

The implementation of Model-Agnostic Explainable AI (MA-XAI) methods in finance faces several challenges and limitations. These hurdles are critical and require attention to ensure effective and transparent deployment in real-world financial systems.

10.1 High-dimensional data and temporal dynamics

Financial datasets frequently encompass high-dimensional data involving numerous market variables, economic indicators, and complex temporal structures. MA-XAI techniques, such as SHAP and LIME, struggle to manage high-dimensional and sequential data, as their explanations become less insightful or overly generalized. Additionally, financial models significantly rely on temporal dynamics, where past market behavior heavily influences future outcomes. Traditional MA-XAI methods like LIME and SHAP may inadequately capture or reflect these dynamic temporal dependencies, resulting in partial or misleading interpretations.

11 Abstract and derived features

Financial models frequently utilize abstract or derived features such as principal component analysis (PCA) components, financial ratios, and latent variables, which are inherently challenging to interpret. Although MA-XAI methods highlight the significance of these features, they typically do not elucidate their practical implications in ways comprehensible to financial experts. This limitation reduces the effectiveness of MA-XAI methods, as stakeholders require understandable explanations to make informed decisions.

11.1 Domain knowledge and lack of global interpretability

Interpreting financial AI model outputs often necessitates significant domain expertise. MA-XAI methods predominantly focus on feature importance but rarely provide insights into the underlying complex relationships without external expert interpretation. Additionally, most MA methods, such as LIME and SHAP, emphasize local interpretability (individual predictions) rather than global model behavior. Stakeholders, however, may require a holistic view of model decision patterns (global interpretability) to comprehend broader financial risk trends or model behaviors, which existing MA methods insufficiently address.

11.2 Local inconsistency and scalability

Given the inherent volatility and noise present in financial data, local explanations generated by methods like LIME can vary significantly across similar instances. Such inconsistency reduces stakeholder confidence and complicates the validation of model predictions. Furthermore, financial datasets often involve high-frequency data with extensive features, rendering some MA-XAI techniques—particularly SHAP—computationally intensive and less scalable, thus unsuitable for real-time financial decision-making scenarios.

11.3 Computational efficiency and real-time constraints

Many MA-XAI techniques, notably SHAP and Counterfactual Explanations, are computationally demanding, especially with large datasets typical in financial environments. This limitation impedes their practical integration into real-time decision-making processes such as algorithmic trading and immediate fraud detection.

11.4 Fairness and bias mitigation

Another critical limitation involves the ability of MA-XAI methods to effectively detect and explain biases embedded within financial models, especially when minority groups are underrepresented. XAI methods must be further enhanced to ensure fairness and prevent discriminatory practices in automated financial decision-making, aligning with ethical and regulatory standards.

11.5 Simplification of complex relationships

Financial models often embody intricate nonlinear relationships among variables. MA methods such as LIME and PDPs typically approximate these complex interactions linearly, potentially resulting in overly simplified and less accurate interpretations. Misrepresentation of nonlinear financial relationships could lead to misguided decision-making.

11.6 Static vs. dynamic relationships

MA-XAI approaches usually address static explanations and frequently neglect dynamic feature interdependencies common in finance. For instance, the interplay between asset prices and market volatility or investor sentiment shifts dynamically and cannot be fully captured through static XAI explanations. Thus, explanations provided may not sufficiently address the evolving nature of financial markets.

11.7 Broader context and strategic decision-making

Financial decisions often require understanding the broader contextual influences such as geopolitical events, regulatory changes, and macroeconomic shifts. Existing MA-XAI methods generally fail to incorporate these broader contexts in explanations, limiting their utility for strategic decision-making.

11.8 Potential solutions and recommendations for overcoming challenges

To enhance the practicality and robustness of MA-XAI methods in finance, this study proposes several targeted solutions and areas for future research:

11.9 Optimization of computational efficiency

Given the computational intensity of methods such as SHAP and Counterfactual Explanations, it is advisable to explore optimization strategies including:

11.9.1 Model distillation

Simplifying complex models into interpretable surrogate models.

Quantization and approximation methods, reducing computational overhead without significantly sacrificing accuracy.

11.9.2 Hybrid XAI approaches

The development of hybrid models integrating high-performing AI methods (e.g., deep neural networks) with MA-XAI techniques offers a balance between predictive accuracy and interpretability. These hybrid approaches can provide more consistent and understandable explanations suitable for regulatory audits.

11.10 Domain-specific adaptations

Tailoring MA-XAI methods to specific financial domains (e.g., risk management, fraud detection) can enhance their effectiveness. Leveraging domain expertise through interactive interfaces and incorporating expert-driven feature explanations can significantly improve the quality and acceptance of model outputs.

11.11 Real-time computational optimization

Future research should focus on optimizing computationally intensive methods (such as SHAP and LIME) to achieve real-time or near-real-time interpretability. Techniques like model distillation and quantization may enable real-time XAI integration, particularly beneficial for high-frequency trading and live credit scoring.

11.12 Ensuring regulatory compliance

The integration of MA-XAI with regulatory frameworks (Basel III, GDPR, FCRA) should be prioritized. Future research could develop standardized auditing tools based on XAI, facilitating transparent, auditable, and compliant financial AI practices. Such frameworks could ensure transparent, accountable, and ethically responsible use of AI.

11.13 Fairness and ethical AI

Finally, ethical considerations such as bias detection and fairness should become integral components of financial AI. Implementing adversarial debiasing, fairness-aware modelling, and continuous explainability audits can significantly enhance the trustworthiness and accountability of AI-driven financial decision-making systems.

12 Significance of the survey and contributions

This study offers a comprehensive analysis of Model-Agnostic XAI (MA-XAI) methods applied in financial decision-making, addressing the limitations and challenges associated with explainability in AI-driven financial models. The key contributions of this study are as follows:

12.1 Extensive literature review and systematic categorization

We reviewed 60 high-quality articles and provided an in-depth analysis of MA-XAI applications in finance.

Structured XAI methodologies into a systematic tabular format, offering a comparative overview of the different interpretability techniques used in financial applications.

12.2 Simplified explanation of MA-XAI methods for financial applications

Each XAI approach is explained intuitively, avoiding complex mathematical equations, making it accessible to both financial experts and AI practitioners.

12.3 Analysis of the most frequently used MA-XAI methods

LIME, SHAP, and Counterfactual Explanations were identified as the most widely adopted techniques for understanding financial datasets.

They evaluated the effectiveness of these methods in credit scoring, fraud detection, risk assessment, and stock market prediction.

12.4 Examination of financial datasets and AI model trends

The most used datasets in financial applications were analyzed, highlighting their role in risk assessment and investment strategies.

They found that credit management is the dominant area of research, with most selected studies focusing on AI-based credit risk assessment.

It was identified that Artificial Neural Networks (ANNs) and Boosting ML algorithms (XGBoost, LightGBM, and CatBoost) dominate financial AI research, accounting for 50% of the total applications.

12.5 Identification of challenges in the adoption of XAI in finance

They highlighted the trade-off between explainability and model accuracy, particularly in DL models.

Scalability and computational efficiency issues in post-hoc explanation methods, such as SHAP and LIME, have been addressed.

Regulatory compliance concerns were discussed, emphasizing the need for audit-friendly AI explanations to meet the requirements of the GDPR, Basel III, and Fair Credit Reporting Act (FCRA).

12.6 Practical implications for financial institutions

Explains how XAI enhances trust, regulatory alignment, and financial transparency in decision-making.

Showed that XAI improves fraud detection, risk management, and customer confidence in AI-driven financial services.

The role of human-centered XAI in improving interpretability and fairness is emphasized.

12.7 Analysis of the most frequently used MA-XAI methods

LIME, SHAP, and Counterfactual Explanations were identified as the most widely adopted techniques for understanding financial datasets.

They evaluated the effectiveness of these methods in credit scoring, fraud detection, risk assessment, and stock market prediction.

13 Discussion and future directions

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into financial services has enhanced predictive capabilities and operational efficiency, facilitated by advancements in Big Data analytics and the increased availability of large-scale financial datasets. AI models have improved market forecasting, reduced information asymmetry, and supported better risk management practices, such as credit risk assessment, bankruptcy prediction, and fraud detection. Furthermore, AI-powered early-warning systems contribute significantly to regulatory compliance and financial oversight by anticipating market disruptions and enabling timely interventions. Despite these benefits, AI models often operate as "black boxes," lacking transparency and limiting stakeholder trust. Explainable AI (XAI) has emerged as an essential solution to these challenges, offering interpretability and regulatory compliance by providing human-understandable explanations of AI-driven decisions. However, this study acknowledges several limitations. First, a trade-off exists between explainability and predictive accuracy, as interpretable models (e.g., decision trees or linear regression) generally achieve lower accuracy compared to complex models, such as deep neural networks (DNNs). Addressing this, future research should focus on hybrid approaches combining rule-based models with DNNs to effectively balance interpretability and accuracy. Second, computational complexity and scalability remain critical concerns, especially for computationally intensive post-hoc explainability methods like SHAP and

LIME. Future studies should investigate hardware acceleration and optimization techniques (e.g., GPU and TPU utilization, approximate algorithms) to enhance the efficiency and scalability of these methods on large financial datasets. Third, the generalizability of XAI methods across diverse financial contexts is still uncertain. Future research should develop adaptive frameworks that tailor explanations specifically to financial applications such as stock prediction, credit scoring, and fraud detection, thereby improving consistency and practical relevance. Fourth, regulatory compliance and trustworthiness are crucial in the highly regulated financial sector, requiring alignment with frameworks such as GDPR, Basel III, and the Fair Credit Reporting Act (FCRA). Future efforts should standardize XAI-driven auditing tools, such as SHAP-based audits or Counterfactual-based compliance checks, to strengthen regulatory adherence and accountability. Additionally, integrating XAI methods into areas like risk management and anti-money laundering (AML) can enhance the transparency and fairness of high-risk financial decisions. Techniques such as Partial Dependence Plots (PDPs) and LIME can help detect and mitigate biases, thus improving trust and accountability in automated financial processes. Further research is also required to investigate underexplored applications of Model-Agnostic XAI methods, such as portfolio optimization, internet financing platforms, and advanced fraud detection mechanisms. Moreover, combining global and local interpretability methods (e.g., SHAP and LIME) could address challenges associated with high-dimensional data and complex decision structures. Future studies should also examine how XAI impacts organizational performance, specifically investigating the effects of enhanced explainability on brand equity, customer trust, and investor confidence. Collaborative research involving AI developers, financial professionals, and regulators will be crucial in advancing ethical AI practices, ensuring compliance with financial regulations, and promoting reliable, fair, and transparent AI-driven financial decision-making.

14 Conclusion

This systematic review critically evaluated the adoption and application of Model-Agnostic Explainable Artificial Intelligence (MA-XAI) methods in financial domains. The analysis identified prominent MA-XAI techniques, including SHAP, LIME, Counterfactual Explanations, and Partial Dependence Plots (PDPs), highlighting their widespread use across diverse financial scenarios such as credit scoring, fraud detection, risk assessment, and portfolio management. Additionally, the review introduced a unified taxonomy to standardize classification and facilitate broader adoption of these methods. Despite their evident benefits, significant challenges persist, notably the balance between interpretability and predictive accuracy, computational demands, scalability constraints, and meeting evolving regulatory standards. To address these challenges, future research should specifically explore hybrid XAI models that effectively combine interpretability with predictive performance, computational optimizations for real-time interpretability, regulatory-aligned XAI frameworks, and ethical strategies for bias mitigation. Advancements in these areas will significantly enhance transparency, accountability, and trustworthiness in AI-driven financial decision-making. Regulatory Alignment and Compliance: Develop standardized XAI auditing frameworks aligned explicitly with regulatory mandates (e.g., Basel III, GDPR, Fair

Credit Reporting Act) to facilitate transparency, accountability, and compliance in financial AI systems.

Acknowledgements All authors contributed equally to the preparation of the manuscript. The authors thank Abdullah Al Salem University (AASU) for their support in the publication of this article.

Author contributions Conceptualization: Farhina Sardar Khan, Formal Analysis: Syed Shahid Mazhar, Dhoha Al Saleh, methodology: Kashif Mazhar, Supervision: Syed Shahid Mazhar Validation: Syed Shahid Mazhar Investigation: Amir Mazhar, Funding acquisition: Dhoha Al Saleh, Writing—review & editing: Syed Shahid Mazhar, Farhina Sardar Khan, Dhoha Al Saleh, Kashif Mazhar, Amir Mazhar.

Funding This work was supported by Abdullah Al Salem University (AASU), Kuwait.

Data availability There is no dataset available to accompany this review paper

Declarations

Competing interest The authors declare that they have no competing financial interests or personal relationships that could have influenced the work reported in this study.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

Abedin MZ, Guotai C, Moula FE, Azad AS, Khan MS (2019) Topological applications of multilayer perceptrons and support vector machines in financial decision support systems. Int J Finance Econ 24(1):474–507. https://doi.org/10.1002/ijfe.1675

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahmed I, Jeon G, Piccialli F (2022) From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. IEEE Trans Ind Inf 18(8):5031–5042. https://doi.org/10.1109/TII.2022.3146552

Alblooshi M, Alhajeri H, Almatrooshi M, Alaraj M (2024) Unlocking transparency in credit scoring: leveraging XGBoost with XAI for informed business decision-making. In: 2024 International conference on artificial intelligence, computer, data sciences and applications (ACDSA), IEEE. 1–6. https://doi.org/10.1109/ACDSA59508.2024.10467573.

Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, Guidotti R, Del Ser J, Díaz-Rodríguez N, Herrera F (2023a) Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fus 99:101805. https://doi.org/10.1016/j.inffus.2023.101805

Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M (2023b) The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review. Comput Biol Med 166:107555. https://doi.org/10.1016/j.compbiomed.2023.107555

AlSaleh DA (2019) The role of technology-based services in establishing brand equity within the private hospitals sector in Kuwait. J Transnatl Manag 24(1):21–39. https://doi.org/10.1080/15475778.2018.1562298

Alshamsi AS (2014) Predicting car insurance policies using random forest. In: 2014 10th International conference on innovations in information technology (IIT). IEEE, 128–132. https://doi.org/10.1109/INNOVATIONS.2014.6987575.

Amelot LM, Agathee US, Sunecher Y (2021) Time series modelling, NARX neural network and hybrid KPCA–SVR approach to forecast the foreign exchange market in mauritius. Afr J Econ Manag Stud 12(1):18–54. https://doi.org/10.1108/AJEMS-04-2019-0161

Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM (2021) Explainable artificial intelligence: an analytical review. Wires Data Min Know Discovery. https://doi.org/10.1002/widm.1424

Ardekani AM, Bertz J, Bryce C, Dowling M, Long SC (2024) FinSentGPT: a universal financial sentiment engine? Int Rev Financ Anal 94:103291. https://doi.org/10.1016/j.irfa.2024.103291

Arrieta B, Alejandro N-R, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fus 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Babaei G, Giudici P, Raffinetti E (2022) Explainable artificial intelligence for crypto asset allocation. Financ Res Lett 47:102941. https://doi.org/10.1016/j.frl.2022.102941

Babaei G, Giudici P, Raffinetti E (2023) Explainable FinTech lending. J Econ Bus 125–126:106126. https://doi.org/10.1016/j.jeconbus.2023.106126

Babaei G, Giudici P, Raffinetti E (2025) A rank graduation box for SAFE AI. Expert Syst Appl 259:125239. https://doi.org/10.1016/j.eswa.2024.125239

Baecke P, Bocca L (2017) The value of vehicle telematics data in insurance risk selection processes. Decis Support Syst 98:69–79. https://doi.org/10.1016/j.dss.2017.04.009

Bahoo S, Cucculelli M, Goga X, Mondolo J (2024) Artificial intelligence in finance: a comprehensive review through bibliometric and content analysis. SN Busin Econ 4(2):23. https://doi.org/10.1007/s43546-023-00618-x

Barenkamp M, Rebstadt J, Thomas O (2020) Applications of AI in classical software engineering. AI Perspec 2(1):1. https://doi.org/10.1186/s42467-020-00005-4

Baudry M, Robert CY (2019) A machine learning approach for individual claims reserving in insurance. Appl Stoch Model Bus Ind 35(5):1127–1155. https://doi.org/10.1002/asmb.2455

Bauer K, Hinz O, van der Aalst W, Weinhardt C (2021) Expl(AI)n it to me—explainable ai and information systems research. Bus Inf Syst Eng 63(2):79–82. https://doi.org/10.1007/s12599-021-00683-2

Behera S, Desik PA, Soma P, Sundari N (2016) Segmentation-based predictive modeling approach in insurance marketing strategy. https://api.semanticscholar.org/CorpusID:168910032

Benhamou E, Ohana J-J, Saltiel D, Guez B (2021) Explainable AI (XAI) models applied to planning in financial markets. SSRN Electron J. https://doi.org/10.2139/ssrn.3862437

Bermúdez Ll, Pérez JM, Ayuso M, Gómez E, Vázquez FJ (2008) A Bayesian dichotomous model with asymmetric link for fraud in insurance. Insur: Math Econ 42(2):779–786. https://doi.org/10.1016/j.insmatheco.2007.08.002

Bermúdez L, Anaya D, Belles-Sampera J (2023) Explainable AI for paid-up risk management in life insurance products. Financ Res Lett 57:104242. https://doi.org/10.1016/j.frl.2023.104242

Bhattacharjee B, Sridhar A, Shafi M (2017) An artificial neural network-based ensemble model for credit risk assessment and deployment as a graphical user interface. Int J Data Min Modell Manag 9(2):122. https://doi.org/10.1504/IJDMMM.2017.085643

Bhowmik A, Sannigrahi M, Chowdhury D, Dwivedi AD, Mukkamala RR (2022) DBNex: deep belief network and explainable ai based financial fraud detection. In: 2022 IEEE international conference on big data (big data). IEEE, 3033–42. https://doi.org/10.1109/BigData55660.2022.10020494.

Bian Y, Chen Yang J, Zhao L, Liang L (2018) Good drivers pay less: a study of usage-based vehicle insurance models. Transp Res Part a: Policy Pract 107:20–34. https://doi.org/10.1016/j.tra.2017.10.018

Bichler M, Gupta A, Ketter W (2010) Research commentary—designing smart markets. Inf Syst Res 21(4):688–699. https://doi.org/10.1287/isre.1100.0316

Biddle R, Liu S, Tilocca P, Xu G (2018) Automated underwriting in life insurance: predictions and optimisation. 135–146. https://doi.org/10.1007/978-3-319-92013-9_11

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

Bogina V, Hartman A, Kuflik T, Shulner-Tal A (2022) Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. Int J Artif Intell Educ 32(3):808–833. https://doi.org/10.1007/s40593-021-00248-0

Bonisone PP, Subbu R, Aggour KS (2002) Evolutionary optimization of fuzzy decision systems for automated insurance underwriting. In: 2002 IEEE world congress on computational intelligence. 2002 IEEE international conference on fuzzy systems. FUZZ-IEEE'02. Proceedings (Cat. No.02CH37291). IEEE, 1003–1008. https://doi.org/10.1109/FUZZ.2002.1006641

Bonissone PP (2015) Machine learning applications. Springer handbook of computational intelligence. Springer, Berlin, pp 783–821

Boodhun N, Jayabalan M (2018) Risk prediction in life insurance industry using supervised learning algorithms. Complex Intell Syst 4(2):145–154. https://doi.org/10.1007/s40747-018-0072-1

Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, Nensa F (2023) Explainable AI in medical imaging: an overview for clinical practitioners—saliency-based XAI approaches. Eur J Radiol 162:110787. https://doi.org/10.1016/j.ejrad.2023.110787

Bove C, Aigrain J, Lesot MJ, Tijus C, Detyniecki M (2021) Contextualising local explanations for non-expert users: an XAI pricing interface for insurance. In IUI Workshops. https://api.semanticscholar.org/CorpusID:235958016

Bruckert S, Finzel B, Schmid U (2020) The next generation of medical decision support: a roadmap toward transparent expert companions. Front Artif Intell. https://doi.org/10.3389/frai.2020.507973

Buijsman S (2022) Defining explanation and explanatory depth in XAI. Mind Mach 32(3):563–584. https://doi.org/10.1007/s11023-022-09607-9

Bussmann N, Giudici P, Marinelli D, Papenbrock J (2020) Explainable AI in fintech risk management. Front Artif Intell. https://doi.org/10.3389/frai.2020.00026

Bussmann N, Giudici P, Marinelli D, Papenbrock J (2021) Explainable machine learning in credit risk management. Comput Econ 57(1):203–216. https://doi.org/10.1007/s10614-020-10042-0

Cabitza F, Campagner A, Ciucci D (2019) New frontiers in explainable AI: understanding the GI to interpret the GO. 27–47. https://doi.org/10.1007/978-3-030-29726-8_3

Calders T, Ntoutsi E, Pechenizkiy M, Rosenhahn B, Ruggieri S (2021) Introduction to the special section on bias and fairness in AI. ACM SIGKDD Explor Newsl 23(1):1–3. https://doi.org/10.1145/3468507.3468509

Calzarossa MC, Giudici P, Zieni R (2025) An assessment framework for explainable AI with applications to cybersecurity. Artif Intell Rev 58(5):150. https://doi.org/10.1007/s10462-025-11141-w

Cao X, Li S, Katsikis V, Khan AT, He H, Liu Z, Zhang L, Peng C (2024) Empowering financial futures: large language models in the modern financial landscape. EAI Endorsed Trans AI Robot. https://doi.org/10.4108/airo.6117

Cao X, Peng C, Zheng Y, Li S, Ha TT, Shutyaev V, Katsikis V, Stanimirovic P (2024) Neural networks for portfolio analysis in high-frequency trading. IEEE Trans Neural Netw Learn Syst 35(12):18052–18061. https://doi.org/10.1109/TNNLS.2023.3311169

Cao X, Yang Y, Li S, Stanimirović PS, Katsikis VN (2025) Artificial neural dynamics for portfolio allocation: an optimization perspective. IEEE Trans Syst, Man, Cybernet: Syst 55(3):1960–1971. https://doi.org/10.1109/TSMC.2024.3514919

Cao H, Zhang R (2019) Using PCA to improve the detection of medical insurance fraud in SOFM neural networks. In: Proceedings of the 2019 3rd international conference on management engineering, software engineering and service sciences. ACM, New York, 117–22. https://doi.org/10.1145/3312662.3312713

Carfora MF, Martinelli F, Mercaldo F, Nardone V, Orlando A, Santone A, Vaglini G (2019) A 'pay-how-you-drive' car insurance approach through cluster analysis. Soft Comput 23(9):2863–2875. https://doi.org/10.1007/s00500-018-3274-y

Carta S, Podda AS, Reforgiato Recupero D, Stanciu MM (2022) Explainable AI for financial forecasting. 51–69. https://doi.org/10.1007/978-3-030-95470-3_5

Carvalho DV, Pereira EM, Cardoso JS (2019) Machine learning interpretability: a survey on methods and metrics. Electronics 8(8):832. https://doi.org/10.3390/electronics8080832

Çelik TB, İcan Ö, Bulut E (2023) Extending machine learning prediction capabilities by explainable AI in financial time series prediction. Appl Soft Comput 132:109876. https://doi.org/10.1016/j.asoc.2022.109876

Černevičienė J, Kabašinskas A (2022) Review of multi-criteria decision-making methods in finance using explainable artificial intelligence. Front Artif Intell. https://doi.org/10.3389/frai.2022.827584

Černevičienė J, Kabašinskas A (2024) Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artif Intell Rev 57(8):216. https://doi.org/10.1007/s10462-024-10854-8

Chang WT, Lai KH (2021) A neural network-based approach in predicting consumers' intentions of purchasing insurance policies. Acta Inf Pragensia 10(2):138–154. https://doi.org/10.18267/j.aip.152

Chen M-Y (2011) Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. Comput Math Appl 62(12):4514–4524. https://doi.org/10.1016/j.camwa.2011.10.030

Cheng X, Jin Z, Yang H (2020) Optimal insurance strategies: a hybrid deep learning markov chain approximation approach. ASTIN Bull 50(2):449–477. https://doi.org/10.1017/asb.2020.9

Choi I, Kim WC (2024) Unlocking ETF price forecasting: exploring the interconnections with statistical dependence-based graphs and XAI techniques. Knowl-Based Syst 305:112567. https://doi.org/10.1016/j.knosys.2024.112567
Christmann A (2004) An approach to model complex high? Dimensional insurance data. Allgemeines Statist Archiv 88(4):375–396. https://doi.org/10.1007/s101820400178
Chromik M (2021) Making SHAP rap: bridging local and global insights through interaction and narratives. 641–51. https://doi.org/10.1007/978-3-030-85616-8_37
Clement T, Kemmerzell N, Abdelaal M, Amberg M (2023) XAIR: a systematic metareview of explainable AI (XAI) aligned to the software development process. Mach Learn Knowl Extract 5(1):78–108. https://doi.org/10.3390/make5010006
Cremer CZ (2021) Deep limitations? Examining expert disagreement over deep learning. Progress Artif Intell 10(4):449–464. https://doi.org/10.1007/s13748-021-00239-1
Daníelsson J, Macrae R, Uthemann A (2022) Artificial intelligence and systemic risk. J Bank Finance 140:106290. https://doi.org/10.1016/j.jbankfin.2021.106290
Das A, Rad P (2020) Opportunities and challenges in explainable artificial intelligence (XAI): a survey
Dastile X, Celik T (2021) Making deep learning-based predictions for credit scoring explainable. IEEE Access 9:50426–50440. https://doi.org/10.1109/ACCESS.2021.3068854
Dastile X, Celik T, Vandierendonck H (2022) Model-agnostic counterfactual explanations in credit scoring. IEEE Access 10:69543–69554. https://doi.org/10.1109/ACCESS.2022.3177783
David M (2015) Auto insurance premium calculation using generalized linear models. Procedia Econ Finance 20:147–156. https://doi.org/10.1016/S2212-5671(15)00059-3
Ben David D, Resheff YS, Tron T (2021) Explainable AI and adoption of financial algorithmic advisors. In: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society. ACM, New York, 390–400 https://doi.org/10.1145/3461702.3462565
De T, Giri P, Mevawala A, Nemani R, Deo A (2020) Explainable AI: a hybrid approach to generate human-interpretable explanation for deep learning prediction. Procedia Comput Sci 168:40–48. https://doi.org/10.1016/j.procs.2020.02.255
Delong Ł, Wüthrich MV (2020) Neural networks for the joint development of individual payments and claim incurred. Risks 8(2):33. https://doi.org/10.3390/risks8020033
Demajo LM, Vella V, Dingli A (2020) Explainable AI for interpretable credit scoring. https://doi.org/10.5121/csit.2020.101516
Deprez P, Shevchenko PV, Wüthrich MV (2017) Machine learning techniques for mortality modeling. Eur Actuar J 7(2):337–352. https://doi.org/10.1007/s13385-017-0152-4
Devriendt S, Antonio K, Reynkens T, Verbelen R (2021) Sparse regression with multi-type regularized feature modeling. Insur: Math Econ 96:248–261. https://doi.org/10.1016/j.insmatheco.2020.11.010
Din ZA, Venugopalan H, Lin H, Wushensky A, Liu S, King ST (2021) Doing good by fighting fraud: ethical anti-fraud systems for mobile payments
Dixon M, Klabjan D, Bang JH (2017) Classification-based financial markets prediction using deep neural networks. Algorithm Finance 6(3–4):67–77. https://doi.org/10.3233/AF-170176
Došilović FK, Brčić M, Hlupić N (2018) Explainable artificial intelligence: a survey. In: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE, 0210–15. https://doi.org/10.23919/MIPRO.2018.8400040
Dunis CL, Laws J, Karathanasopoulos A (2013) GP algorithm versus hybrid and mixed neural networks. Eur J Finance 19(3):180–205. https://doi.org/10.1080/1351847X.2012.679740
Durango-Gutiérrez JH, Durango-Cohen PL, Velez-Ospina JA (2021) Pricing strategies in thepresence of strategic consumers and competition: a real options approach. Int J Finance Econ 26(4):4933–4956. https://doi.org/10.1002/ijfe.2475
Duval F, Pigeon M (2019) Individual loss reserving using a gradient boosting-based approach. Risks 7(3):79. https://doi.org/10.3390/risks7030079
Ebid AM (2021) 35 Years of (AI) in geotechnical engineering: state of the art. Geotech Geol Eng 39(2):637–690. https://doi.org/10.1007/s10706-020-01536-7
Elliott K, Price R, Shaw P, Spiliotopoulos T, Ng M, Coopamootoo K, Moorsel A (2021) Towards an equitable digital society: artificial intelligence (AI) and corporate digital responsibility (CDR). Society 58(3):179–88. https://doi.org/10.1007/s12115-021-00594-8.
Eluwole OT, Akande S (2022) Artificial intelligence in finance: possibilities and threats. In: 2022 IEEE international conference on industry 4.0, artificial intelligence, and communications technology (IAICT). IEEE, 268–73. https://doi.org/10.1109/IAICT55358.2022.9887488
Episcopos A, Pericli A, Jianxun Hu (1998) Commercial mortgage default: a comparison of logit with radial basis function networks. J Real Estate Finance Econ 17(2):163–178. https://doi.org/10.1023/A:1007701420328

Ernst E, Merola R, Samaan D (2019) Economics of artificial intelligence: implications for the future of work. IZA J Labor Policy. https://doi.org/10.2478/izajolp-2019-0004
Feldman D, Gross S (2005) Mortgage default: classification trees analysis. J Real Estate Finance Econ 30(4):369–396. https://doi.org/10.1007/s11146-005-7013-7
Fernández JA (2020) United States banking stability: an explanation through machine learning. Banks Bank Syst 15(4):137–149. https://doi.org/10.21511/bbs.15(4).2020.12
Fontes M, Almeida JDSD, Cunha A (2024) Application of example-based explainable artificial intelligence (XAI) for analysis and interpretation of medical imaging: a systematic review. IEEE Access 12:26419–26427. https://doi.org/10.1109/ACCESS.2024.3367606
Freeborough W, van Zyl T (2022) Investigating explainability methods in recurrent neural network architectures for financial time series data. Appl Sci 12(3):1427. https://doi.org/10.3390/app12031427
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat. https://doi.org/10.1214/aos/1013203451
Galeshchuk S, Mukherjee S (2017) Deep networks for predicting direction of change in foreign exchange rates. Intell Syst Account, Finance Manag 24(4):100–110. https://doi.org/10.1002/isaf.1404
Gan G (2013) Application of data clustering and machine learning in variable annuity valuation. SSRN Electron J. https://doi.org/10.2139/ssrn.2322863
Gan G, Valdez EA (2017) Valuation of large variable annuity portfolios: Monte Carlo simulation and synthetic datasets. Dep Model 5(1):354–374. https://doi.org/10.1515/demo-2017-0021
Gan G, Huang JX (2017) A data mining framework for valuing large portfolios of variable annuities. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, 1467–75. https://doi.org/10.1145/3097983.3098013
Gandhar A, Gupta K, Pandey AK, Raj D (2024) Fraud detection using machine learning and deep learning. SN Comput Sci 5(5):453. https://doi.org/10.1007/s42979-024-02772-x
Garg P, Chakravarthy AS, Mandal M, Narang P, Chamola V, Guizani M (2021) ISDNet: ai-enabled instance segmentation of aerial scenes for smart cities. ACM Trans Internet Technol 21(3):1–18. https://doi.org/10.1145/3418205
Gawantka F, Just F, Ullrich M, Savelyeva M, Lässig J (2024) Evaluation of XAI methods in a FinTech context. 143–154. https://doi.org/10.1007/978-3-031-49552-6_13
Gepp A, Kumar K, Bhattacharya S (2010) Business failure prediction using decision trees. J Forecast 29(6):536–555. https://doi.org/10.1002/for.1153
Ghosh I, Dragan P (2023) Can financial stress be anticipated and explained? Uncovering the hidden pattern using EEMD-LSTM, EEMD-prophet, and XAI methodologies. Complex Intell Syst 9(4):4169–4193. https://doi.org/10.1007/s40747-022-00947-8
Gianfagna L, Di Cecco A (2021) Model-agnostic methods for XAI. Explainable AI with python. Springer, Cham, pp 81–113
Gil D, Hobson S, Mojsilović A, Puri R, Smith JR (2020) AI for management: an overview. The future of management in an AI world. Springer, Cham, pp 3–19
Gimpel H, Rau D, Röglinger M (2018) Understanding FinTech start-ups—a taxonomy of consumer-oriented service offerings. Electron Mark 28(3):245–264. https://doi.org/10.1007/s12525-017-0275-0
Giudici P, Piergallini A, Recchioni MC, Raffinetti E (2024) Explainable artificial intelligence methods for financial time series. Physica A 655:130176. https://doi.org/10.1016/j.physa.2024.130176
Gkolemis V, Dalamagas T, Diou C (2022) DALE: differential accumulated local effects for efficient and accurate global explanations
Gleicher M (2016) A framework for considering comprehensibility in modeling. Big Data 4(2):75–88. https://doi.org/10.1089/big.2016.0007
Goldstein A, Kapelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 24(1):44–65. https://doi.org/10.1080/10618600.2014.907095
Govindaraj V, Jaganathan HV, Prakash P (2023) Explainable transformers in financial forecasting. World J Adv Res Rev 20(2):1434–1441. https://doi.org/10.30574/wjarr.2023.20.2.1956
Gramegna A, Giudici P (2020) Why to buy insurance? An explainable artificial intelligence approach. Risks 8(4):137. https://doi.org/10.3390/risks8040137
Gramegna A, Giudici P (2021) SHAP and LIME: an evaluation of discriminative power in credit risk. Front Artif Intell. https://doi.org/10.3389/frai.2021.752558
Guelman L (2012) Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Syst Appl 39(3):3659–3667. https://doi.org/10.1016/j.eswa.2011.09.058
Guidotti R (2024) Counterfactual explanations and how to find them: literature review and benchmarking. Data Min Knowl Disc 38(5):2770–2824. https://doi.org/10.1007/s10618-022-00831-6

Guidotti R, Monreale A, Giannotti F, Pedreschi D, Ruggieri S, Turini F (2019a) Factual and counterfactual explanations for black box decision making. IEEE Intell Syst 34(6):14–23. https://doi.org/10.1109/MIS.2019.2957223

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2019b) A survey of methods for explaining black box models. ACM Comput Surv 51(5):1–42. https://doi.org/10.1145/3236009

Gunnarsson ES, Isern HR, Kaloudis A, Risstad M, Vigdel B, Westgaard S (2024) Prediction of realized volatility and implied volatility indices using ai and machine learning: a review. Int Rev Financ Anal 93:103221. https://doi.org/10.1016/j.irfa.2024.103221

Gupta A, Dengre V, Kheruwala HA, Shah M (2020) Comprehensive review of text-mining applications in finance. Financ Innov 6(1):39. https://doi.org/10.1186/s40854-020-00205-1

Gweon H, Li S, Mamon R (2020) An effective bias-corrected bagging method for the valuation of large variable annuity portfolios. ASTIN Bull 50(3):853–871. https://doi.org/10.1017/asb.2020.28

Han J, Li Y (2023) Asset allocation strategy based on announcements and machine learning—an approach in chinese market. Highl Busin, Econ Manag 5:251–263. https://doi.org/10.54097/hbem.v5i.5083

Hanif A (2021) Towards explainable artificial intelligence in banking and financial services

Hashemi M, Fathi A (2020) PermuteAttack: counterfactual explanation of machine learning credit scorecards

Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, Scardapane S, Spinelli I, Mahmud M, Hussain A (2024) Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput 16(1):45–74. https://doi.org/10.1007/s12559-023-10179-8

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York

Herm L-V, Heinrich K, Wanner J, Janiesch C (2023) Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. Int J Inf Manag 69(April):102538. https://doi.org/10.1016/j.ijinfomgt.2022.102538

Heston SL, Sinha NR (2017) News vs. sentiment: predicting stock returns from news stories. Financ Anal J 73(3):67–83. https://doi.org/10.2469/faj.v73.n3.3

Hoang D, Wiegratz K (2023) Machine learning methods in finance: recent applications and prospects. Eur Financ Manag 29(5):1657–1701. https://doi.org/10.1111/eufm.12408

Holzinger A, Saranti A, Molnar C, Biecek P, Samek W (2022) Explainable AI methods—a brief overview. Springer, Cham, pp 13–38

Houlihan P, Creamer GG (2021) Leveraging social media to predict continuation and reversal in asset prices. Comput Econ 57(2):433–453. https://doi.org/10.1007/s10614-019-09932-9

Hu ZF, Kuflik T, Mocanu IG, Najafian S, Shulner Tal A (2021) Recent studies of XAI - review. In: Adjunct proceedings of the 29th ACM conference on user modeling, adaptation and personalization. ACM, New York, 421–431. https://doi.org/10.1145/3450614.3463354

Huang Y, Meng S (2019) Automobile insurance classification ratemaking based on telematics driving data. Decis Support Syst 127:113156. https://doi.org/10.1016/j.dss.2019.113156

Huang J, Chai J, Cho S (2020) Deep learning in finance and banking: a literature review and classification. Front Bus Res China 14(1):13. https://doi.org/10.1186/s11782-020-00082-6

Huang S, Simaan M, Tang Yi (2024) Measuring bank complexity using Xai. SSRN Electron J. https://doi.org/10.2139/ssrn.4785689

Huang CY (2018) Financial trading as a game: a deep reinforcement learning approach

Islam MR, Ahmed MU, Barua S, Begum S (2022) A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Appl Sci 12(3):1353. https://doi.org/10.3390/app12031353

Jain R, Alzubi JA, Jain N, Joshi P (2019) Assessing risk in life insurance using ensemble learning. J Intell Fuzzy Syst 37(2):2969–2980. https://doi.org/10.3233/JIFS-190078

Jain R, Vanzara R, Sarvakar K (2024) The rise of AI and ML in financial technology: an in-depth study of trends and challenges. 329–341. https://doi.org/10.1007/978-981-99-7137-4_32

Jalal N, Mehmood A, Choi GS, Ashraf I (2022) A novel improved random forest for text classification using feature ranking and optimal number of trees. J King Saud Univ Comput Inf Sci 34(6):2733–2742. https://doi.org/10.1016/j.jksuci.2022.03.012

Jeong H, Gan G, Valdez EA (2018) Association rules for understanding policyholder lapses. Risks 6(3):69. https://doi.org/10.3390/risks6030069

Ji Y (2021) Explainable AI methods for credit card fraud detection: evaluation of LIME and SHAP through a user study

Jiang X, Pan S, Long G, Xiong F, Jiang J, Zhang C (2019) Cost-sensitive parallel learning framework for insurance intelligence operation. IEEE Trans Ind Electron 66(12):9713–9723. https://doi.org/10.1109/TIE.2018.2873526

Jiang Y, Olmo J, Atwi M (2024) Deep reinforcement learning for portfolio selection. Glob Financ J 62:101016. https://doi.org/10.1016/j.gfj.2024.101016

Jiang Z, Liang J (2016) Cryptocurrency portfolio management with deep reinforcement learning

Jones S, Johnstone D, Wilson R (2015) An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. J Bank Finance 56(July):72–85. https://doi.org/10.1016/j.jbankfin.2015.02.006

Jones S, Johnstone D, Wilson R (2017) Predicting corporate bankruptcy: an evaluation of alternative statistical frameworks. J Bus Financ Acc 44(1–2):3–34. https://doi.org/10.1111/jbfa.12218

Jung Y-J, Han S-H, Choi H-J (2021) Explaining CNN and RNN using selective layer-wise relevance propagation. IEEE Access 9:18670–18681. https://doi.org/10.1109/ACCESS.2021.3051171

Jurgovsky J, Granitzer M, Ziegler K, Calabretto S, Portier P-E, He-Guelton L, Caelen O (2018) Sequence classification for credit-card fraud detection. Expert Syst Appl 100:234–245. https://doi.org/10.1016/j.eswa.2018.01.037

Kalasampath K, Spoorthi KN, Sajeev S, Kuppa SS, Ajay K, Maruthamuthu A (2025) A Literature review on applications of explainable artificial intelligence (XAI). IEEE Access 13:41111–41140. https://doi.org/10.1109/ACCESS.2025.3546681

Kapale R, Deshpande P, Shukla S, Kediya S, Pethe Y, Metre S (2024) Explainable AI for fraud detection: enhancing transparency and trust in financial decision-making. In: 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI). IEEE, 1–6. https://doi.org/10.1109/IDICAIEI61867.2024.10842874

Karamizadeh F, Zolfagharifar SA (2016) Using the clustering algorithms and rule-based of data mining to identify affecting factors in the profit and loss of third party insurance, insurance company auto. Indian J Sci Technol. https://doi.org/10.17485/ijst/2016/v9i7/87846

Kašćelan V, Kašćelan L, Burić MN (2016) A nonparametric data mining approach for risk prediction in car insurance: a case study from the montenegrin market. Econ Res-Ekonomska Istraživanja 29(1):545–558. https://doi.org/10.1080/1331677X.2016.1175729

Katsikis VN, Mourtas SD, Stanimirović PS, Li S, Cao X (2021) Time-varying mean-variance portfolio selection under transaction costs and cardinality constraint problem via beetle antennae search algorithm (BAS). Operat Res Forum 2(2):18. https://doi.org/10.1007/s43069-021-00060-5

Kenny EM, Ford C, Quinn M, Keane MT (2021) Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies. Artif Intell 294:103459. https://doi.org/10.1016/j.artint.2021.103459

Khan AH, Cao X, Katsikis VN, Stanimirovic P, Brajevic I, Li S, Kadry S, Nam Y (2020) Optimal portfolio management for engineering problems using nonconvex cardinality constraint: a computing perspective. IEEE Access 8:57437–57450. https://doi.org/10.1109/ACCESS.2020.2982195

Khan AT, Cao X, Brajevic I, Stanimirovic PS, Katsikis VN, Li S (2022a) Non-linear activated beetle antennae search: a novel technique for non-convex tax-aware portfolio optimization problem. Expert Syst Appl 197:116631. https://doi.org/10.1016/j.eswa.2022.116631

Khan AT, Cao X, Li S, Katsikis VN, Brajevic I, Stanimirovic PS (2022b) Fraud detection in publicly traded U.S. firms using beetle antennae search: a machine learning approach. Expert Syst Appl 191:116148. https://doi.org/10.1016/j.eswa.2021.116148

Khan W, Ghazanfar MA, Azam MA, Karami A, Alyoubi KH, Alfakeeh AS (2022c) Stock market prediction using machine learning classifiers and social media, news. J Ambient Intell Humaniz Comput 13(7):3433–3456. https://doi.org/10.1007/s12652-020-01839-w

Khandani AE, Kim AJ, Andrew WL (2010) Consumer credit-risk models via machine-learning algorithms. J Bank Finance 34(11):2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Khodairy MA, Abosamra G (2021) Driving behavior classification based on oversampled signals of smartphone embedded sensors using an optimized stacked-LSTM neural networks. IEEE Access 9:4957–4972. https://doi.org/10.1109/ACCESS.2020.3048915

Khoozani S, Zahra AQ, Sabri Md, Seng WC, Seera M, Eg KY (2024) Navigating the landscape of concept-supported XAI: challenges, innovations, and future directions. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-17666-y

Kim K-J (2003) Financial time series forecasting using support vector machines. Neurocomputing 55(1–2):307–319. https://doi.org/10.1016/S0925-2312(03)00372-2

Kim E-S (2020) Deep learning and principal-agent problems of algorithmic governance: the new materialism perspective. Technol Soc 63:101378. https://doi.org/10.1016/j.techsoc.2020.101378

Kim S, Woo J (2021) Explainable AI framework for the financial rating models. In: 2021 10th International conference on computing and pattern recognition. ACM, New York, 252–255. https://doi.org/10.1145/3497623.3497664

Kitchenham B (2007) Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report EBSE-2007-01

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

Kose I, Gokturk M, Kilic K (2015) An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. Appl Soft Comput 36:283–299. https://doi.org/10.1016/j.asoc.2015.07.018

Kulesza T, Burnett M, Wong WK, Stumpf S (2015) Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th international conference on intelligent user interfaces. ACM New York, 126–37. https://doi.org/10.1145/2678025.2701399

Kumar P, Hota L, Tikkiwal VA, Kumar A (2024) Analysing forecasting of stock prices: an explainable AI approach. Procedia Comput Sci 235:2009–2016. https://doi.org/10.1016/j.procs.2024.04.190

Kumar M, Ghani R, Mei ZS (2010) Data mining to predict and prevent errors in health insurance claims processing. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, 65–74.https://doi.org/10.1145/1835804.1835816

Kute DV, Pradhan B, Shukla N, Alamri A (2021) Deep learning and explainable artificial intelligence techniques applied for detecting money laundering—a critical review. IEEE Access 9:82300–82317. https://doi.org/10.1109/ACCESS.2021.3086230

Kwak BI, Han ML, Kim HK (2021) Driver identification based on wavelet transform using driving patterns. IEEE Trans Ind Inf 17(4):2400–2410. https://doi.org/10.1109/TII.2020.2999911

La Gatta V, Moscato V, Postiglione M, Sperlì G (2021a) CASTLE: cluster-aided space transformation for local explanations. Expert Syst Appl 179:115045. https://doi.org/10.1016/j.eswa.2021.115045

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

Lahmiri S (2016) Features selection, data mining and finacial risk classification: a comparative study. Intell Syst Account, Finance Manag 23(4):265–275. https://doi.org/10.1002/isaf.1395

Lamberti, WF (2023) An overview of explainable and interpretable AI. In: AI Assurance. Elsevier, 55–123. https://doi.org/10.1016/B978-0-32-391919-7.00015-9

Lariviere B, Vandenpoel D (2005) Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Syst Appl 29(2):472–484. https://doi.org/10.1016/j.eswa.2005.04.043

Larsson S, Heintz F (2020) Transparency in artificial intelligence. Internet Policy Rev. https://doi.org/10.14763/2020.2.1469

Le HH, Viviani J-L (2018) Predicting bank failure: an improvement by implementing a machine-learning approach to classical financial ratios. Res Int Bus Financ 44:16–25. https://doi.org/10.1016/j.ribaf.2017.07.104

Lepri B, Oliver N, Pentland A (2021) Ethical machines: the human-centric use of artificial intelligence. Iscience 24(3):102249. https://doi.org/10.1016/j.isci.2021.102249

Letrache K, Ramdani M (2023) Explainable artificial intelligence: a review and case study on model-agnostic methods. In: 2023 14th International conference on intelligent systems: theories and applications (SITA). IEEE, 1–8. https://doi.org/10.1109/SITA60746.2023.10373722

Li Y, Yan C, Liu W, Li M (2018) A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. Appl Soft Comput 70:1000–1009. https://doi.org/10.1016/j.asoc.2017.07.027

Li X, Xiong H, Li X, Xuanyu Wu, Zhang X, Liu Ji, Bian J, Dou D (2022) interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. Knowl Inf Syst 64(12):3197–3234. https://doi.org/10.1007/s10115-022-01756-8

Li Y, Wang S, Ding H, Chen H (2023) Large language models in finance: a survey. In: 4th ACM international conference on AI in finance. ACM, New York, 374–82. https://doi.org/10.1145/3604237.3626869

Lin KY, Liu Y, Li L, Dou R (2021) A review of explainable artificial intelligence. 574–584. https://doi.org/10.1007/978-3-030-85910-7_61

Linardatos P, Papastefanopoulos V, Kotsiantis S (2020) Explainable AI: a review of machine learning interpretability methods. Entropy 23(1):18. https://doi.org/10.3390/e23010018

Lipton ZC (2018) The mythos of model interpretability. Queue 16(3):31–57. https://doi.org/10.1145/3236386.3241340

Liu C (2024) Research on corporate financial risk prediction and early warning system based on big data analysis. 209–218. https://doi.org/10.1007/978-3-031-70598-4_20

Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Del Ser J, Guidotti R et al (2024) Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. Inf Fus 106:102301. https://doi.org/10.1016/j.inffus.2024.102301

Love PED, Fang W, Matthews J, Porter S, Luo H, Ding L (2023) Explainable artificial intelligence (XAI): precepts, models, and opportunities for research in construction. Adv Eng Inform 57:102024. https://doi.org/10.1016/j.aei.2023.102024

Lu Y-H, Lin Y-C (2024) The determinants of voluntary disclosure: integration of extreme gradient boost (XGBoost) and explainable artificial intelligence (XAI) techniques. Int Rev Financ Anal 96:103577. https://doi.org/10.1016/j.irfa.2024.103577

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in neural information processing systems, edited by I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Luo C, Desheng Wu, Dexiang Wu (2017) A deep learning approach for credit scoring using credit default swaps. Eng Appl Artif Intell 65:465–470. https://doi.org/10.1016/j.engappai.2016.12.002

Lyu L, Jiangshan Yu, Nandakumar K, Li Y, Ma X, Jin J, Han Yu, Ng KS (2020) Towards fair and privacy-preserving federated deep models. IEEE Trans Parallel Distrib Syst 31(11):2524–2541. https://doi.org/10.1109/TPDS.2020.2996273

Madapatha S, Fernando P (2024) A systematic literature review of XAI-based approaches on brain disease detection using brain mri images. In: 2024 4th international conference on advanced research in computing (ICARC). IEEE, 19–24. https://doi.org/10.1109/ICARC61713.2024.10499752

Malhi A, Knapic S, Främling K (2020) Explainable agents for less bias in human-agent decision making. 129–146. https://doi.org/10.1007/978-3-030-51924-7_8

Mandeep Agarwal A, Bhatia A, Malhi A, Kaler P, Pannu HS (2022) Machine learning based explainable financial forecasting. In: 2022 4th International conference on computer communication and the internet (ICCCI). IEEE, 34–38. https://doi.org/10.1109/ICCCI55554.2022.9850272

Maree C, Modal JE, Omlin CW (2020) Towards responsible AI for financial transactions. In: 2020 IEEE symposium series on computational intelligence (SSCI). IEEE, 16–21. https://doi.org/10.1109/SSCI47803.2020.9308456

Markus AF, Kors JA, Rijnbeek PR (2021) The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 113:103655. https://doi.org/10.1016/j.jbi.2020.103655

Martin KE (2017) Algorithms: owning mistakes & designing accountability. SSRN Electron J. https://doi.org/10.2139/ssrn.3056692

Martins T, De Almeida AM, Cardoso E, Nunes L (2024) Explainable artificial intelligence (XAI): a systematic literature review on taxonomies and applications in finance. IEEE Access 12:618–629. https://doi.org/10.1109/ACCESS.2023.3347028

Mashrur A, Luo W, Zaidi NA, Robles-Kelly A (2020) Machine learning for financial risk management: a survey. IEEE Access 8:203203–203223. https://doi.org/10.1109/ACCESS.2020.3036322

Matloob I, Khan SA, Rahman HU (2020) Sequence mining and prediction-based healthcare fraud detection methodology. IEEE Access 8:143256–143273. https://doi.org/10.1109/ACCESS.2020.3013962

Mavrepis P, Makridis G, Fatouros G, Koukos V, Separdani MM, Kyriazis D (2024) XAI for all: can large language models simplify explainable AI?

Mazhar K, Dwivedi P (2024) Decoding the black box: LIME-assisted understanding of convolutional neural network (CNN) in classification of social media tweets. Soc Netw Anal Min 14(1):133. https://doi.org/10.1007/s13278-024-01297-8

Meena R, Mishra A (2023) Need for artificial intelligence (Ai) to be explainable in banking and finance: review of Ai applications, Ai black box, Xai tools and principles

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2022) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35. https://doi.org/10.1145/3457607

Memon J, Sami M, Khan RA, Uddin M (2020) Handwritten optical character recognition (OCR): a comprehensive systematic literature review (SLR). IEEE Access 8:142642–142668. https://doi.org/10.1109/ACCESS.2020.3012542

Mienye ID, Sun Y (2023) A machine learning method with hybrid feature selection for improved credit card fraud detection. Appl Sci 13(12):7254. https://doi.org/10.3390/app13127254

Minh D, Xiang Wang H, Fen Li Y, Nguyen TN (2022) Explainable artificial intelligence: a comprehensive review. Artif Intell Rev 55(5):3503–3568. https://doi.org/10.1007/s10462-021-10088-y

Mirza N, Rizvi SKA, Naqvi B, Umar M (2024) Inflation prediction in emerging economies: machine learning and FX reserves integration for enhanced forecasting. Int Rev Financ Anal 94:103238. https://doi.org/10.1016/j.irfa.2024.103238

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

Mishra AK, Tyagi AK, Richa, Patra SR (2024) Introduction to machine learning and artificial intelligence in banking and finance. 239–290. https://doi.org/10.1007/978-3-031-47324-1_14

Moirangthem DS, Lee M (2021) Hierarchical and lateral multiple timescales gated recurrent units with pre-trained encoder for long text classification. Expert Syst Appl 165:113898. https://doi.org/10.1016/j.eswa.2020.113898

Molnar C, Freiesleben T, König G, Herbinger J, Reisinger T, Casalicchio G, Wright MN, Bischl B (2023) Relating the partial dependence plot and permutation feature importance to the data generating process. 456–479. https://doi.org/10.1007/978-3-031-44064-9_24
Montariol S, Martinc M, Pelicon A, Pollak S, Koloski B, Lončarski I, Valentinčič A (2024) Multi-task learning for features extraction in financial annual reports
Montavon G, Samek W, Müller K-R (2018) Methods for interpreting and understanding deep neural networks. Digit Signal Process 73:1–15. https://doi.org/10.1016/j.dsp.2017.10.011
Moore DH (1987) Classification and regression trees. Cytometry 8(5):534–535. https://doi.org/10.1002/cyto.990080516
Morik K, Hüppe C, Unterstein K (2002) End-user access to multiple sources: incorporating knowledge discovery into knowledge management. Intell Syst Account, Finance Manag 11(4):201–214. https://doi.org/10.1002/isaf.233
Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Interpretable machine learning: definitions, methods, and applications. https://doi.org/10.1073/pnas.1900654116
Mutlu EÇ, Yousefi N, Ozmen Garibay O (2022) Contrastive counterfactual fairness in algorithmic decision-making. In: Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society. ACM, New York, 499–507. https://doi.org/10.1145/3514094.3534143
Nallakaruppan MK, Chaturvedi H, Grover V, Balusamy B, Jaraut P, Bahadur J, Meena VP, Hameed IA (2024) Credit risk assessment and financial decision support using explainable artificial intelligence. Risks 12(10):164. https://doi.org/10.3390/risks12100164
Nazir S, Dickson DM, Akram MU (2023) Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. Comput Biol Med 156:106668. https://doi.org/10.1016/j.compbiomed.2023.106668
Neumann Ł, Nowak RM, Okuniewski R, Wawrzyński P (2019) Machine learning-based predictions of customers' decisions in car insurance. Appl Artif Intell 33(9):817–828. https://doi.org/10.1080/08839514.2019.1630151
Nizam T, Zafar S (2023) Explainable artificial intelligence (XAI): conception, visualization and assessment approaches towards amenable XAI. 35–51. https://doi.org/10.1007/978-3-031-18292-1_3
Nourallah M, Öhman P, Hamati S (2024) Financial technology and financial capability: study of the European Union. Glob Financ J 62:101008. https://doi.org/10.1016/j.gfj.2024.101008
Ohana JJ, Ohana S, Benhamou E, Saltiel D, Guez B (2021) Explainable AI (XAI) models applied to the multi-agent environment of financial markets. 189–207. https://doi.org/10.1007/978-3-030-82017-6_12
Okoli C (2023) Statistical inference using machine learning and classical techniques based on accumulated local effects (ALE)
Olah C, Mordvintsev A, Schubert L (2017) Feature visualization. Distill. https://doi.org/10.23915/distill.00007
Olden JD, Joy MK, Death RG (2004) An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol Model 178(3–4):389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013
Onasoga B, Hwidi J (2024) Enhancing credit card default prediction: prioritizing recall over accuracy. 441–59. https://doi.org/10.1007/978-981-97-3817-5_32
Owens E, Sheehan B, Mullins M, Cunneen M, Ressel J, Castignani G (2022) Explainable artificial intelligence (XAI) in insurance. Risks 10(12):230. https://doi.org/10.3390/risks10120230
Ozbayoglu AM, Gudelek MU, Sezer OB (2020) Deep learning for financial applications: a survey. Appl Soft Comput 93:106384. https://doi.org/10.1016/j.asoc.2020.106384
Ozkaya I (2020) What is really different in engineering AI-enabled systems? IEEE Softw 37(4):3–6. https://doi.org/10.1109/MS.2020.2993662
Pagliaro C, Mehta D, Shiao HT, Wang S, Xiong L (2021) Investor behavior modeling by analyzing financial advisor notes. In: Proceedings of the second ACM international conference on AI in finance. ACM, New York, 1–8. https://doi.org/10.1145/3490354.3494388
Papadimitriou T, Gogas P, Agrapetidou A (2022) The resilience of the U.S. banking system. Int J Financ Econ 27(3):2819–2835. https://doi.org/10.1002/ijfe.2300
Park MS, Son H, Hyun C, Hwang HJ (2021) Explainability of machine learning models for bankruptcy prediction. IEEE Access 9:124887–124899. https://doi.org/10.1109/ACCESS.2021.3110270
Park S, Yang J-S (2022) Interpretable deep learning LSTM model for intelligent economic decision-making. Knowl-Based Syst 248(July):108907. https://doi.org/10.1016/j.knosys.2022.108907
Pathak J, Vidyarthi N, Summers SL (2005) A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims. Manag Audit J 20(6):632–644. https://doi.org/10.1108/02686900510606119
Pawelczyk M, Broelemann K, Kasneci G (2019) Learning model-agnostic counterfactual explanations for tabular data. https://doi.org/10.1145/3366423.3380087

Popa S, Claudia D, Popa DN, Bogdan V, Simut R (2021) composite financial performance index prediction—a neural networks approach. J Bus Econ Manag 22(2):277–296. https://doi.org/10.3846/jbem.2021.14000

Raees M, Meijerink I, Lykourentzou I, Khan V-J, Papangelis K (2024) From explainable to interactive AI: a literature review on current trends in human-AI interaction. Int J Hum Comput Stud 189:103301. https://doi.org/10.1016/j.ijhcs.2024.103301

Rahim R, Chishti MA (2024) Artificial intelligence applications in accounting and finance. In: 2024 ASU international conference in emerging technologies for sustainability and intelligent systems (ICETSIS). IEEE, 1782–1786. https://doi.org/10.1109/ICETSIS61505.2024.10459526

Rai A (2020) Explainable AI: from black box to glass box. J Acad Mark Sci 48(1):137–141. https://doi.org/10.1007/s11747-019-00710-5

Rai A, Constantinides P, Sarker S (2019) Next generation digital platforms: toward human-AI hybrids. Manag Inf Syst Quart 43:3

Raisch S, Krakowski S (2021) Artificial intelligence and management: the automation–augmentation paradox. Acad Manag Rev 46(1):192–210. https://doi.org/10.5465/amr.2018.0072

Rane N, Choudhary S, Rane J (2023) Explainable artificial intelligence (XAI) approaches for transparency and accountability in financial decision-making. SSRN Electron J. https://doi.org/10.2139/ssrn.4640316

Rashid A, Asif S, Butt NA, Ashraf I (2013) Feature level opinion mining of educational student feedback data using sequential pattern mining and association rule mining. Int J Comput Appl 81(10):31–38. https://doi.org/10.5120/14050-2215

Redelmeier A, Jullum M, Aas K (2020) Explaining predictive models with mixed features using shapley values and conditional inference trees. 117–137. https://doi.org/10.1007/978-3-030-57321-8_7

Ribeiro MT, Singh S, Guestrin C (2016a) Model-agnostic interpretability of machine learning

Ribeiro MT, Singh S, Guestrin C (2016b) 'Why should i trust you?': Explaining the predictions of any classifier

Ribeiro MT, Singh S, Guestrin C (2018) Anchors: high-precision model-agnostic explanations. Proc AAAI Conf Artif Intell. https://doi.org/10.1609/aaai.v32i1.11491

Rieg T, Frick J, Baumgartl H, Buettner R (2020) Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms. PLoS ONE 15(12):e0243615. https://doi.org/10.1371/journal.pone.0243615

Roy A, Sun J, Mahoney R, Alonzi L, Adams S, Beling P (2018) Deep learning detecting fraud in credit card transactions. In: 2018 Systems and information engineering design symposium (SIEDS). IEEE, 129–134. https://doi.org/10.1109/SIEDS.2018.8374722

Rupapara V, Rustam F, Amaar A, Washington PB, Lee E, Ashraf I (2021) Deepfake tweets classification using stacked Bi-LSTM and words embedding. PeerJ Comput Sci 7:e745. https://doi.org/10.7717/peerj-cs.745

Rupapara V, Rustam F, Aljedaani W, Shahzad HF, Lee E, Ashraf I (2022) Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model. Sci Rep 12(1):1000. https://doi.org/10.1038/s41598-022-04835-6

Sahakyan M, Aung Z, Rahwan T (2021) Explainable artificial intelligence for tabular data: a survey. IEEE Access 9:135392–135422. https://doi.org/10.1109/ACCESS.2021.3116481

Saleem R, Yuan Bo, Kurugollu F, Anjum A, Liu Lu (2022) Explaining deep neural networks: a survey on the global interpretation methods. Neurocomputing 513:165–180. https://doi.org/10.1016/j.neucom.2022.09.129

Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G, Bokoro PN, Sharma R (2022) Explainable AI for healthcare 5.0: opportunities and challenges. IEEE Access 10:84486–84517. https://doi.org/10.1109/ACCESS.2022.3197671

Sarker IH (2021) Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput Sci 2(6):420. https://doi.org/10.1007/s42979-021-00815-1

Saw SN, Yan YY, Ng KH (2025) Current status and future directions of explainable artificial intelligence in medical imaging. Eur J Radiol 183:111884. https://doi.org/10.1016/j.ejrad.2024.111884

Schmid U, Finzel B (2020) Mutual explanations for cooperative decision making in medicine. KI - Künstliche Intelligenz 34(2):227–233. https://doi.org/10.1007/s13218-020-00633-2

Schmitt M, Cummins M (2023) Beyond accuracy in artificial intelligence based credit scoring systems: explainability and sustainability in decision support. SSRN Electron J. https://doi.org/10.2139/ssrn.4536400

Schwalbe G, Finzel B (2023) A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Min Knowl Discovery. https://doi.org/10.1007/s10618-022-00867-8

Sevim Ş, Yıldız B, Dalkılıç N (2016) Risk assessment for accounting professional liability insurance. Sosyoekonomi. https://doi.org/10.17233/se.2016.06.004

Sezer OB, Ozbayoglu AM (2018) Algorithmic financial trading with deep convolutional neural networks: time series to image conversion approach. Appl Soft Comput 70:525–538. https://doi.org/10.1016/j.asoc.2018.04.024

Sezer OB, Ozbayoglu M, Dogdu E (2017) A deep neural-network based stock trading system based on evolutionary optimized technical analysis parameters. Procedia Comput Sci 114:473–480. https://doi.org/10.1016/j.procs.2017.09.031

Shah P, Guez A (2009) Mortality forecasting using neural networks and an application to cause-specific data for insurance purposes. J Forecast 28(6):535–548. https://doi.org/10.1002/for.1111

Shah A, Raj P, Pushpam Kumar SP, Asha HV (2020) FinAID, a financial advisor application using AI. Int J Recent Technol Eng (IJRTE) 9(1):2282–2286. https://doi.org/10.35940/ijrte.A2951.059120

Shaheen MY (2021) Applications of artificial intelligence (AI) in healthcare: a review. https://doi.org/10.14293/S2199-1006.1.SOR-.PPVRY8K.v1

Sheehan B, Murphy F, Ryan C, Mullins M, Liu HY (2017) Semi-autonomous vehicle motor insurance: a Bayesian network risk transfer approach. Transp Res Part c: Emerg Technol 82:124–137. https://doi.org/10.1016/j.trc.2017.06.015

Shi S, Li J, Li G, Pan P, Liu K (2021) XPM. In: Proceedings of the 30th ACM international conference on information & knowledge management. ACM, New York, 1661–1670. https://doi.org/10.1145/3459637.3482494

Shi Si, Tse R, Luo W, D'Addona S, Pau G (2022) Machine learning-driven credit risk: a systemic review. Neural Comput Appl 34(17):14327–14339. https://doi.org/10.1007/s00521-022-07472-2

Siami M, Naderpour M, Jie Lu (2021) A mobile telematics pattern recognition framework for driving behavior extraction. IEEE Trans Intell Transp Syst 22(3):1459–1472. https://doi.org/10.1109/TITS.2020.2971214

Sigrist F, Hirnschall C (2019) Grabit: gradient tree-boosted tobit models for default prediction. J Bank Finance 102:177–192. https://doi.org/10.1016/j.jbankfin.2019.03.004

Smith H (2021) Clinical AI: opacity, accountability, responsibility and liability. AI Soc 36(2):535–545. https://doi.org/10.1007/s00146-020-01019-6

Smith KA, Willis RJ, Brooks M (2000) An analysis of customer retention and insurance claim patterns using data mining: a case study. J Oper Res Soc 51(5):532–541. https://doi.org/10.1057/palgrave.jors.2600941

Smyth GK, Jørgensen B (2002) Fitting tweedie's compound poisson model to insurance claims data: dispersion modelling. ASTIN Bull 32(1):143–157. https://doi.org/10.2143/AST.32.1.1020

Sohail M, Peres P, Li Y (2021) Feature importance analysis for customer management of insurance products. In: 2021 International joint conference on neural networks (IJCNN), 1–8. IEEE. https://doi.org/10.1109/IJCNN52387.2021.9533893

Soleymani F, Vasighi M (2022) Efficient portfolio construction by means of CVaR and k -means++ clustering analysis: evidence from the NYSE. Int J Financ Econ 27(3):3679–3693. https://doi.org/10.1002/ijfe.2344

Speith T (2022) A review of taxonomies of explainable artificial intelligence (XAI) methods. In: 2022 ACM conference on fairness, accountability, and transparency. ACM, New York, 2239–2250. https://doi.org/10.1145/3531146.3534639

Stahl BC (2021) Conclusion. 117–122. https://doi.org/10.1007/978-3-030-69978-9_8

Sun C, Yan Z, Li Q, Zheng Y, Xudong Lu, Cui L (2019) Abnormal group-based joint medical fraud detection. IEEE Access 7:13589–13596. https://doi.org/10.1109/ACCESS.2018.2887119

Swathi Y, Challa M (2023) A comparative analysis of explainable ai techniques for enhanced model interpretability. In: 2023 3rd international conference on pervasive computing and social networking (ICPCSN). IEEE, 229–34. https://doi.org/10.1109/ICPCSN58827.2023.00043

Talukder Md, Alamin MK, Uddin MA (2024) An integrated multistage ensemble machine learning model for fraudulent transaction detection. J Big Data 11(1):168. https://doi.org/10.1186/s40537-024-00996-5

Tao H, Zhixin L, Xiaodong S (2012) Insurance fraud identification research based on fuzzy support vector machine with dual membership. In: 2012 International conference on information management, innovation management and industrial engineering. IEEE, 457–460. https://doi.org/10.1109/ICIII.2012.6340016

Tao W, Zhu H, Tan K, Wang J, Liang Y, Jiang H, Yuan P, Lan Y (2024) FinQA: a training-free dynamic knowledge graph question answering system in finance with LLM-based revision. 418–423. https://doi.org/10.1007/978-3-031-70371-3_32

Thakur R, AlSaleh D, Hale D (2023) Digital disruption: a managers' eye view. J Busin Ind Mark 38(1):53–70. https://doi.org/10.1108/JBIM-05-2021-0273

Thanathamathee P, Sawangarreerak S, Chantamunee S, Nizam DN (2024) SHAP-instance weighted and anchor explainable AI: enhancing XGBoost for financial fraud detection. Emerg Sci J 8(6):2404–2430. https://doi.org/10.28991/ESJ-2024-08-06-016

Tian Y, Liu G (2020) MANE: model-agnostic non-linear explanations for deep learning model. In: 2020 IEEE world congress on services (SERVICES). IEEE, 33–36. https://doi.org/10.1109/SERVICES48979.2020.00021
Tillmanns S, Ter Hofstede F, Krafft M, Goetz O (2017) How to separate the wheat from the chaff: improved variable selection for new customer acquisition. J Mark 81(2):99–113. https://doi.org/10.1509/jm.15.0398
Tomsett R, Braines D, Harborne D, Preece A, Chakraborty S (2018) Interpretable to whom? A Role-based model for analyzing interpretable machine learning systems
Tyagi S (2022) Analyzing machine learning models for credit scoring with explainable AI and optimizing investment decisions
Ullah I, Rios A, Gala V, Mckeever S (2021) Explaining deep learning models for tabular data using layer-wise relevance propagation. Appl Sci 12(1):136. https://doi.org/10.3390/app12010136
Van Der Burgt J (2020) Explainable AI in banking. J Digit Bank 4(4):344–350
Van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA (2022) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 79:102470. https://doi.org/10.1016/j.media.2022.102470
Van Roy V, Vertesy D, Damioli G (2020) AI and robotics innovation. Handbook of labor, human resources and population economics. Springer, Cham, pp 1–35
Varadarajan MN, Priya S (2024) AI and ML in finance: revolutionizing the future of banking and investments. In: 2024 6th International conference on energy, power and environment (ICEPE). IEEE, 1–5. https://doi.org/10.1109/ICEPE63236.2024.10668910
Verma S, Sharma R, Deb S, Maitra D (2021) Artificial intelligence in marketing: systematic review and future research direction. Int J Inf Manag Data Insights 1(1):100002. https://doi.org/10.1016/j.jjimei.2020.100002
Viaene S, Derrig RA, Baesens B, Dedene G (2002) A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. J Risk Insur 69(3):373–421. https://doi.org/10.1111/1539-6975.00023
Viaene S, Derrig RA, Dedene G (2004) A case study of applying boosting naive bayes to claim fraud diagnosis. IEEE Trans Knowl Data Eng 16(5):612–620. https://doi.org/10.1109/TKDE.2004.1277822
Viaene S, Dedene G, Derrig RA (2005) Auto claim fraud detection using bayesian learning neural networks. Expert Syst Appl 29(3):653–666. https://doi.org/10.1016/j.eswa.2005.04.030
Vilone G, Longo L (2020) Explainable artificial intelligence: a systematic review
Viswan V, Shaffi N, Mahmud M, Subramanian K, Hajamohideen F (2024) Explainable artificial intelligence in Alzheimer's disease classification: a systematic review. Cogn Comput 16(1):1–44. https://doi.org/10.1007/s12559-023-10192-x
Wang HD (2020) Research on the features of car insurance data based on machine learning. Procedia Comput Sci 166:582–587. https://doi.org/10.1016/j.procs.2020.02.016
Wang Y, Wei Xu (2018) Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decis Support Syst 105:87–95. https://doi.org/10.1016/j.dss.2017.11.001
Wang J, Sun T, Liu B, Cao Y, Zhu H (2021) CLVSA: a convolutional LSTM based variational sequence-to-sequence model with attention for predicting trends of financial markets. https://doi.org/10.24963/ijcai.2019/514
Watson D (2022) Rational shapley values. In: 2022 ACM Conference on fairness, accountability, and transparency. ACM, New York, 1083–1094. https://doi.org/10.1145/3531146.3533170
Weber P, Valerie Carl K, Hinz O (2024) Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. Manag Rev Quart 74(2):867–907. https://doi.org/10.1007/s11301-023-00320-0
Wen Q, Zhou T, Zhang C, Chen W, Ma Z, Yan J, Sun L (2022) Transformers in time series: a survey
West D (2000) Neural network credit scoring models. Comput Oper Res 27(11–12):1131–1152. https://doi.org/10.1016/S0305-0548(99)00149-5
White A, Garcez A (2019) Measurable counterfactual local explanations for any classifier
Wu TY, Wang YT (2021) Locally interpretable one-class anomaly detection for credit card fraud detection. In: 2021 International conference on technologies and applications of artificial intelligence (TAAI). IEEE, 25–30. https://doi.org/10.1109/TAAI54685.2021.00014
Xiao B, Benbasat I (2007) E-commerce product recommendation agents: use, characteristics, and impact. MIS Q 31(1):137. https://doi.org/10.2307/25148784
Xiao J, Zhong Yu, Jia Y, Wang Y, Li R, Jiang X, Wang S (2024) A novel deep ensemble model for imbalanced credit scoring in internet finance. Int J Forecast 40(1):348–372. https://doi.org/10.1016/j.ijforecast.2023.03.004
Xu D, Zhang X, Feng H (2019) Generalized fuzzy soft sets theory-based novel hybrid ensemble credit scoring model. Int J Financ Econ 24(2):903–921. https://doi.org/10.1002/ijfe.1698

Yan K, Li Y (2024) Machine learning-based analysis of volatility quantitative investment strategies for american financial stocks. Quant Finance Econ 8(2):364–386. https://doi.org/10.3934/QFE.2024014

Yang G, Ye Q, Xia J (2022) Unbox the black-box for the medical explainable ai via multi-modal and multicentre data fusion: a mini-review, two showcases and beyond. Inf Fus 77:29–52. https://doi.org/10.1016/j.inffus.2021.07.016

Yang Y, Uy MC, Huang A (2020) FinBERT: a pretrained language model for financial communications

Ye Y, Pei H, Wang B, Chen PY, Zhu Y, Xiao J, Li B (2020) Reinforcement-learning based portfolio management with augmented asset movement prediction states. Proc AAAI Conf Artif Intell 34(01):1112–1119. https://doi.org/10.1609/aaai.v34i01.5462

Yeo WJ, van der Heever W, Mao R, Cambria E, Satapathy R, Mengaldo G (2023) A comprehensive review on financial explainable AI

Yin H, Xingying Wu, Kong SX (2022) Daily investor sentiment, order flow imbalance and stock liquidity: evidence from the chinese stock market. Int J Financ Econ 27(4):4816–4836. https://doi.org/10.1002/ijfe.2402

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. 818–833. https://doi.org/10.1007/978-3-319-10590-1_53.

Zhang Z, Zohren S, Roberts S (2020) Deep learning for portfolio optimization. J Financ Data Sci 2(4):8–20. https://doi.org/10.3905/jfds.2020.1.042

Zhang Y, Chu G, Shen D (2021) The role of investor attention in predicting stock prices: the long short-term memory networks perspective. Financ Res Lett 38:101484. https://doi.org/10.1016/j.frl.2020.101484

Zhang Z, Chong Wu, Shiyou Qu, Chen X (2022) An explainable artificial intelligence approach for financial distress prediction. Inf Process Manag 59(4):102988. https://doi.org/10.1016/j.ipm.2022.102988

Zhang B, Kong D (2020) Dynamic estimation model of insurance product recommendation based on naive bayesian model. In: Proceedings of the 2020 international conference on cyberspace innovation of advanced technologies. ACM, New York, 219–224. https://doi.org/10.1145/3444370.3444575

Zhao Y, Stasinakis C, Sermpinis G, Shi Y (2018) Neural network copula portfolio optimization for exchange traded funds. Quant Finance 18(5):761–775. https://doi.org/10.1080/14697688.2017.1414505

Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, Wang S, Yin D, Mengnan Du (2024a) Explainability for large language models: a survey. ACM Trans Intell Syst Technol 15(2):1–38. https://doi.org/10.1145/3639372

Zhao H, Liu Z, Wu Z, Li Y, Yang T, Shu P, Xu S, Dai H, Zhao L, Mai G, Liu N et al. (2024b) Revolutionizing finance with LLMs: an overview of applications and insights

Zhou Z, Hu M, Salcedo M, Gravel N, Yeung W, Venkat A, Guo D, Zhang J, Kannan N, Li S (2023) XAI meets biology: a comprehensive review of explainable AI in bioinformatics applications

Zolotareva E (2021) Aiding long-term investment decisions with XGBoost machine learning model

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Farhina Sardar Khan[1] · Syed Shahid Mazhar[2] · Kashif Mazhar[3] · Dhoha A. AlSaleh[4] · Amir Mazhar[2]

✉ Syed Shahid Mazhar
shahid.dphil@gmail.com

✉ Dhoha A. AlSaleh
dhoha.alsaleh@aasu.edu.kw

Farhina Sardar Khan
farhina.skhan05@gmail.com

Kashif Mazhar
kashif.mazhar@ddn.upes.ac.in

Amir Mazhar
amirmazhar126@gmail.com

[1] Department of Commerce, Integral University, Lucknow, UP, India

[2] Department of Business Management, Integral University, Lucknow, UP, India

[3] School of Computer Science, The University of Petroleum and Energy Studies (UPES), Dehradun, Uttarakhand, India

[4] College of Business and Entrepreneurship, Abdullah Al Salem University, Kuwait City, Kuwait