



Kent Academic Repository

Lyu, Qi and Wu, Shaomin (2025) *Explainable artificial intelligence for business and economics: methods, applications and challenges*. *Expert Systems*, 42 (4). ISSN 0266-4720.

Downloaded from

<https://kar.kent.ac.uk/108904/> The University of Kent's Academic Repository KAR

The version of record is available from

<https://doi.org/10.1111/exsy.70017>

This document version

Publisher pdf

DOI for this version

Licence for this version

CC BY (Attribution)

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in **Title of Journal**, Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).

REVIEW OPEN ACCESS

Explainable Artificial Intelligence for Business and Economics: Methods, Applications and Challenges

Qi Lyu | Shaomin Wu 

Kent Business School, University of Kent, Canterbury, Kent, UK

Correspondence: Shaomin Wu (s.m.wu@kent.ac.uk)**Received:** 25 September 2024 | **Revised:** 23 December 2024 | **Accepted:** 29 January 2025**Keywords:** business and economics | explainable artificial intelligence | machine learning | new challenges

ABSTRACT

In recent years, artificial intelligence (AI) has made significant strides in research and shown great potential in various application fields, including business and economics (B&E). However, AI models are often black boxes, making them difficult to understand and explain. This challenge can be addressed using eXplainable Artificial Intelligence (XAI), which helps humans understand the factors driving AI decisions, thereby increasing transparency and confidence in the results. This paper aims to provide a comprehensive understanding of the state-of-the-art research on XAI in B&E by conducting an extensive literature review. It introduces a novel approach to categorising XAI techniques from three different perspectives: samples, features and modelling method. Additionally, the paper identifies key challenges and corresponding opportunities in the field. We hope that this work will promote the adoption of AI in B&E, inspire interdisciplinary collaboration, foster innovation and growth and ensure transparency and explainability.

1 | Introduction

Artificial Intelligence (AI) technologies have seen a surge of research interest in recent years and are increasingly used in various fields of business and economics (B&E) (see, e.g., Johnson et al. 2022). Specifically, in the context of AI, B&E are referred in particular to the disciplines where AI technologies are applied to improve decision-making effectiveness, optimise business processes and provide strategic insights. AI serves as an efficient tool in those fields, enabling data-driven decision-making and offering innovative solutions to complex problems.

- *Business:* It refers to the activities undertaken by individuals or organisations to produce, buy, or sell products and services with the goal of generating value, achieving operational efficiency, and meeting consumer demands. In the context of AI, business applications focus on leveraging AI technologies to improve productivity, innovation, and competitiveness (Bharadiya 2023).

- *Economics:* It is the study of how resources are produced, distributed, and consumed, focusing on decision-making at individual, organisational, and societal levels. In AI, economics explores how intelligent systems impact markets, labour, and public policy, while optimising outcomes in resource allocation (Varian 2018).

With advancements in data collection technology, B&E data has grown exponentially, leading to more AI applications. The rapidly evolving business landscape and workforce dynamics have made AI an integral part of daily business operations. According to Ransbotham et al. (2017), 85% of CEOs believe that AI creates new opportunities for B&E, although 40% also express concerns about the risks associated with AI models.

Numerous scholarly papers have explored AI applications in B&E, focusing on finance, healthcare management, human resource management, marketing and supply chain management in the business domain, and macroeconomics and microeconomics in the economics domain. Table 1 provides examples of

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Expert Systems* published by John Wiley & Sons Ltd.

AI applications in B&E. Many AI models function as black boxes, making them difficult for humans to understand and trust. In practice, it is often crucial for users to comprehend how an AI model derives a particular decision, especially when the decision has high stakes. While AI undoubtedly benefits B&E, there is a risk in blindly trusting or applying the recommendations,

insights or predictions provided by AI models. Therefore, it is essential to understand the reasoning and logic behind these models, leading to the development of eXplainable AI (XAI) methods.

There is no strict definition of XAI. A widely cited definition suggests that XAI aims to improve 'the degree to which a human can understand the cause of a decision' (Miller 2019). Generally, XAI is a field of AI focused on developing systems that make AI models more interactive and transferable, and their results more accessible, reliable, causal, fair, informative and trustworthy. A more detailed description of these concepts is provided in Table 2.

It is worth mentioning that the concepts of explainability, interpretability, and transparency are quite ambiguous, which may confuse the beginners of XAI. To describe those concepts, we use an example from the financial industry where a bank uses an AI model to assess loan applications. In this scenario, we will look at how explainability, interpretability, transparency and trust play out in a real B&E background.

Generally, *explainability* refers to the ability to explain the internal mechanisms of an AI model or the rationale behind its decisions in a way that is understandable to humans (Roscher et al. 2020). In this financial example, explainability means that the AI model can provide specific, understandable reasons for approving or denying a loan application. For instance, if the

TABLE 1 | Cases of the applications of AI for B&E.

Field	Application	Reference
Business	Credit scoring	Cao (2022)
	Fraud detecting	Zarifis et al. (2023)
	Candidate selection	Chowdhury et al. (2023)
	Inventory management	Toorajipour et al. (2021)
	Advertising management	Ford et al. (2023)
	Consumer service	Vaid et al. (2023)
	Recommendation system	da Silva et al. (2023)
Economics	Resource allocation	Yudkowsky (2013)
	Policy making	Acemoglu (2024)

TABLE 2 | Goals of XAI Ali et al. (2023); Arrieta et al. (2020).

Goal	Description	Target audience
Accessibility	Accessibility refers to the involvement of (non-technical) end users in the AI modelling process.	④⑥
Confidence	Confidence describes the robustness and stability of a model, including its working regime.	①②③
Causality	Causality among data variables means finding cause-effect relationships leading to higher model comprehensiveness.	①④⑤⑦
Fairness	Fairness tries to prohibit the unfair or unethical use of model results and outputs by ethical analysis and illumination of results affecting relations.	④⑦
Informativeness	Informativeness is concerned with the distinction between the original human decision-making problem and the problem solved by a given model, including its inner mechanisms.	①⑤
Interactivity	Interactivity deals with the level of interaction between end users and XAI models to improve the latter.	①④
Privacy	Privacy awareness is about enlightening possible breaches by informing users.	④⑦
Regulatory Compliance	Regulatory compliance refers to the adherence of AI systems and their applications to the legal, ethical, and policy frameworks set forth by regulatory bodies.	⑦
Transferability	Transferability deals with uncovering boundary constraints of models to better assess their applicability in other cases.	①⑤
Trustworthiness	Trustworthiness refers to the degree of confidence a model will react as expected when opposing a specific problem.	①④

Note: ①: domain experts; ②: domain developers; ③: domain managers; ④: users of the model affected by decisions; ⑤: data scientists; ⑥: product owners, managers; ⑦: regulatory entities or agencies.

model denies a loan, the bank might explain that the decision was based on the applicant's high debt-to-income ratio and recent credit score decline.

Interpretability is the extent to which a human can understand how a model makes decisions or how different inputs lead to certain outputs (Murdoch et al. 2019). Specifically, it means how much the bank's loan officers and risk managers understand the overall structure and functioning of the model. For a model to be interpretable, it might need to be simpler or have visualizable pathways that clearly show how different variables interact to lead to different outcomes.

Transparency refers to the openness and accessibility of information about an AI model's design, structure, and decision-making processes (Roscher et al. 2020). In the bank example, it means that the bank openly shares information about how the loan approval model was designed, the factors it considers and any limitations it might have.

XAI is becoming a focal research field within AI. This shift is driven by real-world demands across different industries, including B&E like healthcare, finance, and insurance. In other words, XAI can be seen as the future of AI, which means XAI can:

- Reduce regulatory pressure: Governments and regulatory departments are asking for more transparency in AI applications, making XAI essential for compliance.
- Improve trust and adoption: XAI helps build trust between humans and AI, which is crucial for the widespread adoption of AI across B&E.
- Mitigate bias and promise fairness: XAI tools will play a vital role in identifying and mitigating bias in AI models, ensuring the ethical of AI applications.

- Improve performance optimization: XAI enables developers to debug and optimise AI systems, improving their overall performance and robustness.
- Keep security: XAI can be applied to improve the security of AI applications by revealing potential vulnerabilities to adversarial attacks.

The research on XAI has been growing exponentially, which has provided a basis for some literature review studies. The existing literature review papers focus on introducing XAI methods just from a technical perspective, or a XAI taxonomy perspective (see, e.g., Tchuente et al. (2024); Černevičienė and Kabašinskas (2024)). However, XAI is a highly applied research field, and it is also very important to study the specific applications of XAI in B&E to derive best practices for better implementation and adoption, which motivates the writing of this paper. Compared with the existing literature review, this paper is the first literature review work that considers both the technical perspective and applications. For the domain experts in B&E, this paper will help them to gain a holistic understanding of XAI, comprehend the importance of XAI in B&E, culture their ability to choose the appropriate XAI methods, and then apply the XAI methods into industry applications.

In this study, we followed the PRISMA reporting guidelines, which primarily provide guidance for the reporting of systematic reviews evaluating the effects of interventions, to relevant literature between 2018 and 2024. Initially, we retrieved 8878 records from Scopus, PubMed, IEEE Xplore. After removing 2891 duplicates, 5987 records were screened based on title and abstract, excluding 3234 records due to irrelevance. Subsequently, 2753 full-text articles were assessed, and 2324 were further excluded as they do not meet eligibility criteria. Finally, 579 studies were included in our systematic review (Figure 1).

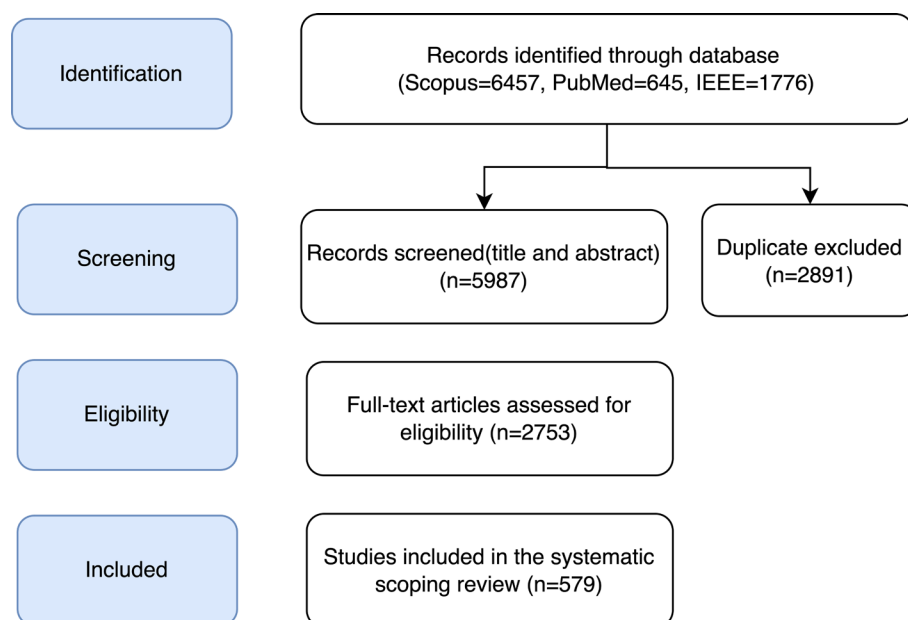


FIGURE 1 | PRISMA diagram.

The contributions of this paper include

- To gain a comprehensive understanding of state-of-the-art research on XAI in B&E, this paper conducts a thorough literature review on the technological development of XAI in B&E.
- It introduces a novel approach to categorising XAI techniques from three different perspectives: samples, features and modelling method.
- Additionally, the paper identifies key challenges and research opportunities in the field.

The remainder of this paper is structured as follows. Section 2 reviews the development in XAI on B&E and proposes a new taxonomy in XAI. Section 3 discusses the applications of XAI for B&E. Section 4 identifies challenges and opportunities in the research of XAI in B&E. Section 5 concludes the paper.

2 | XAI Techniques

Denote $\mathbf{Z} = (\mathbf{X}, \mathbf{y})$ as the dataset on which an AI model is trained, where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^\top$, $\mathbf{y} = (y_1, y_2, \dots, y_n)^\top$, and n is the sample size. That is, $\mathbf{X}$ is the n observations of the m predictors, and $\mathbf{y}$ is the n observations of the dependent variable. Suppose that each sample has m features, and $\mathbf{x}_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$ (with $i \in \{1, 2, \dots, n\}$). The i th sample is denoted by $(\mathbf{x}_i, y_i)$. Then, the objective of a supervised learning is to find a function $f(\mathbf{X})$,

$$\begin{aligned} \hat{\mathbf{y}} &= f(\mathbf{X}) \\ &= (f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_n)) \end{aligned} \quad (1)$$

to minimise the cost function $g(f(\mathbf{X}), \mathbf{y})$, where $\hat{\mathbf{y}}$ is an estimate of $\mathbf{y}$, the function g can be a measure of the distance between $f(\mathbf{X})$ and $\mathbf{y}$, such as the mean squared error $\frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2$.

The AI model, $\hat{\mathbf{y}} = f(\mathbf{X})$, can take a simple form, such as a linear regression model with small m or a shallow decision tree, or a complex form, such as a deep learning model with billions of parameters. The former is referred to as transparent models and the as black-box models. As mentioned in Section 1, XAI aims to improve users' understanding of $\hat{\mathbf{y}} = f(\mathbf{X})$.

Many XAI techniques have been developed for various purposes. To better understand and evaluate the different types of XAI techniques and their applications in B&E, this section reviews existing XAI techniques and proposes a new taxonomy.

Unlike conventional taxonomies, which may focus on specific needs faced by B&E, our taxonomy offers a unique and more practical perspective. As illustrated in Figure 2, the proposed taxonomy bridges the gap between XAI research and real-world applications, facilitating the use of XAI in B&E.

As mentioned at the beginning of this section, an AI model involves samples (i.e., $\mathbf{Z}$), features (i.e., $\mathbf{X}$), types of the model $f(\mathbf{X})$, and estimating the parameters in the model.

- A *sample* represents an observation for an AI model, which can be, for instance, a consumer's online shopping transaction. To understand consumers' shopping preferences, it is useful to build an AI model based on the samples.
- A *feature* is all samples' specific common attribute. For instance, in mortgage applications, a borrower's income can be an important feature in the AI model for credit scoring.
- A *Modelling method* is the specific model structure and calculation or function behind AI models. It includes two subclasses: transparent structure and inference mechanism. The transparent structure offers intuitive and understandable structure for the whole model, thus makes model understandable. Inference mechanism offers the internal

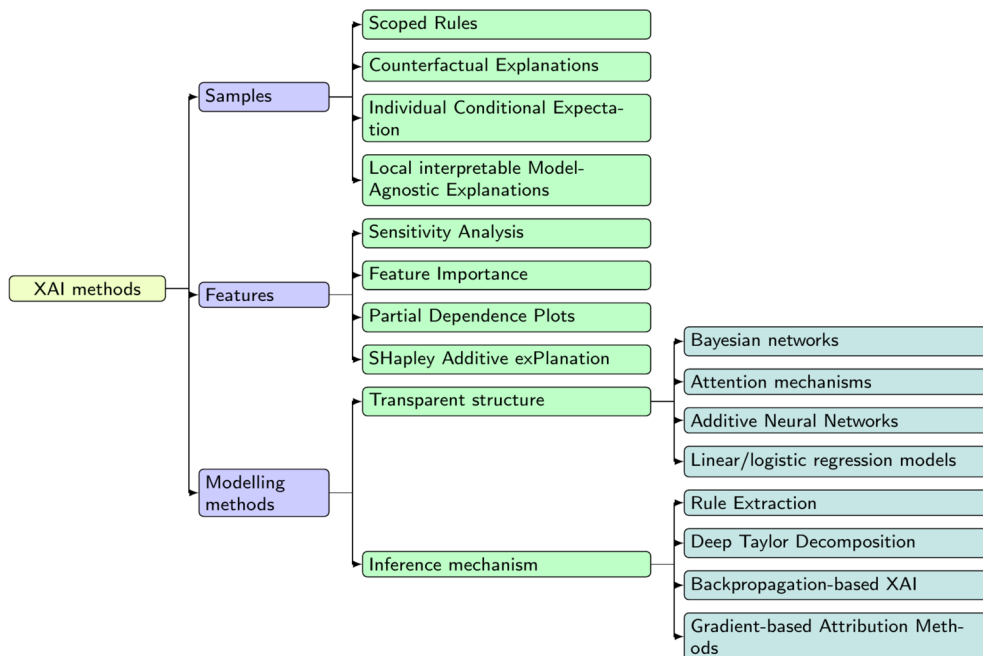


FIGURE 2 | A new taxonomy.

computational logic of AI models to make model understandable. Suppose that AI model is a black-box, we know nothing about the model, transparent structure is a glass box that makes the things in the box can be seen and understood from outside, and inference mechanism is a tool that can help to open black-box.

Therefore, an XAI method can be categorised into one of the three classes: Samples-based, Features-based, and modelling method-based, as shown in Figure 2, where samples-based XAI aims to explain the model based on each sample $\mathbf{x}_i$; features-based XAI corresponds to explain AI model based on each feature $x_{i,j}$; and modelling method-based XAI further explains the modelling method of f .

The following subsection provides a more detailed introduction and comprehensive review of the samples-based XAI, features-based XAI and modelling method-based XAI techniques, respectively.

2.1 | Samples-Based XAI Techniques

In this section, we will review samples-based XAI techniques and identify knowledge gaps.

2.1.1 | The Techniques

Sample-based XAI techniques focus on explaining individual predictions or results made by AI models. These methods aim to provide insights into why a specific prediction was made for a particular sample, enhancing transparency and interpretability. Most sample-based XAI techniques can be considered local explanations within the interpretability scope. Below are some examples of sample-based XAI techniques:

Scoped Rules (SR): Also known as Local Rules or Sample-specific Rules, these rule-based models generate easily understandable and interpretable rules for individual samples or subsets of data (van der Waa et al. 2021). SR are tailored to explain the behaviour of the model for a particular sample or a local region of the feature space. They are particularly useful in applications where understanding the rationale behind individual predictions is crucial, such as healthcare, finance, and autonomous systems.

Counterfactual Explanations (CE): CE provide insights by generating alternative samples where the prediction changes (Wachter et al. 2017). These explanations answer the question, “What changes to the input features would result in a different model’s prediction?” Freiesleben (2022) discusses the relationship between CE and adversarial examples, finding that adversarial examples can be seen as misclassified counterfactuals. Carrizosa et al. (2024) addresses a more general setting in which a group of CEs is sought for a group of samples. CE are particularly useful in sensitivity analysis, where understanding the factors influencing model’s predictions is critical.

Individual Conditional Expectation (ICE): ICE is a graphical visualisation technique used to understand the relationship

between a feature and the model’s predictions for individual samples (Goldstein et al. 2015). Studies such as Fan et al. (2023) exploit ICE to provide visual strategies and explanations for financial distress prediction and risk assessment in auditing. ICE curves explore how predictions change as a specific feature varies while keeping all other features constant, facilitating model interpretation and building trust in AI systems.

Local Interpretable Model-agnostic Explanation (LIME): LIME is a model-agnostic method proposed by Ribeiro et al. (2016a) that explains the predictions of any classifier by perturbing the input and observing the changes in the output. LIME generates local surrogate models around a specific sample to explain its prediction. Numerous works have explored LIME’s applications and theoretical foundations. Additionally, Li et al. (2023) have developed improved LIME models such as BMB-LIME, S-LIME, and G-LIME, respectively.

Furthermore, Garreau and Luxburg (2020) discuss the theoretical analysis of LIME, deriving closed-form expressions for the coefficients of the interpretable model when the function to explain is linear, and proving LIME’s effectiveness in discovering meaningful features.

Overall, samples-based XAI techniques help improve the interpretability and trustworthiness of AI models by providing explanations for individual predictions. This enables users to understand the reasoning behind specific decisions made by the model and has already been applied to B&E problems (Wang et al. 2020).

2.1.2 | Comments on Samples-Based XAI Techniques

Based on the previous discussion, users can gain explanations using sample-based XAI methods, which can be easily understood by laypeople. However, there are still significant knowledge gaps and unresolved problems in the field of sample-based XAI and corresponding challenges.

Sample dependence: Samples-based XAI methods offer explanations based on selected samples, where the explanations deeply depend on selected samples. Thus, sample engineering is vital for samples-based XAI methods. One promising research opportunity is to develop tools for choosing the suitable samples for reducing sample bias.

Stability and consistency of explanations: When applying XAI methods in various fields, explanations for different scenarios at different times must be consistent. Sample-based XAI methods may show inconsistencies between different samples due to the randomness of model training, resulting in varied interpretations of similar samples. Research on improving the consistency and stability of interpretations is crucial. Thus, the stability and reliability of explanation methods require further study, and developing stable explanation generation algorithms is necessary.

Misleading explanations: Sample-based XAI can offer explanations for individual users. Inaccurate or misleading explanations may cause users to misunderstand the model. To avoid this, XAI

methods should verify the accuracy and effectiveness of the explanations through experiments and tests. Providing multi-angle explanations to enhance comprehensive understanding is a promising research opportunity.

Computational resource consumption: Due to the increasingly accumulated data, high-quality explanations for sample-based XAI may require significant computing resources. Developing interpretation methods that can efficiently handle large-scale data and complex models is a challenge. Techniques like LIME and SHAP often have high computational complexity, making them difficult to use in large-scale applications. Therefore, developing more efficient algorithms to reduce computing resource consumption is a promising research opportunity, potentially involving more efficient computing architectures and technologies like distributed computing.

2.2 | Features-Based XAI Techniques

This section will review features-based XAI techniques and identify knowledge gaps.

2.2.1 | The Techniques

Feature-based XAI techniques focus on explaining AI models by analyzing the contribution of individual features to the models' predictions. These methods provide insights into how each feature influences the decision-making process of the model.

Feature Importance (FI): is a fundamental technique in XAI that helps users understand decisions and predictions from a feature's perspective. By quantifying the importance of features, users can identify which features are most influential in driving the predictions. Permutation Importance (PI) is a popular FI technique that measures the change in a model's performance when the values of a feature are randomly shuffled (Altmann et al. 2010). Besides interpretability, PI is also a powerful and widely used feature selection method for AI models in B&E (Hapfelmeier et al. 2023).

Partial Dependence Plots (PDPs): are another important technique for valuing FI. They provide a tool for visualising the relationship between a feature and the model's predictions (Greenwell et al. 2017). PDPs and PI can be used together to assess feature importance: PI provides a quantitative measure, while PDPs offer visual insights. *Accumulated Local Effect (ALE)* plots are another approach for visualising feature effects, showing how a model's predictions change as a feature varies (Apley and Zhu 2020). While PDPs provide a global view of feature effects, ALE plots offer a local, nuanced understanding, especially regarding feature interactions and non-linearities.

SHAP (SHapley Additive exPlanation): SHAP, proposed by Lundberg and Lee (2017), is a widely used feature importance technique built on the Shapley value from cooperative game theory. It assigns a value to each player (feature) based on their contribution to the total payoff (model's prediction).

SHAP has been applied to complex scenarios in B&E. For example, Buyuktepe et al. (2023) developed application cases of fraud risk prediction using XAI methods, where SHAP values highlighted the most important features for this fraud risk detection.

It should be noted that these different categories are not strictly mutually exclusive: one XAI method can belong to different categories. For example, LIME can be categorised into the features-based XAI, and SHAP is also regarded as a samples-based XAI method.

Sensitivity analysis (SA): SA studies how the outputs of a system are related to and influenced by its inputs (Razavi et al. 2021). SA and feature-based XAI share the goal of understanding feature impact. Sobol' (1990) proposed an output variance SA method-based on ANOVA decomposition, which has been widely used (Sobol' et al. 2011). The Fourier Amplitude Sensitivity Test (FAST) is another established SA technique (Cukier et al. 1973), defining sensitivity based on conditional variances to indicate individual or joint effects of inputs on the output. FAST computes the "main effect" contribution of each input feature to the output variance.

Feature interactions: are the contextual dependence between features that jointly impact predictions. Tsang et al. (2018) develop a framework for detecting statistical interactions captured by a feed-forward multi-layer neural network through directly interpreting its learned weights. Janizek et al. (2021) present another feature interaction XAI method (named integrated Hessians), an extension of integrated Gradients, proposed by Tsang et al. (2018), that explains pairwise feature interactions in neural networks. Integrated Hessians overcomes several theoretical limitations of previous methods, and is not limited to a specific architecture or class of neural network. Tsang et al. (2020) propose an interaction attribution and detection framework called Archipelago, which is interpretable, non-model-specific, axiomatic and scalable in real-world settings.

In summary, features-based XAI technology is currently the most widely accepted and used method. Due to its easy-to-use property and strong understandability, it has been widely used in B&E problems.

2.2.2 | Comments on Features-Based XAI Techniques

The aforementioned feature-based XAI methods provide explanations by evaluating the contribution of features to model's predictions and decisions, identifying the most important features for decision-making. Compared to sample-based XAI methods, feature-based XAI techniques are more suitable for big data analysis, offering insights into the overall behaviour of AI models and the importance of features.

However, these methods still face many knowledge gaps and challenges:

Explainability in high-dimensional data: With the development of digital B&E, feature dimensions are increasing exponentially. Effectively generating explanations for high-dimensional

data is crucial and challenging. Some work focuses on high-dimensional data in fields like computer vision and natural language processing (Kenny and Keane 2021), but these methods are difficult to present directly to users. Research into high-dimensional data XAI methods, such as dimensionality reduction and visualisation techniques, is promising.

Balance between local and global explanations: Most current feature-based explanation methods focus on local explanations, that is, single prediction explanations, lacking a global perspective. More research is needed to integrate local and global explanations, providing a comprehensive understanding of the model while maintaining the accuracy of local explanations.

Stability and consistency of explanations: Similar to sample-based XAI methods, feature-based XAI methods face challenges in stability and consistency. Explanation results should not fluctuate greatly with small changes in input data. However, many current methods do not perform well in this regard. Improving the stability and robustness of explanation methods and developing methods to quantify and improve the consistency between different explanation methods are crucial.

Handling model complexity and nonlinearity: Interpreting complex and highly nonlinear models (such as deep learning models) remains unsolved. Existing methods often struggle to explain these models. There is a need for methods that capture nonlinear relationships and provide intuitive interpretations. Addressing the interpretation complexity brought by high-dimensional data and making the interpretation concise and understandable is essential.

Feature-based XAI methods are valuable in XAI for B&E, but they still have many knowledge gaps and challenges. Solving these problems requires interdisciplinary research collaboration, combining theory with practice, and continuously improving and innovating interpretation methods to meet the needs of different fields and applications.

2.3 | Modelling Method-Based XAI Techniques

Modelling method-based XAI techniques exploit the modelling method of an AI model to improve the model's interpretability. It includes two types: transparent structure and inference mechanism. In this section, we will introduce them, respectively.

2.3.1 | Transparent Structure-Based XAI Techniques

Transparent structure-based XAI refers to methods that leverage the inherent structure or architecture of AI models to provide explanations for their decisions. This can involve understanding of how the architecture of an AI model influences its decision-making process, or extracting interpretable information directly from the transparent structure.

The most classical transparent structure-based XAI modelling method is the Linear/Logistic Regression model (LR) (Ribeiro et al. 2016b). LR models assume a linear relationship between the predictor and dependent variables, making the model

interpretable. Decision trees are another classical transparent model that easily fulfils every transparency constraint due to their hierarchical structures (Mahbooba et al. 2021). It is worth noting that some researchers may do not think LR as XAI method, because it is a machine learning model which is transparent and easily to be understood by consumers, and some researchers think these kinds of transparent machine learning model as XAI method, because they can be applied as transparent agent models for black-box AI models (Zhang et al. 2021a; 2021b).

Decision trees partition the feature space based on a series of binary decisions, providing transparent explanations in the form of decision paths that show which features were considered and how they influenced the final decision. It is worth noting that these classical transparent structures have been used as agent models for black-box AI models, known as model-agnostic approaches.

Attention Mechanisms: were originally developed for natural language processing tasks (Vaswani et al. 2017). They have since been adapted for various AI models in B&E (Yilmaz and Esra Buyuktahtakin 2024), including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and graph neural networks (GNNs), to improve interpretability and understanding. Attention mechanisms calculate the weights of features, which can be interpreted as importance scores. High attention weights indicate that the corresponding input features/elements strongly influence the model's decisions/results, making them crucial for understanding how the model arrives at its predictions. Fukui et al. (2019) proposes the Attention Branch Network (ABN), which extends an explanation model by introducing a branch structure with an attention mechanism, focusing on the attention map for visual explanation and representing high response values as attention locations in image recognition. Liu et al. (2023) explore using attention as guidance to combine explanatory information in object detector models to best present an XAI saliency map that is interpretable (plausible) to humans. Wang et al. (2021) uses self-attention mechanisms and an interpretable one-dimensional CNN to generate intra-class and inter-class explanations of the model's actions. The intra-class explanation captures the relative importance of different features within that class, while the inter-class explanation captures the relative importance between classes.

Despite significant research on XAI based on attention mechanisms, there remains a debate about 'Is Attention Interpretable?' (Serrano and Smith 2019). To this day, there is no definitive answer to this question.

Additive neural networks are a class of AI models that decompose the prediction into a sum of contributions from individual features or basis functions (Agarwal et al. 2021). These neural additive models (NAM) are interpretable and easy to understand, revealing the relationships between input features and the output. Friedman and Popescu (2008) introduce rule ensembles, a class of models that includes additive logistic regression as a special case. It presents methods for fitting additive models to classification tasks and providing interpretable predictions. Harezlak et al. (2018) presents generalised additive models (GAMs), which extend linear models by allowing nonlinear

relationships between the predictors and the response. GAMs are based on the principle of additivity and are widely used in regression analysis (Wood 2017). Kotschieder et al. (2015) introduces deep neural decision forests (NDFs), combining decision forests with deep neural networks. NDFs allow for the integration of additive components into deep learning architectures, providing both interpretability and flexibility.

Transparent structure-based XAI techniques use transparent architectures and inner working theories to provide explanations for predictions, allowing users to understand model decisions and predictions. In conclusion, transparent structure-based XAI techniques play a crucial role in enhancing the transparency and interpretability of AI models. By providing transparent, modelling method-based insights into AI model predictions, these methods enable users to validate model behaviour, identify potential biases and ensure fairness in AI systems.

2.3.2 | Inference Mechanism-Based XAI Techniques

Inference mechanism-based XAI techniques refer to a category of methods that explain the behaviours or decisions of AI models by internal structures, such as activation functions or optimization processes. These techniques focus on how the model transforms input data into output predictions/results, providing insights into the logistics of structures (i.e., optimizations, calculations) in AI models.

The main optimization processes of AI models are backpropagation-based, where backpropagation calculates the gradients of the loss function with respect to the model's parameters, enabling the model to adjust its weights during training. Thus, an intuitive approach to inference mechanism-based XAI is to understand black-box models through backpropagation.

Gradient-based Attribution Methods compute the gradient of a model's output with respect to the input features to determine each feature's contribution to the model's decision (Ancona et al. 2017). Methods such as the vanilla gradient (Simonyan et al. 2013), smooth gradient (Smilkov et al. 2017), integrated gradient (Sundararajan et al. 2017) and Grad-CAM (Selvaraju et al. 2017) all leverage gradients to provide explanations for model decisions. They differ in their specific gradient calculation approaches. For instance, the vanilla gradient method directly computes the gradients of the output class score with respect to the input features, while the smooth gradient method averages gradients. The choice of these methods often depends on the nature of the model, the type of input data, and the desired level of interpretability and robustness for the task at hand.

Backpropagation-based XAI methods leverage the backpropagation algorithm to gain insight into the decision-making process of complex neural networks. These methods aim to explain the behaviour of neural networks by attributing the contribution of input features or neurons to the model's predictions. Montavon et al. (2017) develop an approach to interpret multi-layer neural networks by decomposing the network classification decision into the contributions of its input elements, exploiting the network structure by backpropagating the interpretation from the output layer to the input layer. Shrikumar et al. (2017) propose

the Deep Learning Important Features (DeepLIFT) method, which decomposes the output prediction of a neural network for a specific input by backpropagating the contributions of all neurons in the network to each feature of the input (Dwivedi et al. 2023).

Layer-wise Relevance Propagation (LRP) is another backpropagation-based technique that scales to highly complex deep neural networks. It operates by propagating the prediction backward in a neural network using a set of purposely designed propagation rules. It assigns relevance scores to neurons or feature maps, indicating their contribution to the final prediction (Montavon et al. 2019).

Unlike LRP, deep Taylor decomposition (DTD) approximates the function locally around the input data point using a Taylor series expansion and assigns relevance scores to different input features. Montavon et al. (2017) present DTD to explain non-linear classification decisions of AI models. Hassan et al. (2023) apply DTD to explain COVID-19 diagnosis with a deep network model. Koh and Liang (2020) introduce influence functions to understand black-box predictions made by AI models. While not directly related to DTD, it provides insights into techniques for explaining and interpreting the decisions of complex models.

Rule extraction aims to extract human-interpretable rules from complex models, providing insights into how the model makes predictions and enhancing transparency and interpretability. Qiao et al. (2021) propose a paradigm for learning independent logical rules in disjunctive normal form as interpretable models for classification. Vaughan et al. (2018) propose an explainable neural network (xNN), designed to learn interpretable features with direct extraction and display capabilities. Wu et al. (2018) introduce a novel tree-regularisation technique that allows domain experts to quickly understand and approximately compute the complexity of a model. Several model simplification techniques have been proposed to improve neural networks' explainability. The DeepRED algorithm (Zilke et al. 2016) extends a rule extraction approach presented by Sato and Tsukimoto (2001) for multi-layer neural networks by adding more decision trees and rules. Martens et al. (2007) extract rules from trained SVMs (support vector machines) with minimal loss of accuracy, ranking it highly among comprehensible classification techniques. Dumitrescu et al. (2022) propose a penalised logistic tree regression (PLTR) model, extracting rules from various short-depth decision trees built with original predictive variables as predictors in a penalised logistic regression model. PLTR captures nonlinear effects in credit scoring data while preserving the intrinsic interpretability of the logistic regression model.

Overall, inference mechanism-based XAI techniques offer interpretable explanations using specific inference mechanisms in models, such as leveraging gradients computed during the training process. By understanding how each input feature contributes to the model's decision, users can gain insights into the model's behaviours, improving transparency and trust in AI systems.

To gain a holistic view of existing XAI techniques, Table 3 provides a summary of all XAI techniques according to our research.

TABLE 3 | Summary of XAI methods.

MC		XAI technique	AS	IS	MD	Reference
Samples-based		BMB/G/S-LIME	Po	Local	MA	Hung and Lee (2024)
		Counterfactual explanations (CE)	Po	Local	Both	Guidotti (2022)
		Individual conditional expectation curves	Po	Local	MS	Goldstein et al. (2015)
		Local interpretable model-agnostic explanations (LIME)	Po	Local	MA	Ribeiro et al. (2016a)
		Scoped rules	Po	Local	Both	van der Waa et al. (2021)
		SHapley Additive exPlanations (SHAP)	Po	Both	MA	Lundberg and Lee (2017)
Features-based		Accumulated local	Po	Local	MA	Biecek (2018)
		Archipelago	Po	Local	MA	Tsang et al. (2020)
		Feature interaction	Po	Global	MA	Tsang et al. (2020)
		Local interpretable MA explanations (LIME)	Po	Local	MA	Ribeiro et al. (2016a)
		Partial dependence plot (PDP)	Po	Global	MA	Greenwell et al. (2017)
		Permutation feature importance (PFI)	Po	Global	MA	Altmann et al. (2010)
		Sensitivity analysis	Po	Both	MA	Sobol' (1990)
		SHapley Additive exPlanations (SHAP)	Po	Both	MA	Lundberg and Lee (2017)
Modelling method-based	Transparent	Attention branch network (ABN)	Pr	Both	MA	Fukui et al. (2019)
		Attention mechanism	Po	Local	MS	Wiegrefe and Pinter (2019)
		Deep neural decision forests (NDFs)	Pr	Both	MA	Kontschieder et al. (2015)
		General additive models (GAMs)	Pr	Both	MA	Harezlak et al. (2018)
		Linear/logistic regression	Pr	Both	MS	Ribeiro et al. (2016b)
		Neural additive models (NAMs)	Pr	Both	MA	Agarwal et al. (2021)
	Inference mechanism	Tree based: decision tree	Pr	Both	MS	Quinlan (1986)
		Backpropagation based: deep LIFT rescale	Po	Local	MA	Shrikumar et al. (2017)
		Backpropagation based: deep shap	Po	Local	MA	Chen et al. (2019)
		BackpropagationBased: DeepTaylor	Po	Local	MA	Montavon et al. (2017)
		DeepRED	Pr	Both	MS	Zilke et al. (2016)
		Explainable neural network (xNN)	Pr	Both	MS	Vaughan et al. (2018)
		Gradient based: Grad-CAM	Po	Local	MA	Selvaraju et al. (2017)
		Gradient based: integrated gradients	Po	Local	MA	Sundararajan et al. (2017)
		Gradient based: smooth gradients	Po	Local	MA	Smilkov et al. (2017)
		Gradient based: vanilla gradient	Po	Local	MA	Yuan et al. (2022)
		Layer-wise relevance propagation (LRP)	Po	Local	MA	Montavon et al. (2019)
		Penalised logistic tree regression (PLTR)	Pr	Both	MS	Dumitrescu et al. (2022)
		Rule extraction	Po	Local	MA	Averkin and Yarushev (2021)

Abbreviations: AS: application stage; IS: interpretability scope; MA: model-agnostic; MC: model component; MD: model dependency; MS: model specific; Po: post hoc; Pr: pre hoc.

2.3.3 | Comments on Modelling Method-Based XAI Techniques

Based on the discussion of modelling method-based XAI techniques, the transparent structure is easily to be understood, which means there is no need for more explanations for the outputs of transparent structure. For example, the decision path and weights of decision trees and linear regression models can be clearly presented to users. However, there are still some points that need to be explored.

Data and task complexity limitations: transparent structure-based XAI usually have less dependence on data and can still perform well when the amount of data is insufficient or the data quality is not high. However, with the increasing of data-scale, including but not limited data sets' dimensions, data types and sample size, It is inevitable that XAI methods should have the ability to capture complex patterns and nonlinear relationships for large-scale data and complex tasks with well performed explanations.

Hence, the discussions on how to optimise transparent structure-based XAI methods should be developed to improve the ability to handle large-scale and complex tasks while maintaining interpretability.

While for inference mechanism-based XAI techniques, they are not easy to be understood by consumers directly. But, these inference mechanism-based XAI perform better at large data sets. Therefore, a natural idea is to explore ensemble models that combine transparent and black-box models to achieve high-performance while maintaining a certain degree of interpretability.

Model complexity limitations: in some cases, transparent models may be too simple and cannot provide sufficient prediction accuracy and generalisation ability. Thus, a research direction is to explore how to enhance a model's ability to handle complex data while maintaining model transparency, such as developing more complex but still interpretable transparent models. Noticed that, inference mechanism-based XAI techniques are usually used into deep learning models, and having higher performance and flexibility than transparent structure-based XAI techniques and can handle complex nonlinear relationships and large-scale data. They are also applicable to complex tasks such as image recognition and natural language processing, with strong versatility. Thus ensemble of black-box and transparent structure-based XAI may be an opportunity of XAI to balance the interpretability and performance/efficiency.

2.4 | Evaluation Methods

As XAI is developing speedily and widely applied in industries, more and more XAI techniques are proposed to offer understanding predictions and decisions to AI models' users. However, model users may still not trust the results of AI models if we just have explained the AI models, without providing the quality of the explanation.

To overcome this barrier, properly using XAI to enhance the systematic evaluation of AI models, organisations can ensure that their AI systems not only have high-performance but also are transparent, fair and trustworthy.

Evaluation of XAI can help build trust from consumers, comply with regulatory requirements, identify and mitigate bias, and make more informed and ethical decisions. As AI continues to be integrated into all aspects of B&E, the importance of XAI assessments will grow, driving responsible and effective use of AI technologies.

Thus, this section will therefore review evaluation metrics and make comments on them.

2.4.1 | Evaluation Metrics

Due to the diversity of data (NLP, time series, panel data, etc.) and XAI techniques (samples-based, features-based, modelling method-based, etc.), it is difficult to evaluate and compare the developed XAI techniques. There remains a notable absence of robust and trustworthy evaluations regarding the impact of explanations on users' experiences and behaviours (van der Waa et al. 2021). The literature by Anjomshoe et al. (2019) revealed that 97% of the 62 reviewed articles acknowledged the importance of explanations in meeting user needs. However, a significant proportion (41%) of these articles did not evaluate their explanations with actual users. Furthermore, among the papers that have performed user evaluations, a relatively small percentage provided comprehensive discussions on the context (27%), results (19%) and limitations (14%) of their experiments. Another survey conducted by Adadi and Berrada (2018), which evaluated 381 papers, found that only 5% papers explicitly focused on evaluating XAI techniques. These findings indicate that while evaluations of XAI are being researched, many of them lack sufficient detail to serve as robust foundations for further research in the field of XAI.

Generally, when we comprehend an XAI model and see it as an optimisation problem, there are two part in the objective function Λ :

$$\Lambda = h(f(\mathbf{X}), \mathbf{y}) + I(f(\mathbf{X}), \mathbf{y}, \hat{\mathbf{y}}) \quad (2)$$

where $h(f(\mathbf{X}), \mathbf{y})$ is the original prediction/classification performance, and $I(f(\mathbf{X}), \mathbf{y}, \hat{\mathbf{y}})$ is the XAI models' interpretable performance.

The classification performance metrics include 'Accuracy', 'Precision', 'Recall', 'F1 Score', 'AUC' area under the ROC curve (receiver operating curve), and the prediction performance include 'Mean Squared Error (MSE)', 'Root Mean Squared Error (RMSE)', and 'Mean Absolute Error (MAE)' (Novaković et al. 2017).

It is necessary to turn our attention to the quality of interpretability, which is the evaluation of XAI techniques. As introduced in Table 2, there are several goals of XAI, intuitively, the

corresponding evaluation methods should also align with the respective goals.

However, related research has not been well established. Only a few scholarly papers discuss the evaluation of XAI. van der Waa et al. (2021) discuss and propose some methods for evaluating/ accessing the performance of XAI. Lozano-Murcia et al. (2023) compare different kinds of evaluation methods on several datasets, and gives corresponding evaluation methods for feature importance, consistency, stability and robust, computation time and efficiency, fairness and bias and regulatory compliance. Agarwal et al. (2022) propose a comprehensive and extensible open-source framework for evaluating and benchmarking post hoc XAI methods, without considering other types of XAI methods.

Below briefly summarises existing universal XAI evaluation methods.

Stability and Robustness check the stability of the explainability and the overall model across different training samples or data perturbations. Robust models should produce consistent results and consistent interpretability. Then, the relationship of the interpretability technique among the different scenarios is computed (Lozano-Murcia et al. 2023):

$$SR(E) = \frac{1}{n(n-1)/2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{corr}(E(S_i), E(S_j))$$

where E is a specific XAI method, $E(S_i)$ and $E(S_j)$ are the results of applying E into two different training samples or data perturbations, S_i and S_j represent two data samples after perturbations, and $\text{corr}(E(S_i), E(S_j))$ is the Pearson correlation coefficient between $E(S_i)$ and $E(S_j)$.

Fairness and Bias evaluate AI models in terms of their potential biases or the unfair treatments, especially regarding some sensitive attributes (e.g.: age, gender and race). Thus, the core of fairness and bias is to evaluate whether the explanations provided satisfy these criteria for each sensitive feature (Angerschmid et al. 2022):

$$FB(E) = \sum_{i=1}^n \omega_i F_i$$

where ω_i is the weight assigned to the i th fairness criterion F_i based on its relevance or importance.

Consistency is proposed based on the relationship between the XAI techniques results after its applications to different algorithms. This metric ensures that the explanations should be consistent across similar inputs (Lozano-Murcia et al. 2023).

$$CO(E) = \frac{1}{m(m-1)/2} \sum_{i=1}^{m-1} \sum_{j=i+1}^m \text{corr}(E(A_i), E(A_j))$$

where $C_{score}(E)$ is the score of the XAI technique E , m is the number of applied AI models, and $E(A_i)$ and $E(A_j)$ are the results/ predictions of applying E to AI model A_i and A_j , respectively.

The sums iterate through all distinct pairs of diverse AI models, and the entire equation computes the average over the count of these pairs.

Computation time and efficiency is vital as some XAI techniques can be highly complex. The trade-off between computation time and the quality of the explanations provided has to be proposed (Dwivedi et al. 2023).

$$CE = \omega_1 T + \omega_2 (1 - F_{score})$$

where T denotes the computation time, F_{score} represents the quality of the explanation, and the weights ω_1 and ω_2 can be adjusted based on special conditions. This equation assumes that a lower CE score is better (since lower computation time and higher F_{score} are desired).

Fidelity refers to how accurately the explanation reflects the true behaviour of the model being explained. It measures the accuracy or truthfulness of the explanation provided by XAI techniques relative to the original model (Miró-Nicolau et al. 2025). High fidelity means that the explanation closely agrees with the prediction or behaviour of the original model, while low fidelity implies that the explanation may be misleading or inaccurate.

However, there is no widely accepted definition of fidelity. Fidelity can be calculated based on the consistency of feature importance, or based on prediction accuracy, or even by explaining consistency.

For example, fidelity can be provided by comparing the prediction accuracy of the XAI model and the original model, which is generally being used for sample-based XAI techniques (like LIME) and global interpretability techniques (such as decision tree models) (Miró-Nicolau et al. 2025) gives the following definition:

$$FI = \frac{n}{n+1} \sum_{i=1}^n |f(X_i) - \hat{f}(X_i)|$$

Except these general evaluation methods, there are some evaluation metrics for specific XAI techniques. *Counterfactual Similarity for CE* refers to the similarity of the counterfactual examples provided by the model to the original instance given the prediction of a sample in the CE method (Wachter et al. 2017).

$$CS = \frac{1}{1 + DS(\mathbf{X}, \mathbf{X}')}$$

where DS represents the distance between original data $\mathbf{X}$ and counterfactual examples $\mathbf{X}'$, which is usually Euclidean distance.

Feature Contribution (FC) for features-based techniques measures the impact of each feature on the model's prediction in features-based XAI methods, such as SHAP and LIME (Lundberg 2017).

$$FC = \sum_{S \in \mathcal{N} \setminus \{i\}} \frac{|S|! (N - |S| - 1)!}{|N|!} |f(S \cup \{i\}) - f(S)|$$

where S is the Shapley value of the i th feature, S is the set that contains all subsets of features, N is the set of all features, and $f(S)$ is the prediction of AI model based on the set of S .

Relevance Score for LRP calculates the relevance score of each layer to represent the contribution of each input feature to the model prediction (Bach et al. 2015). For a neural network, the formula for calculating the contribution of each neuron is:

$$RS_j^l = \sum_i \frac{z_{ij}^l}{\sum_k z_{ik}^l} RS_i^l$$

where RS_j^l is the relevance score of j th neuron in layer l , z_{ij}^l is the weight of i th neuron to j th neuron in layer l , and RS_i^l is the relevance from the last neuron.

More metrics about evaluation methods of XAI have been discussed by Agarwal et al. (2022), which provides a systematically comprehensive and extensible open source framework to evaluate and benchmark post hoc XAI methods, with 22 quantitative metrics for evaluating faithfulness, stability, robustness, and fairness of explanation methods. Mohseni et al. (2021) present a survey and framework to support diverse evaluation methods in XAI research, and develop a framework with step-by-step design guidelines paired with evaluation methods. Pawlicki et al. (2024) execute a comprehensive systematic review of XAI methods, various evaluation metrics, and existing frameworks to assess their utility and relevance, delivering key insights into their practical utility and effectiveness.

In general, with the advancement of AI technology, evaluation methods are still under development. From the current research progress, they are still lagging the development of XAI technology, which presents a huge, yet challenging research opportunity.

2.4.2 | Comments on Evaluation Methods of XAI

Through these XAI methods, we can gain explanations of AI models. However, there is no well-agreed understanding of what exactly XAI is or what an optimal explanation should be (Samek et al. 2021; Islam et al. 2024).

Although many XAI techniques have been proposed, there is no clear definition of XAI and no unanimous measures to evaluate different XAI techniques. As discussed in Section 2, attention-based methods are widely used to explain black-box deep learning models. However, can attention really explain these models effectively? Serrano and Smith (2019) scrutinised this question. Without a robust evaluation system in XAI, it is risky to fully trust an AI model, even if some predictions are explained. Moreover, the definition of XAI and the development of evaluation methods in XAI lag behind the development of AI technology, indicating a promising research area full of challenges.

Objective evaluation of explanation quality: There is a lack of standardised indicators to evaluate the quality and effectiveness

of explanations. Most current evaluation methods rely on subjective judgement, making it very challenging in comparing different explanation methods. Different fields have their own requirements for explanations, and interpretability is a relative concept, so it is difficult to universalize the existing evaluation standards.

Developing objective and standardised evaluation indicators and benchmarks to measure the quality of explanations is necessary. This includes quantifying the accuracy, stability and operability of explanations. Another research opportunity lies in studying the different requirements for explanations in various fields and developing field-specific evaluation standards and frameworks.

User-centred evaluation methods: Current evaluation methods are mostly based on technical indicators and lack full consideration of the needs and feedback of end users (such as business personnel, doctors, engineers, etc.). Nguyen et al. (2024) evaluate the XAI methods using plausibility and faithfulness metrics to measure how well the explanations align with human intuition without considering the applicability to other domains. Thus, future work should investigate the generalizability of the evaluation approach across different fields and evaluate its performance and interpretability on various datasets.

Users' understanding and needs are vital, and existing methods struggle to fully evaluate the user experience of explanations. Thus, developing user-centred evaluation methods, combining user research and experiments, presents challenges and opportunities in XAI evaluation methods to understand users' needs and preferences for explanations. Designing user experience tests and surveys to evaluate the effectiveness and ease of use of explanation methods in real-world scenarios can address the problem of the lack of user consideration.

Explanation stability and consistency: The stability and consistency of explanation methods have not been well evaluated, as discussed in sample-based and feature-based XAI techniques. The main reason for the lack of stability and consistency is the absence of systematic methods to quantify and compare these aspects of explanations. Furthermore, the consistency between global and local explanation is also promising, for example, XAI techniques like SHAP, LIME can both provide global and local interpretability, how to gain the stable and consistent explanation for both global and local interpretability is also a challenging topic in XAI for B&E.

To improve the stability and consistency of XAI techniques, we should develop evaluation metrics and methods that can quantify these qualities. More focus should be placed on enhancing the stability and consistency of explanation methods to ensure that the explanation results are reliable and credible.

Overall, there are still significant knowledge gaps and challenges in the evaluation of XAI. Future research needs to make breakthroughs in standardised evaluation indicators, user-centred evaluation, stability and consistency, model complexity adaptability, and cross-domain applicability. Through systematic research and innovation, a more comprehensive and effective evaluation method can be developed to promote the practical

application of explainable artificial intelligence, improving the transparency, credibility and user acceptance of AI systems.

3 | Applications of XAI in B&E

Section 1 provides examples of AI applications in B&E. This section will discuss real-world applications of XAI techniques in B&E in detail, bridging the gap between the methodological development of XAI techniques and practical use.

3.1 | Finance

Using AI can improve the accuracy and efficiency of financial products in the most popular tasks like credit management, stock price predictions, and fraud detection (Černevičienė and Kabašinskas 2024; Weber et al. 2024). It can also improve market forecasting, ensuring credit scoring fairness and trustworthiness, identifying factors associated with fraud detection, and reducing potential costs caused by AI biases or errors (Urazova 2023).

For instance, if financial institutions provide inconsistent and unexplainable decisions in similar situations, they risk losing customer trust, potentially leading to a crisis of confidence and significant financial loss. Hashemi and Fathi (2020) apply CE to credit scorecards and financial text classification problems, offering sample-based explanation methods for black-box financial AI models. Szeląg and Słowiński (2024) develops monotonic decision rules to understand bank data and characterise loyal customers. By providing sample-specific explanations, SR enables users to make informed decisions and take appropriate actions based on the model's predictions.

Credit card fraud aims to cheat the users of the credit cards, Adhegaonkar et al. (2024) classify legitimate and fraudulent business transactions with three XAI methods: decision tree, logistic regression and support machine. Meanwhile, Chen et al. (2024) applies sample and features-based XAI techniques (SHAP and LIME) to credit scoring using 2016–2020 UK residential mortgage data to address the class imbalance classification problem. The results show that interpretations from LIME and SHAP become less stable as class imbalance increases, indicating that class imbalance adversely affects AI model interpretability. Due to the less of accuracy and transparency in traditional financial distress prediction, Fan et al. (2023) proposes a comprehensive framework that enhances the performance of both prediction and interpretability. With the applications of sample and feature-based XAI: PDP, ICE and SHAP, this work provides global and local interpretations that help businesses, financial institutions, and regulators make informed decisions.

3.2 | Marketing

AI is increasingly vital in marketing, offering opportunities to maximise ROI through business insights and making marketing more intelligent, efficient, consumer-friendly, and effective (Herhausen et al. 2024). However, marketers need to trust AI's recommendations, making XAI essential in marketing problems.

Marketers can use AI model explanations to gain deeper consumer understanding, providing the best possible experience and increasing ROI by thoroughly examining consumer data and understanding what consumers truly want (Haleem et al. 2022). Chen et al. (2023) propose an interpretable feature-based XAI method to predict hotel booking cancellations, influencing hoteliers' managerial decisions. Wang et al. (2023) applies SHAP for feature-based XAI to explore leverage points in customer churn prediction, providing more explainable insights into customer behaviours.

For click-through rate (CTR) prediction, which is crucial for advertising agencies to make appropriate recommendations and maximise profit, Jose and Shetty (2022) use a transparent structure-based additive neural network to provide interpretable insights into features and their interactions, demonstrating effectiveness and efficiency with reduced computational costs. With the development of web marketing, online coupon distribution has become a significant marketing measure that leads to increased sales, Yoneda et al. (2024) proposes an experimental design ML model to analyze potential purchase intention, and applies the feature-based XAI (SHAP) to estimate the effect of coupons and analyze the causal relationship between coupons and results. For the online advertising, applying XAI into prediction of user behaviour could help to understand the drivers behind user actions. Al-Khafaji and Karan (2023) recognise the opaque nature of AI models, then leverages sample and feature-based SHAP and LIME, tools of explainable AI, ensuring that AI models' decisions remain interpretable.

Chien et al. (2022) address fake news detection on social media, applying inference mechanism-based LRP to explain predictions and increase transparency in human-AI interaction.

3.3 | Insurance

AI revenues in insurance are expected to grow by 23 Bora et al. (2022) uses SHAP and LIME to provide sample and feature-based interpretable predictions for health insurance costs, enhancing user experience and building trust between users and AI models. Tzougas and Kutzkov (2023) utilise LIME to explain a binary classification AI model for predicting claims in a French motor third-party insurance portfolio. Yankol-Schalck (2022) constructs a score for personal automobile policies that evolves over the life of a claim, using LIME to interpret AI model results and improve fraud detection confidence. Gramegna and Giudici (2020) also employ LIME to enhance interpretability for AI models and explain consumer decisions regarding non-life insurance. Ramachandran et al. (2023) use random forests to provide feature-based importance scores for medical insurance cost prediction, aiding in understanding underlying relationships and identifying key factors driving outcomes.

3.4 | Supply Chain

As supply chains become more complex and globalised, AI plays a critical role in demand forecasting, supplier selection, route optimization, and inventory management (Pournader et al. 2021). However, without XAI, users may struggle to understand why

specific routes are chosen or management decisions are made, considering factors such as traffic patterns, delivery time windows and vehicle capacity constraints, highlighting the importance of XAI in the supply chain industry.

Sadeghi et al. (2024) present a feature-based explainable method (SHAP) to explore themes in tweets discussing XAI in decision support systems, providing empirical evidence of XAI's impact on supply chain decision-making processes. Bhatia and Albarrak (2023) present an XAI-based Faster RCNN model to evaluate food item contents, aiding policy-makers, manufacturers and merchants in efficient decision-making and improving public health and welfare. Olan et al. (2024) emphasise XAI's ability to measure uncertainty and predict users' information requirements, showing both prediction capability and logical reasoning in planning and object manipulation. Kumar and Kumar (2023) apply SHAP for sample and feature-based explanations in supplier selection under procurement, identifying biases, drifts, and data gaps, increasing consumer trust by making information more transparent and understandable.

Chang et al. (2024) apply features-based XAI to identify two effective algorithms (Random Forest and Gradient Boosting models), for credit risk detection to avoid the lack of interpretability or transparency makes decision-makers sceptical. This study also contributes to the literature on explainable credit risk detection in supply chain finance and provides practical implications for the decision-making of financial institutions.

3.5 | Human Resource

AI is increasingly adopted in human resources (HR) due to its potential to create value for consumers, employees, and organisations (Silva et al. 2022). However, recent studies show that organisations are not yet experiencing the expected benefits from AI adoption, as using AI directly in HR management is risky (Delecraz et al. 2022). XAI can be used in HR management for recruitment, performance evaluation, training, and development, as well as avoiding or mitigating bias, unfairness, and distrust in AI decisions.

According to Hofeditz et al. (2022), final hiring decisions are likely to remain with humans, but human biases could cause discrimination based on age, sex, race, and so forth. They develop a feature-based XAI approach to moderate these discriminations and explore the impact of AI-based candidate recommendations on candidate selection decisions. A 2022 UNESCO publication on 'The Impact of Artificial Intelligence on Women's Working Lives' reports that AI in recruitment processes excludes women from promotions (Collett et al. 2022). The report finds that setting the user's gender to 'female' resulted in fewer ads related to higher-paying jobs compared to users setting the gender to 'male'. This persistent bias impacts the application of AI and threatens critical workforce factors like diversity, equity and inclusion, which is a main goal of XAI.

The technological capabilities of Human Resource Analytics (HRA), enhanced by recent innovations in AI, offer exciting opportunities. However, organisations often fail to realise this

potential due to limited understanding of why individuals choose to adopt or ignore the corresponding tools. Thus, Hülter et al. (2024) find that fairness and non-discrimination were not critically questioned, even when potential biases were highlighted by XAI visualisations and the interviewees were explicitly asked about them.

Employee attrition and high turnover are significant challenges in today's competitive job market. Marín Díaz et al. (2023) explore the application of XAI in identifying potential employee turnover and devising data-driven solutions. Their work discusses both sample-based (FI and LIME) and feature-based (SHAP and PDP) XAI techniques to explain employee turnover, aiding decision-makers in understanding model predictions and developing targeted retention and recruitment strategies.

3.6 | Healthcare

AI has been extensively implemented in healthcare (Javaid et al. 2022). However, doctors are often unable to explain why certain decisions are made, limiting AI's applicability in healthcare. With XAI, doctors can explain why certain patients are at high risk for hospital admission and determine the most suitable treatments.

Ge et al. (2023) suggest that AI may over-promise real-world performance due to inflationary effects and uses Parkinson's disease as evidence to propose an improved evaluation for AI models in healthcare, an XAI attempt to enhance AI model explanations. Agbozo and Balungu (2024) note that recent advancements in black-box AI models lack understandable explanations, limiting fairness, confidence and transparency in AI decisions. They address the need for XAI to explain AI's decisions in biomedicine, such as AI-assisted clinical diagnoses, using the Shapley value to illustrate predictions from a liver disease detection model.

Peng et al. (2021) use SHAP, LIME and PDP to improve model interpretation of liver disease, combining sample and feature-based interpretable methods to enhance transparency and gain insights into complex models' judgements, guiding treatment strategies and improving hepatitis patient prognosis. Janssens et al. (2024) focus on social media rumour detection to tackle societal impacts from potential misinformation, explaining models with LIME and assessing explanation quality via fidelity and stability. In healthcare, clinicians find it difficult to understand and trust complex AI models due to a lack of intuitive explanations. ElShawi et al. (2021) apply LIME to provide insights into prediction processes, explaining how results were generated from different types of real-world healthcare data.

Chen et al. (2019) highlight the significant impact of AI model predictions on patient welfare. They present the inference mechanism-based DeepSHAP for complex AI models, a framework for layer-wise propagation of Shapley values that builds upon DeepLIFT to make complex healthcare models explainable.

For the modelling method-based XAI, Shi et al. (2020) propose an Explainable Attention-based Model (EXAM) for COVID-19

automatic diagnosis with convincing visual interpretations, where channel-wise and spatial-wise attention mechanisms are combined to effectively extract key features and suppress irrelevant information about COVID-19.

3.7 | Macroeconomics

AI is beneficial in modelling macroeconomics across multiple areas. However, XAI in macroeconomics is crucial for maintaining trustworthiness, ensuring regulatory compliance, improving model accuracy, enhancing decision quality, and mitigating potential negative outcomes. By elucidating the black-box nature of AI models, XAI supports more responsible and effective deployment of AI in shaping economic policies.

Chapman and Desai (2023) analyze comprehensive payments data for macroeconomic predictions in Canada, deriving feature-based XAI to explain predictions and assess AI models' predictive value. Yue and Au (2023) introduce a financial forecasting method using AI models, applying SHAP and Shapley values for sample and feature-based interpretability, converting complex ML predictions into human-understandable forecasting reports, and offering valuable insights for investors, traders, and analysts in a fast-moving economic environment. Ghosh and Jana (2024) adopt AI models to investigate clean energy investment predictability in the US market, using PDP to explore explanatory variable contribution patterns and obtain their relative importance, a widely applied sample-based XAI method. Park and Yang (2022) provide two methods for better economic prediction and decision-making, using feature-based SHAP to explain country-specific economic growth and crisis patterns.

3.8 | Microeconomics

Microeconomics researchers have already adopted AI across various industries, revolutionising consumer economic behaviour analysis, market structure simulation and competitive agent strategy studies. Hakami (2023) reviews existing research on AI in microeconomics, highlighting the multidimensional nature of AI integration and prompting reflections on ethical, societal and economic dimensions.

Zhang et al. (2021a) and Zhang et al. (2021b) develop AI approaches for microeconomic modelling, showing that AI can effectively characterise underlying nonlinear relationships and significantly improve fitting and prediction. Brathwaite et al. (2017) provide a microeconomic framework for decision trees (a transparent structure-based XAI method), exploring how decision trees represent a non-compensatory decision protocol. Sachan et al. (2020) propose an XAI system to automate the loan underwriting process using a belief-rule-base (BRB) system, a sample-based scoped rule system. Based on a business case study, the BRB system finds an optimal trade-off between accuracy and interpretability, two sub-objectives of XAI.

While existing research illustrates the importance and necessity of AI for microeconomics, there is less attention on XAI in this field, and related research still faces significant challenges.

3.9 | Comparative Applications and Trends

Previous sections show that the landscape of XAI applications continues to expand as AI techniques evolve, and the trends driving this evolution often depend on the comparative strengths and weaknesses of various models.

This section will investigate these trends, providing a comparative analysis of how XAI is applied in real-world B&E applications and identifying the factors shaping their adoption. As we can see from Figure 3, the reference number of B&E with XAI between 2018 and 2024 are increasing rapidly, which suggests researchers are increasingly aware of the importance of interpretability and XAI in B&E.

From the perspective of application fields, as shown in Table 4, various domains within B&E have begun leveraging XAI to achieve significant benefits. However, the adoption of XAI in B&E is not without any challenges. Key limitations include data privacy concerns in finance, data sparsity in insurance, and the lack of real-time data in supply chain management. These challenges present both obstacles and opportunities for researchers to address and innovate.

Similarly, Table 5 lists the most applied XAI methods in B&E, which shows the application of XAI in B&E is currently focused on relatively straightforward XAI techniques. More complex and advanced methodologies have yet to be extensively explored, indicating that there are significant opportunities for further innovation and development in this field.

3.10 | Application Tools

To help users apply XAI into practical problems, some tools are developed. However, as far as we know, there is no current a framework that fully integrates XAI technologies or behaves like Artificial General Intelligence (AGI), but there are some initiatives and frameworks that attempt to move in that direction.

While there is no single universal framework that integrates all XAI methods, some research has been made to bring multiple XAI tools together:

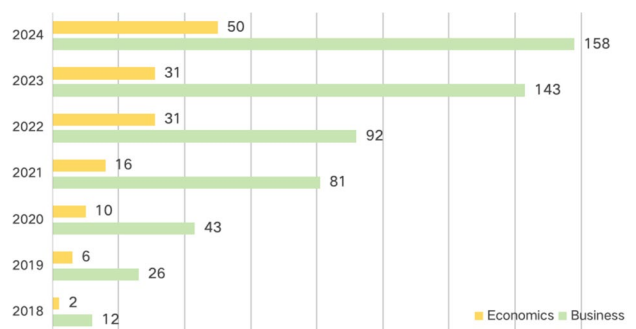


FIGURE 3 | Reference number of business and economics with XAI.

TABLE 4 | Summary of XAI's application tasks in B&E.

Field	Domain	Task	Benefit	Limitation	Reference
Business	Finance	Credit score prediction; loyal customer identification; credit fraud detection; financial distress prediction	Transparency in scoring and predictions; identifying key factors driving predictions	Data privacy concerns; complexity in interpreting financial data	Hashemi and Fathi (2020); Szelag and Slowinski (2024); Chen et al. (2024); Adhegaonkar et al. (2024); Fan et al. (2023)
	Marketing	CTR prediction; user behaviour prediction; Purchase intention prediction	Better targeting of advertisements; understanding user preferences	Dynamic and changing user behaviour; model overfitting risk	Haleem et al. (2022); Wang et al. (2023); Jose and Shetty (2022); Chien et al. (2022); Yoneda et al. (2024); Al-Khafaji and Karan (2023)
	Insurance	Insurance cost prediction; claims prediction; insurance score prediction	Accurate pricing of premiums; fraud detection	Data sparsity for rare events; regulatory constraints	Bora et al. (2022); Tzongas and Kutzkov (2023); Yankol-Schalck (2022); Gramegna and Giudici (2020); Ramachandran et al. (2023)
	Supply chain	Supply chain decision; capability prediction; supplier selection	Enhanced efficiency in logistics; better supplier performance	Limited interpretability of multi-stage decisions; lack of real-time data	Sadeghi et al. (2024); Bhatia and Albarrak (2023); Olan et al. (2024); Kumar and Kumar (2023); Chang et al. (2024)
	Human resource	Candidate selection; potential employee turnover prediction	Fair and transparent hiring; early identification of retention risks	Bias in training data; privacy concerns in employee data	Hofeditz et al. (2022); Marín Díaz et al. (2023); Hülter et al. (2024)
	Healthcare	Disease detection; automatic diagnosis; treatment recommendation systems; drug discovery	Improved diagnostic accuracy; faster drug discovery	Trust issues among practitioners; challenges in real-world validation	Ge et al. (2023); Agbozo and Balungu (2024); Peng et al. (2021); Janssens et al. (2024); ElShawi et al. (2021); Chen et al. (2019); Shi et al. (2020)
Economics	Macroeconomics	Macroeconomic prediction; financial forecasting; investment prediction; economic prediction	Better policy decisions; improved investment strategies	High uncertainty in long-term predictions; sensitivity to exogenous shocks	Chapman and Desai (2023); Yue and Au (2023); Ghosh and Jana (2024); Park and Yang (2022)
	Microeconomics	Supply and demand forecasting; labour market analysis; risk assessment; production function and efficiency analysis	Improved resource allocation; transparency in economic modelling	Complexity in capturing all influencing factors; interpretability challenges in large datasets	Zhang et al. (2021); Brathwaite et al. (2017); Sachan et al. (2020)

Abbreviation: CTR: click-through rate.

TABLE 5 | Summary of XAI's application techniques in B&E.

XAI techniques	Method	Applicable scenario	Suitable data type	Performance metric
Samples-based	LIME	Explaining individual predictions in decision-making systems (Mizanur Rahman and Alam 2023; Chen et al. 2024; Tzougas and Kutzkov 2023)	Tabular, text, image	Fidelity, robustness, interpretability
	CE	Exploring alternative scenarios for better decisions (Hashemi and Fathi 2020)	Tabular, text	Classification performance, counterfactual similarity
	ICE	Understanding how specific features influence predictions (Fan et al. 2023)	Tabular	Fidelity, interpretability
Features-based	SHAP	Identifying key drivers/factors behind model predictions (Kumar and Kumar 2023; Sadeghi et al. 2024; Bora et al. 2022)	Tabular, text	Feature importance, fidelity
	PDP	Exploring average effects of features in AI decision models (Fan et al. 2023)	Tabular	Feature contribution, error metrics (e.g., RMSE)
	PFI	Ranking important factors influencing outcomes (Marín Marín Díaz et al. 2023)	Tabular, image	Feature contribution, fidelity
Modelling method-based	GAMs	(Jose and Shetty 2022)	Tabular	Predictive performance, interpretability
	LRP	Explaining neural network predictions in complex domains (Chien et al. 2022)	Image, text	Relevance score, classification performance
	TBMs	Building interpretable decision rules for ranking or selection (Chang et al. 2024)	Tabular	Predictive performance, feature importance

Abbreviations: CE: counterfactual explanations; GAMs: general additive models; ICE: individual conditional expectation; LIME: local interpretable model-agnostic explanations; LRP: layer-wise relevance propagation; PDP: partial dependence plot; PFI: permutation feature importance; SHAP: shapley additive explanations; TBMs: tree based models.

- *IBM AI Explainability 360 Toolkit*: This is an open-source library that provides a collection of algorithms for explaining machine learning models. It supports multiple XAI techniques, allowing users to select methods appropriate for their models (Arya et al. 2021).
- *Aequitas*: This is an open-source bias and fairness auditing toolkit that helps assess whether AI models treat demographic groups fairly. It focuses on detecting discrimination in AI models based on sensitive attributes such as race, gender, or age (Saleiro et al. 2018).
- *H2O.ai's Driverless AI with XAI*: This platform includes built-in explainability tools such as partial dependence plots (PDP), feature importance, and SHAP values to help interpret AI models automatically generated by the platform (Hall et al. 2017).
- *Microsoft InterpretML*: This is another open-source tool designed to help users understand machine learning models. It combines different XAI methods, such as SHAP and LIME, and includes both model-agnostic and interpretable models (Nori et al. 2019).
- *SHAP (SHapley Additive exPlanations)*: It uses cooperative game theory to explain the contribution of each feature to a model's predictions. It provides both global explanations (how features influence the model overall) and local explanations (for individual predictions) (Lundberg 2017).
- *LIME (Local Interpretable Model-Agnostic Explanations)*: It provides local explanations for any model's predictions by perturbing the input data and analyzing how changes affect the model's output. It can explain predictions from black-box models like neural networks or gradient-boosted trees (Ribeiro et al. 2016a).
- *DeepLIFT (Deep Learning Important FeaTures)*: It is an interpretability method specifically designed for deep learning models. It identifies which neurons or input features are most influential in producing a given output, making it especially useful for neural networks (Shrikumar et al. 2017).

- *OmniXAI: A Library for Explainable AI*: It is a Python machine learning library for XAI that provides a full range of capabilities to address the pain points of explaining machine learning model decisions in practice. OmniXAI aims to be a one-stop comprehensive library that makes it easy for data scientists, machine learning researchers, and practitioners to implement XAI, who need to explain various types of data, models, and explanation methods at different of the AI process (Yang et al. 2022).
- *OpenXAI*: It is the first general-purpose lightweight library that provides a comprehensive list of functions for systematically evaluating the quality of explanations generated by attribute-based explanation methods. OpenXAI supports the development of new datasets (both synthetic and real-world) and explanation methods, and is committed to promoting systematic, reproducible, and transparent evaluation of explanation methods. (Agarwal et al. 2022)

While choosing the right XAI tool for B&E, there are several factors that need to be noticed carefully:

- *AI model type*: Some tools are better suited for specific models (e.g., DeepLIFT for neural networks, SHAP for any model)
- *Business context*: Industries such as finance, healthcare or e-commerce may have specific regulatory requirements or user needs that some tools (e.g., Aequitas for fairness) can address.
- *Scalability and ease of use*: Enterprise-grade tools such as Fiddler and H2O.ai offer scalability and ease of integration, while open source tools such as LIME and SHAP offer flexibility but may require more technical expertise.

By carefully choosing and evaluating these tools based on B&E needs and the specific AI models used, users can effectively achieve explainability and transparency in AI applications for B&E, thereby improving trust, transparency, and fairness. However, These frameworks combine different XAI techniques, but none can fully act as a universal XAI integration platform or allow an AI to understand what users ask like an AGI. They still rely on human oversight and can't replicate general intelligence.

3.11 | User Guidance

Based on the previous discussion of XAI techniques and applications, this section gives some using guidance for the audiences of XAI methods in B&E.

There are several different target audiences for XAI methods, as can be seen from Table 2, so that we offer guidance for these target audiences to use or apply XAI methods better, as shown in Figure 4.

We then provide a guidance from a technological perspective, as illustrated in Figure 5.

The key distinction between building AI models and XAI models lies in the necessity of selecting and applying XAI methods

when working with black-box models. The criteria for choosing appropriate XAI methods are discussed in Section 2, while evaluation approaches are detailed in Section 2.4. Specifically, from an application perspective, if a specific task has already been identified, Table 4 can serve as a reference. Conversely, if the technique to be used is predetermined, Table 5 provides relevant guidance.

4 | Challenges in XAI on B&E

From previous discussions, it is evident that AI has been extensively employed into various aspects of B&E.

However, there are still challenges in the application of XAI. For example, Saeed and Omlin (2023) identify challenges and potential research directions of XAI from two aspects: (1) general challenges and research directions of XAI and (2) challenges and research directions of XAI based on machine learning life cycle's phases: design, development and deployment. Longo et al. (2024) highlights the ongoing challenges within XAI, emphasising the need for broader perspectives and collaborative efforts. However, these are all general discussions without considering the field of B&E specifically. Thus, the exploration of XAI in B&E remains relatively limited, presenting both challenges and opportunities that require further research in B&E.

In the following sections, we will delve deeply into the specific challenges and opportunities that XAI faces in the fields of B&E.

- *Uncertainty in Prediction*: There is increasing emphasis on AI models that consider uncertainty. Decision-making systems may encounter uncertainty from various sources, each offering a different perspective (Kochenderfer 2015). For example, aleatory uncertainty arises from the inherent randomness of predictions, while epistemic uncertainty stems from insufficient data. In general, incorporating uncertainty enhances a model's reliability by allowing it to recognise situations where it lacks the knowledge needed for accurate predictions. Therefore, developing methods to address uncertainty in AI prediction models is crucial, as it helps users understand the levels of uncertainty associated with different outcomes.
- *Integration with Decision-Making Processes*: In recent years, there has been an increasing need in XAI to build trust and understanding the reasoning behind AI making decisions (Tiwari 2023). Specifically, the following two aspects are the most valuable for the integration with decision-making processes:
 1. There has been an increasing exploration of methods for integrating Explainable AI (XAI) into decision-making processes across various areas in B&E. Additionally, there has been a growing number of publications on this research direction recently (Bertsimas and Kallus 2020).
 2. Furthermore, there is a focus on developing methods for processing and analyzing data to support real-time decision-making, particularly in dynamic and fast-paced environments.
- *Trust and Robustness in XAI*: It addresses the robustness and trustworthiness of explanations, ensuring that they

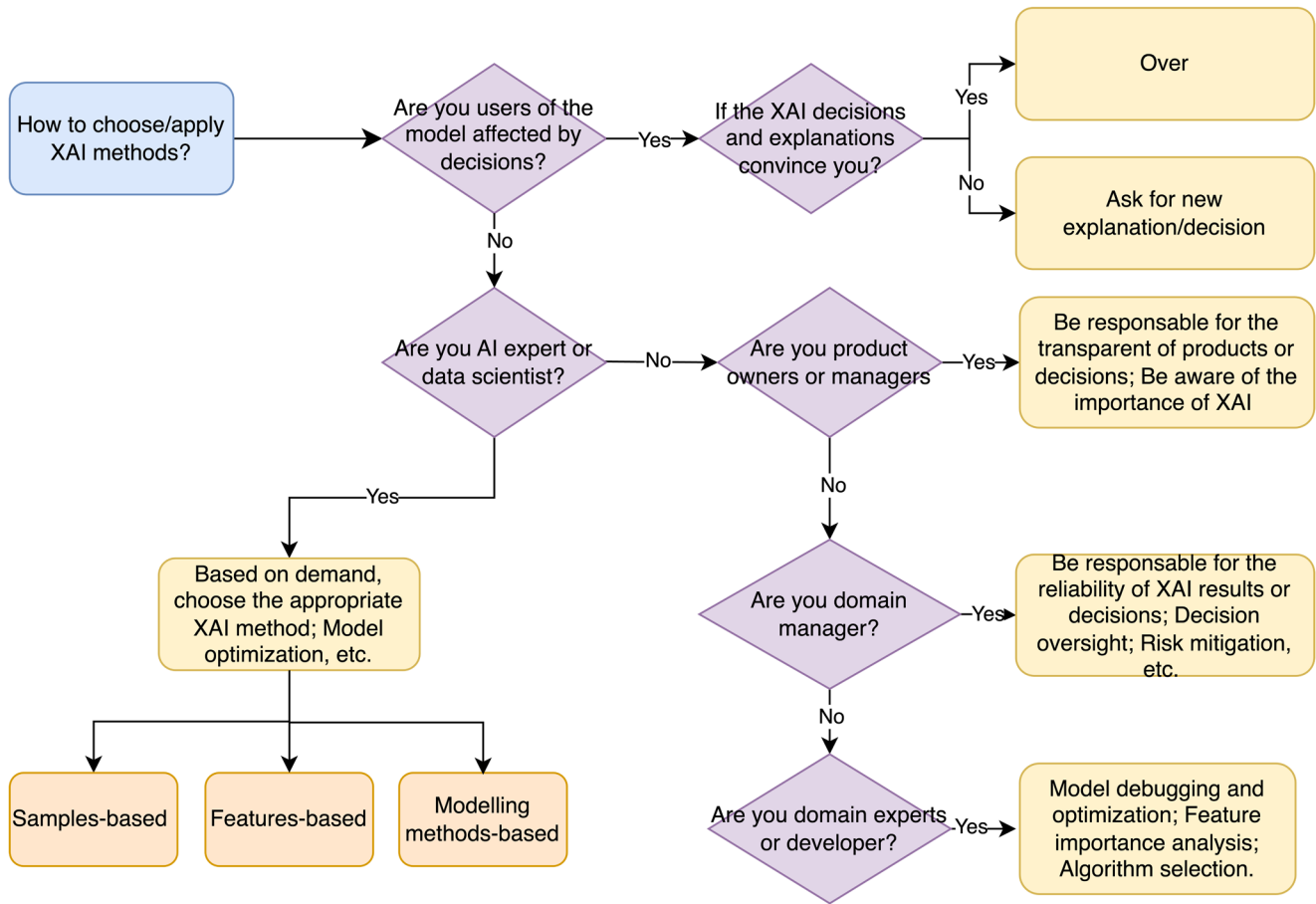


FIGURE 4 | Guidance on how to apply XAI methods by target audience.

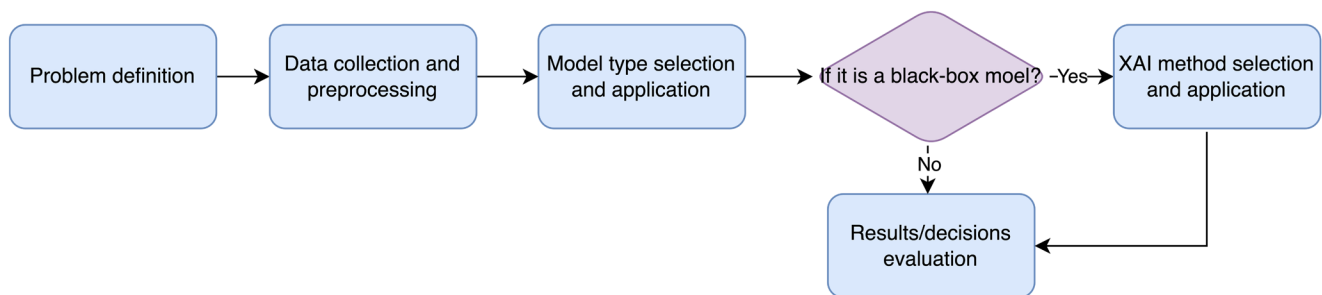


FIGURE 5 | Guidance on how to apply XAI methods by technological perspective.

accurately reflect an AI model's behaviours and are not sensitive to small changes in input data.

For AI to be widely accepted and used, it must not only be explainable but also reliable and trustable. XAI plays a significant role in building trust by ensuring that AI systems are transparent and understandable, while robustness ensures that these systems perform consistently and accurately, even in the face of challenges like noise, adversarial attacks or changing data environments.

In terms of B&E, trust and robustness are key considerations when deploying AI systems, because a lot of scenes are more sensitive and uncertain, such as banking and insurance (Bejger and Elster 2020). For these highly sensitive industries of B&E, it is difficult to be convinced that small/

micro changes make big different decisions, which means robustness is also very important in XAI (van der Cruysen et al. 2023).

To ensure long-term success, businesses need to continuously improve AI models based on users' feedback, keeping them robust and adaptable in dynamic environments, while ensuring they operate in an ethical and transparent manner. By aligning trust with robustness, businesses can create AI systems that are both effective and ethical, paving the way for more responsible AI use in high-risks areas like finance, healthcare, and so forth.

- **Ethical and Responsible AI:** Understanding the ethical implications of data analytics and AI, including issues related to bias, fairness, privacy, and accountability, is essential

for B&E (Dignum 2019). The EU AI Act,¹ which is one of the most comprehensive legislative frameworks for AI, proposed strict requirements around transparency and risk management, which shows clearly why regulatory bodies demand XAI, not merely as a best practice, but as a compliance measure for legal accountability.

Explainable AI (XAI) plays a critical role in addressing ethical imperatives, such as fairness, accountability, and transparency, by making AI systems more understandable and trustworthy (Kaur et al. 2022).

For example, when AI systems provide medical care, loan applications, or employment guidance, they should make the same recommendations to everyone with similar symptoms, financial circumstances, or professional qualifications.

To ensure trust, businesses must prioritise:

1. Fairness and bias mitigation by regularly auditing AI systems with XAI tools.
2. Transparency and accountability, using XAI to trace and explain AI decisions.
3. Ethical data practices, ensuring responsible data use and compliance with privacy regulations. Ultimately, by leveraging XAI to enhance ethical practices, businesses can build trust with stakeholders, regulators, and customers, ensuring that AI systems are both effective and aligned with ethical standards.

Hence, the research of ethical and responsible AI presents both challenges and opportunities in the fields of B&E.

- *Trade-Off Between Performance and Interpretability:* As we all know, AI models, such as decision trees, linear regression, are more explainable/interpretable than deep learning models such as convolutional neural networks, recurrent neural networks, etc. Deep learning models are good at their model performance but poor in interpretability. It can be seen as a seesaw effect in multi-task learning problems (Zhang and Yang 2022), where original problems and models' interpretability could be seen as two different tasks.

From Equation (2), it is obviously that to maximise the overall objective of XAI, we need to maximise both original prediction performance and interpretable performance. There must be a trade-off between these two goals, which is a core challenge in applying AI to B&E. Solving this trade-off requires innovative approaches, such as the use of hybrid models, and human-machine interactive systems. Investment in tailored XAI methods and collaboration between AI experts and business stakeholders can find out a balance between high original performance and explainability. This balance is critical to fostering trust in AI, ensuring regulatory compliance, and making informed decisions in complex B&E problems.

- *XAI for Large Language Model:* Large language models (LLMs), such as BERT, GPT-3, GPT-4, and LLaMA-2, have impressive performance across a wide range of natural language processing (NLP) tasks, which have been widely used into B&E. Leading technology companies, such as Microsoft, Google, and Baidu, have deployed LLMs in their commercial products and services (Zhao et al. 2024). For instance, Microsoft leverages GPT-3.5 to improve search

relevance ranking in new Bing.com. Since LLMs are complex black-box systems, their inner working mechanisms are opaque, and the high complexity makes model interpretation much challenging. Therefore, it is critical to develop explainability to shed light on how these powerful models work. However, due to the complex logistic of LLMs, it is quite hard to achieve this goal.

- *Foundational Theoretical Work in XAI:* As discussed in Section 2, Shapley values and (deep) Taylor decomposition have been proposed as principled frameworks for formalising the task of explanation (Samek et al. 2021). However, many theoretical questions remain. For example, it remains unclear how to incorporate the model and data distribution into the explanation. Related to this is causality, which assumes that there is a causal relationship between two input variables, but has not yet answered whether both variables or only the source variable must constitute the explanation. A deeper formalisation and theoretical understanding of XAI will help to shed light on these vital questions.
- *Interaction with users:* The effectiveness of explanations depends largely on the user's ability to understand. Designing explanation interfaces and interactive methods that are easy for users to understand and operate remains a huge challenge. Different users have different needs and preferences for explanations, and there is currently a lack of in-depth research on how to customise explanations to meet the needs of different users.

Current explanation methods are often unidirectional, that is, the explanation is from the model to the user, excluding user feedback. However, effectively integrate user feedback into the explanation process is also very important and necessary. Mechanisms need to be designed to collect and integrate user feedback to improve the explanation.

- *Less of a framework for Integrating XAI technologies:* As far as we know, there is no existing framework that fully integrates different XAI technologies or behaves like artificial general intelligence (AGI) and ChatGPT, which poses an obstacle for laypeople. But there are some initiatives and frameworks that attempt to move in that direction. While there is no single universal framework that integrates all XAI methods, there is effort and research that aims to integrate multiple XAI tools, such as InterpretML, AI Explainability 360, and TensorFlow Model Analysis.

Overall, these challenges correspond to numerous research opportunities in the adoption of XAI for B&E, and emphasise the relationship between challenges and opportunities, highlighting the necessary in XAI in B&E.

5 | Conclusions

The rapid development of artificial intelligence (AI) has created significant opportunities in various fields, particularly in B&E. However, the application of AI in B&E faces numerous challenges, as AI models often operate as 'black boxes', making their decision-making processes difficult to understand. Explainable Artificial Intelligence (XAI) aims to provide insights into the rationale behind AI decisions, facilitating its application in these fields.

This paper reviewed XAI techniques and their applications across different areas of B&E. It proposed a new taxonomy for understanding the evolution of XAI techniques. To bridge the gap between theoretical taxonomy and practical implementation, the paper summarised various applications of XAI in B&E. By identifying key challenges and opportunities, we aimed to guide future research efforts and promote collaboration in this domain.

In conclusion, realising the potential of XAI in B&E requires a concerted effort to address challenges and seize opportunities. Ongoing research in XAI for B&E is essential to building a more comprehensible, transparent, fair, and intelligent future.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Endnotes

¹ EU AI Act: <https://artificialintelligenceact.eu/wp-content/uploads/2021/08/The-AI-Act.pdf>.

References

- Acemoglu, D. 2024. "The Simple Macroeconomics of AI." no. 32487.
- Adadi, A., and M. Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–52160.
- Adhegaonkar, V. R., A. R. Thakur, N. Varghese, et al. 2024. "Advancing Credit Card Fraud Detection Through Explainable Machine Learning Methods." In *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 792–796. IEEE.
- Agarwal, C., S. Krishna, E. Saxena, et al. 2022. "Openxai: Towards a Transparent Evaluation of Model Explanations." *Advances in Neural Information Processing Systems* 35: 15784–15799.
- Agarwal, R., L. Melnick, N. Forsst, et al. 2021. "Neural Additive Models: Interpretable Machine Learning With Neural Nets." *Advances in Neural Information Processing Systems* 34: 4699–4711.
- Agbozo, E., and D. M. Balungu. 2024. "Liver Disease Classification-An XAI Approach to Biomedical AI." *Informatica* 48: 1.
- Ali, S., T. Abuhmed, S. EI-Sappagh, et al. 2023. "Explainable Artificial Intelligence (XAI): What we Know and What Is Left to Attain Trustworthy Artificial Intelligence." *Information Fusion* 99: 101805.
- Al-Khafaji, A., and O. Karan. 2023. "Explainable AI for Predicting User Behavior in Digital Advertising." In *International Conference on Emerging Trends and Applications in Artificial Intelligence*, 520–531. IEEE.
- Altmann, A., L. Tološi, O. Sander, et al. 2010. "Permutation Importance: A Corrected Feature Importance Measure." *Bioinformatics* 26, no. 10: 1340–1347.
- Ancona, M., E. Ceolini, C. Öztireli, et al. 2017. "Towards Better Understanding of Gradient-based Attribution Methods for Deep Neural Networks." In *6th International Conference on Learning Representations ICLR 2018*, Vancouver, BC, Canada.
- Angerschmid, A., J. Zhou, K. Theuermann, et al. 2022. "Fairness and Explanation in AI-Informed Decision Making." *Machine Learning and Knowledge Extraction* 4, no. 2: 556–579.
- Anjomshoe, S., A. Najjar, D. Calvaresi, et al. 2019. "Explainable Agents and Robots: Results From a Systematic Literature Review." In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, 1078–1088. IEEE.
- Apley, D. W., and J. Zhu. 2020. "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 82, no. 4: 1059–1086.
- Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115.
- Arya, V., R. K. Bellamy, P. Y. Chen, et al. 2021. "AI Explainability 360 Toolkit." In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, 376–379. Association for Computing Machinery.
- Averkin, A., and S. Yarushev. 2021. "Explainable Artificial Intelligence: Rules Extraction From Neural Networks." In *International Conference on Theory and Application of Soft Computing, Computing With Words and Perceptions*, 102–109. Springer.
- Bach, S., A. Binder, G. Montavon, et al. 2015. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation." *PLoS One* 10, no. 7: e0130140.
- Bejger, S., and S. Elster. 2020. "Artificial Intelligence in Economic Decision Making: How to Assure a Trust?" *Ekonomia I Prawo. Economics and Law* 19, no. 3: 411–434.
- Bertsimas, D., and N. Kallus. 2020. "From Predictive to Prescriptive Analytics." *Management Science* 66, no. 3: 1025–1044.
- Bharadiya, J. P. 2023. "Machine Learning and AI in Business Intelligence: Trends and Opportunities." *International Journal of Computer* 48, no. 1: 123–134.
- Bhatia, S., and A. S. Albarrak. 2023. "A Blockchain-Driven Food Supply Chain Management Using Qr Code and Xai-Faster Rcn Architecture." *Sustainability* 15, no. 3: 2579.
- Biecek, P. 2018. "Dalex: Explainers for Complex Predictive Models in R." *Journal of Machine Learning Research* 19, no. 84: 1–5.
- Bora, A., R. Sah, A. Singh, et al. 2022. "Interpretation of Machine Learning Models Using XAI-A Study on Health Insurance Dataset." In *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 1–6. IEEE.
- Brathwaite, T., A. Vij, J. L. Walker, et al. 2017. "Machine Learning Meets Microeconomics: The Case of Decision Trees and Discrete Choice." *arXiv Preprint arXiv:1711.04826*.
- BuyukteCatal, C., G. Kar, Y. Bouzembrak, et al. 2023. "Food Fraud Detection Using Explainable Artificial Intelligence." *Expert Systems* 42: e13387.
- Cao, L. 2022. "AI in Finance: Challenges, Techniques, and Opportunities." *ACM Computing Surveys (CSUR)* 55, no. 3: 1–38.
- Carrizosa, E., J. Ramírez-Ayerbe, D. R. Morales, et al. 2024. "Mathematical Optimization Modelling for Group Counterfactual Explanations." *European Journal of Operational Research* 319, no. 2: 399–412.
- Černevčienė, J., and A. Kabašinskas. 2024. "Explainable Artificial Intelligence (XAI) in Finance: A Systematic Literature Review." *Artificial Intelligence Review* 57, no. 8: 216.
- Chang, V., Q. A. Xu, S. H. Akinloye, et al. 2024. "Prediction of Bank Credit Worthiness Through Credit Risk Analysis: An Explainable Machine Learning Study." *Annals of Operations Research*: 1–25.
- Chapman, J. T. E., and A. Desai. 2023. "Macroeconomic Predictions Using Payments Data and Machine Learning." *Forecast* 5, no. 4: 652–683.
- Chen, H., S. Lundberg, S. I. Lee, et al. 2019. "Explaining Models by Propagating Shapley Values of Local Components." *arXiv Preprint arXiv: 1911.11888*.

- Chen, S., E. W. Ngai, Y. Ku, et al. 2023. "Prediction of Hotel Booking Cancellations: Integration of Machine Learning and Probability Model Based on Feature Interaction." *Decision Support Systems* 170: 113959.
- Chen, Y., R. Calabrese, B. Martin-Barragan, et al. 2024. "Interpretable Machine Learning for Imbalanced Credit Scoring Datasets." *European Journal of Operational Research* 312, no. 1: 357–372.
- Chien, S.-Y., C. J. Yang, F. Yu, et al. 2022. "XFlag: Explainable Fake News Detection Model on Social Media." *International Journal of Human Computer Interaction* 38, no. 18–20: 1808–1827.
- Chowdhury, S., P. Dey, S. Joel-Edgar, et al. 2023. "Unlocking the Value of Artificial Intelligence in Human Resource Management Through AI Capability Framework." *Human Resource Management Review* 33, no. 1: 100899.
- Collett, C., L. G. Gomes, G. Neff, et al. 2022. *The Effects of AI on the Working Lives of Women*. UNESCO Publishing.
- Cukier, R. I., C. M. Fortuin, K. E. Shuler, et al. 1973. "Study of the Sensitivity of Coupled Reaction Systems to Uncertainties in Rate Coefficients. I Theory." *Journal of Chemical Physics* 59, no. 8: 3873–3878.
- Delecraz, S., L. Eltarr, M. Becuwe, et al. 2022. "Responsible Artificial Intelligence in Human Resources Technology: An Innovative Inclusive and Fair by Design Matching Algorithm for Job Recruitment Purposes." *Journal of Responsible Technology* 11: 100041.
- Dignum, V. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Vol. 1. Springer.
- Dumitrescu, E., S. Hué, C. Hurlin, et al. 2022. "Machine Learning for Credit Scoring: Improving Logistic Regression With Non-Linear Decision-Tree Effects." *European Journal of Operational Research* 297, no. 3: 1178–1192.
- Dwivedi, R., D. Dave, H. Naik, et al. 2023. "Explainable AI (XAI): Core Ideas, Techniques, and Solutions." *ACM Computing Surveys* 55, no. 9: 1–33.
- ElShawi, R., Y. Sherif, M. Al-Mallah, et al. 2021. "Interpretability in Healthcare: A Comparative Study of Local Machine Learning Interpretability Techniques." *Computational Intelligence* 37, no. 4: 1633–1650.
- Fan, M., Z. Mo, Q. Zhao, et al. 2023. "Innovative Insights Into Knowledge-Driven Financial Distress Prediction: A Comprehensive XAI Approach." *Journal of the Knowledge Economy* 15: 12554–12595.
- Ford, J., V. Jain, K. Wadhvani, et al. 2023. "AI Advertising: An Overview and Guidelines." *Journal of Business Research* 166: 114124.
- Freisleben, T. 2022. "The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples." *Minds and Machines* 32, no. 1: 77–109.
- Friedman, J. H., and B. E. Popescu. 2008. "Predictive Learning via Rule Ensembles." *Annals of Applied Statistics* 2: 3.
- Fukui, H., T. Hirakawa, T. Yamashita, et al. 2019. "Attention Branch Network: Learning of Attention Mechanism for Visual Explanation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10705–10714. IEEE.
- Garreau, D., and U. Luxburg. 2020. "Explaining the Explainer: A First Theoretical Analysis of LIME." *International Conference on Artificial Intelligence and Statistics* 108: 1287–1296.
- Ge, W., C. Lueck, H. Suominen, et al. 2023. "Has Machine Learning Over-Promised in Healthcare?: A Critical Analysis and a Proposal for Improved Evaluation, With Evidence From Parkinson's Disease." *Artificial Intelligence in Medicine* 139: 102524.
- Ghosh, I., and R. K. Jana. 2024. "Clean Energy Stock Price Forecasting and Response to Macroeconomic Variables: A Novel Framework Using Facebook's Prophet, NeuralProphet and Explainable AI." *Technological Forecasting and Social Change* 200: 123148.
- Goldstein, A., A. Kapelner, J. Bleich, et al. 2015. "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation." *Journal of Computational and Graphical Statistics* 24, no. 1: 44–65.
- Gramegna, A., and P. Giudici. 2020. "Why to Buy Insurance? An Explainable Artificial Intelligence Approach." *Risks* 8: 4.
- Greenwell, B. M. 2017. "Pdp: An R Package for Constructing Partial Dependence Plots." *R Journal* 9, no. 1: 421.
- Guidotti, R. 2022. "Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking." *Data Mining and Knowledge Discovery* 38: 2770–2824.
- Hakami, N. 2023. "Navigating the Microeconomic Landscape of Artificial Intelligence: A Scoping Review." *Migration Letters* 20, no. S2: 1018–1031.
- Haleem, A., M. Javaid, M. A. Qadri, et al. 2022. "Artificial Intelligence (AI) Applications for Marketing: A Literature-Based Study." *International Journal of Intelligent Networks* 3: 119–132.
- Hall, P., M. Kurka, A. Bartz, et al. 2017. *Using H2O Driverless AI*. H2O. ai.
- Hapfelmeier, A., R. Hornung, B. Haller, et al. 2023. "Efficient Permutation Testing of Variable Importance Measures by the Example of Random Forests." *Computational Statistics & Data Analysis* 181: 107689.
- Harezlak, J., D. Ruppert, M. P. Wand, et al. 2018. "Generalized Additive Models." In *Statistical Models With R*, 71–128. Springer.
- Hashemi, M., and A. Fathi. 2020. "PermuteAttack: Counterfactual Explanation of Machine Learning Credit Scorecards." arXiv Preprint arXiv: 2008.10138.
- Hassan, M. M., S. A. AlQahtani, A. Alelaiwi, et al. 2023. "Explaining Covid-19 Diagnosis With Taylor Decompositions." *Neural Computing and Applications* 35, no. 30: 22087–22100.
- Herhausen, D., S. F. Bernritter, E. W. Ngai, et al. 2024. "Machine Learning in Marketing: Recent Progress and Future Research Directions." *Journal of Business Research* 170: 114254.
- Hofeditz, L., A. Rieß, and S. Clausen. 2022. "Applying XAI to an AI-Based System for Candidate Management to Mitigate Bias and Discrimination in Hiring." *Electronic Markets* 32, no. 4: 2207–2233.
- Hülter, S. M., C. Ertel, A. Heidemann, et al. 2024. "Exploring the Individual Adoption of Human Resource Analytics: Behavioural Beliefs and the Role of Machine Learning Characteristics." *Technological Forecasting and Social Change* 208: 123709.
- Hung, Y.-H., and C.-Y. Lee. 2024. "BMB-LIME: LIME With Modeling Local Nonlinearity and Uncertainty in Explainability." *Knowledge-Based Systems* 294: 111732.
- Islam, M. A., M. F. Mridha, M. A. Jahin, et al. 2024. "A Unified Framework for Evaluating the Effectiveness and Enhancing the Transparency of Explainable AI Methods in Real-World Applications." arXiv Preprint arXiv:2412.03884.
- Janizek, J. D., P. Sturmfels, S. I. Lee, et al. 2021. "Explaining Explanations: Axiomatic Feature Interactions for Deep Networks." *Journal of Machine Learning Research* 22, no. 104: 1–54.
- Janssens, B., L. Schetgen, M. Bogaert, et al. 2024. "360 Degrees Rumor Detection: When Explanations Got Some Explaining to Do." *European Journal of Operational Research* 317, no. 2: 366–381.
- Javaid, M., A. Haleem, R. P. Singh, et al. 2022. "Significance of Machine Learning in Healthcare: Features, Pillars and Applications." *International Journal of Intelligent Networks* 3: 58–73.
- Johnson, M., A. Albizri, A. Harfouche, et al. 2022. "Integrating Human Knowledge Into Artificial Intelligence for Complex and Ill-Structured

- Problems: Informed Artificial Intelligence." *International Journal of Information Management* 64: 102479.
- Jose, A., and S. D. Shetty. 2022. "Interpretable Click-Through Rate Prediction Through Distillation of the Neural Additive Factorization Model." *Information Sciences* 617: 91–102.
- Kaur, D., S. Uslu, K. J. Rittichier, et al. 2022. "Trustworthy Artificial Intelligence: A Review." *ACM Computing Surveys (CSUR)* 55, no. 2: 1–38.
- Kenny, E. M., and M. T. Keane. 2021. "Explaining Deep Learning Using Examples: Optimal Feature Weighting Methods for Twin Systems Using Post-Hoc, Explanation-By-Example in XAI." *Knowledge-Based Systems* 233: 107530.
- Kochenderfer, M. J. 2015. *Decision Making Under Uncertainty: Theory and Application*. MIT Press.
- Koh, P. W., and P. Liang. 2020. "Understanding Black-Box Predictions via Influence Functions." arXiv Preprint arXiv: 1703.04730.
- Kontschieder, P., M. Fiterau, A. Criminisi, et al. 2015. "Deep Neural Decision Forests." In *Proceedings of the IEEE International Conference on Computer Vision*, 1467–1475. IEEE.
- Kumar, P., and V. S. Kumar. 2023. "Enhancing Supplier Selection Through Explainable AI: A Transparent and Interpretable Approach." In *2023 IEEE International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering, RKMATE*, 2023. IEEE.
- Li, X., H. Xiong, X. Li, et al. 2023. "G-LIME: Statistical Learning for Local Interpretations of Deep Neural Networks Using Global Priors." *Artificial Intelligence* 314: 103823.
- Liu, G., J. Zhang, A. B. Chan, et al. 2023. "Human Attention-Guided Explainable AI for Object Detection." In *the 4th Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 45.
- Longo, L., M. Brcic, F. Cabitza, et al. 2024. "Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions." *Information Fusion* 106: 102301.
- Lozano-Murcia, C., F. P. Romero, J. Serrano-Guerrero, et al. 2023. "A Comparison Between Explainable Machine Learning Methods for Classification and Regression Problems in the Actuarial Context." *Mathematics* 11, no. 14: 3088.
- Lundberg, S. 2017. "A Unified Approach to Interpreting Model Predictions." arXiv Preprint arXiv:1705.07874.
- Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems*. Curran Associates Inc.
- Mahbooba, B., M. Timilsina, R. Sahal, et al. 2021. "Explainable Artificial Intelligence (XAI) to Enhance Trust Management in Intrusion Detection Systems Using Decision Tree Model." *Complexity* 2021: 1–11.
- Marín Díaz, G., J. J. Galán Hernández, J. L. Galdón Salvador, et al. 2023. "Analyzing Employee Attrition Using Explainable AI for Strategic HR Decision-Making." *Mathematics* 11: 22.
- Martens, D., B. Baesens, T. Van Gestel, et al. 2007. "Comprehensible Credit Scoring Models Using Rule Extraction From Support Vector Machines." *European Journal of Operational Research* 183, no. 3: 1466–1476.
- Miller, T. 2019. "Explanation in Artificial Intelligence: Insights From the Social Sciences." *Artificial Intelligence* 267: 1–38.
- Miró-Nicolau, M., A. Jaume-i-Capó, G. Moyà-Alcover, et al. 2025. "A Comprehensive Study on Fidelity Metrics for XAI." *Information Processing & Management* 62, no. 1: 103900.
- Mizanur Rahman, S. M., and M. G. R. Alam. 2023. "Explainable Loan Approval Prediction Using Extreme Gradient Boosting and Local Interpretable Model Agnostic Explanations." In *International Congress on Information and Communication Technology*, 791–804. Springer.
- Mohseni, S., N. Zarei, E. D. Ragan, et al. 2021. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, no. 3–4: 1–45.
- Montavon, G., S. Lapuschkin, A. Binder, et al. 2017. "Explaining Nonlinear Classification Decisions With Deep Taylor Decomposition." *Pattern Recognition* 65: 211–222.
- Montavon, G., A. Binder, S. Lapuschkin, et al. 2019. "Layer-Wise Relevance Propagation: An Overview." *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* 11700: 193–209.
- Murdoch, W. J., C. Singh, K. Kumbier, et al. 2019. "Interpretable Machine Learning: Definitions, Methods, and Applications." arXiv Preprint arXiv:1901.04592.
- Nguyen, H. T. T., L. P. T. Nguyen, H. Cao, et al. 2024. "XEdgeAI: A Human-Centered Industrial Inspection Framework With Data-Centric Explainable Edge AI Approach." *Information Fusion* 116: 102782.
- Nori, H., S. Jenkins, P. Koch, et al. 2019. "Interpretml: A Unified Framework for Machine Learning Interpretability." arXiv Preprint arXiv:1909.09223.
- Novaković, J. D., A. Veljović, S. S. Ilić, et al. 2017. "Evaluation of Classification Models in Machine Learning." *Theory and Applications of Mathematics & Computer Science* 7, no. 1: 39.
- Olan, F., K. Spanaki, W. Ahmed, et al. 2024. "Enabling Explainable Artificial Intelligence Capabilities in Supply Chain Decision Support Making." *Production Planning and Control*: 1–12.
- Park, S., and J.-S. Yang. 2022. "Interpretable Deep Learning LSTM Model for Intelligent Economic Decision-Making." *Knowledge-Based Systems* 248: 108907.
- Pawlicki, M., A. Pawlicka, F. Uccello, et al. 2024. "Evaluating the Necessity of the Multiple Metrics for Assessing Explainable AI: A Critical Examination." *Neurocomputing* 602: 128282.
- Peng, J., K. Zou, M. Zhou, et al. 2021. "An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients." *Journal of Medical Systems* 45: 5.
- Pournader, M., H. Ghaderi, A. Hassanzadegan, et al. 2021. "Artificial Intelligence Applications in Supply Chain Management." *International Journal of Production Economics* 241: 108250.
- Qiao, L., W. Wang, B. Lin, et al. 2021. "Learning Accurate and Interpretable Decision Rule Sets From Neural Networks." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 4303–4311. AAAI Press.
- Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1: 81–106.
- Ramachandran, V., A. R. Kavitha, R. Pandimeena, et al. 2023. "An Accurate Prediction of Medical Insurance Cost Using Forest Regression Algorithms." In *2023 International Conference on Data Science, Agents and Artificial Intelligence, ICDSAAI 2023*. IEEE.
- Ransbotham, S., D. Kiron, P. Gerbert, et al. 2017. "Reshaping Business With Artificial Intelligence: Closing the Gap Between Ambition and Action." *MIT Sloan Management Review* 59: 1.
- Razavi, S., A. Jakeman, A. Saltelli, et al. 2021. "The Future of Sensitivity Analysis: An Essential Discipline for Systems Modeling and Policy Support." *Environmental Modelling & Software* 137: 104954.
- Ribeiro, M. T., S. Singh, C. Guestrin, et al. 2016a. "“Why Should I Trust You?” Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

- Ribeiro, M. T., S. Singh, C. Guestrin, et al. 2016b. "Model-Agnostic Interpretability of Machine Learning." *arXiv Preprint arXiv: 1606.05386*.
- Roscher, R., B. Bohn, M. F. Duarte, et al. 2020. "Explainable Machine Learning for Scientific Insights and Discoveries." *IEEE Access* 8: 42200–42216.
- Sachan, S., J. B. Yang, D. L. Xu, et al. 2020. "An Explainable AI Decision-Support-System to Automate Loan Underwriting." *Expert Systems With Applications* 144: 113100.
- Sadeghi, K., D. Ojha, P. Kaur, et al. 2024. "Explainable Artificial Intelligence and Agile Decision-Making in Supply Chain Cyber Resilience." *Decision Support Systems* 180: 114194.
- Saeed, W., and C. Omlin. 2023. "Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities." *Knowledge-Based Systems* 263: 110273.
- Saleiro, P., B. Kuester, L. Hinkson, et al. 2018. "Aequitas: A Bias and Fairness Audit Toolkit." *arXiv Preprint arXiv:1811.05577*.
- Samek, W., G. Montavon, S. Lapuschkin, et al. 2021. "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications." *Proceedings of the IEEE* 109, no. 3: 247–278.
- Sato, M., and H. Tsukimoto. 2001. "Rule Extraction From Neural Networks via Decision Tree Induction." In *IJCNN 2001. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, 1870–1875. IEEE.
- Selvaraju, R. R., M. Cogswell, A. Das, et al. 2017. "Grad-Cam: Visual Explanations From Deep Networks via Gradient-Based Localization." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–2951. Association for Computational Linguistics.
- Shi, W., L. Tong, Y. Zhuang, et al. 2020. "Exam: An Explainable Attention-Based Model for Covid-19 Automatic Diagnosis." In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–6. ACM.
- Shrikumar, A., P. Greenside, A. Kundaje, et al. 2017. "Learning Important Features Through Propagating Activation Differences." In *International Conference on Machine Learning*, 3145–3153. IEEE.
- Silva, D., L. B. Pessoa, R. Soltovski, et al. 2022. "Human Resources Management 4.0: Literature Review and Trends." *Computers & Industrial Engineering* 168: 108111.
- da, F. L., B. K. Slodkowski, K. K. A. da Silva, et al. 2023. "A Systematic Literature Review on Educational Recommender Systems for Teaching and Learning: Research Trends, Limitations and Opportunities." *Education and Information Technologies* 28, no. 3: 3289–3328.
- Simonyan, K., A. Vedaldi, A. Zisserman, et al. 2013. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." *arXiv Preprint arXiv: 1312.6034*.
- Smilkov, D., N. Thorat, B. Kim, et al. 2017. "SmoothGrad: Removing Noise by Adding Noise." *arXiv Preprint arXiv: 1706.03825*.
- Sobol', I. M., D. Asotsky, A. Kreinin, et al. 2011. "Construction and Comparison of High-Dimensional Sobol'Generators." *Wilmott* 2011, no. 56: 64–79.
- Sobol', I. M. 1990. "On Sensitivity Estimation for Nonlinear Mathematical Models." *Matematicheskoe Modelirovanie* 2, no. 1: 112–118.
- Sundararajan, M., A. Taly, Q. Yan, et al. 2017. "Axiomatic Attribution for Deep Networks." In *International Conference on Machine Learning*, 3319–3328. IEEE.
- Szeląg, M., and R. Słowiński. 2024. "Explaining and Predicting Customer Churn by Monotonic Rules Induced From Ordinal Data." *European Journal of Operational Research* 317, no. 2: 414–424.
- Tchuente, D., J. Lonlac, B. Kamsu-Foguem, et al. 2024. "A Methodological and Theoretical Framework for Implementing Explainable Artificial Intelligence (XAI) in Business Applications." *Computers in Industry* 155: 104044.
- Tiwari, R. 2023. "Explainable AI (XAI) and Its Applications in Building Trust and Understanding in AI Decision Making." *International Journal of Scientific Research in Engineering and Management* 7: 1.
- Toorajipour, R., and V. Sohrabpour. 2021. "Artificial Intelligence in Supply Chain Management: A Systematic Literature Review." *Journal of Business Research* 122: 502–517.
- Tsang, M., D. Cheng, Y. Liu, et al. 2018. "Detecting Statistical Interactions From Neural Network Weights." In *6th International Conference on Learning Representations ICLR 2018*, Vancouver, BC, Canada.
- Tsang, M., S. Rambhatla, Y. Liu, et al. 2020. "How Does This Interaction Affect Me? Interpretable Attribution for Feature Interactions." In *Advances in Neural Information Processing Systems, 34th Conference on Neural Information Processing Systems*, edited by H. Larochelle, M. A. Ranzato, R. Hadsell, M. F. Balcan, and H. T. Lin, vol. 33, 6147–6159. NeurIPS.
- Tzougas, G., and K. Kutzkov. 2023. "Enhancing Logistic Regression Using Neural Networks for Classification in Actuarial Learning." *Algorithms* 16: 2.
- Urazova, S. A. 2023. *Artificial Intelligence in Banking Systems: Trends and Possible Consequences of Implementation*, 345–355. Springer Nature Singapore.
- Vaid, S., S. Puntoni, A. Khodr, et al. 2023. "Artificial Intelligence and Empirical Consumer Research: A Topic Modeling Analysis." *Journal of Business Research* 166: 114110.
- van der Crujisen, C., J. de Haan, R. Roerink, et al. 2023. "Trust in Financial Institutions: A Survey." *Journal of Economic Surveys* 37, no. 4: 1214–1254.
- van der Waa, J., E. Nieuwburg, A. Cremers, et al. 2021. "Evaluating XAI: A Comparison of Rule-Based and Example-Based Explanations." *Artificial Intelligence* 291: 103404.
- Varian, H. R. 2018. *Artificial Intelligence, Economics, and Industrial Organization*. Vol. 24839. National Bureau of Economic Research Cambridge.
- Vaswani, A. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, 6000–6010. Curran Associates Inc.
- Vaughan, J., A. Sudjianto, E. Brahimi, et al. 2018. "Explainable Neural Networks Based on Additive Index Models." *arXiv Preprint arXiv: 1806.01933*.
- Wachter, S., B. Mittelstadt, C. Russell, et al. 2017. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law & Technology* 31: 841.
- Wang, D. Y. C., L. A. Jordanger, J. C. W. Lin, et al. 2023. "Explainability of Leverage Points Exploration for Customer Churn Prediction." In *Proceedings -2023 IEEE International Conference on Big Data, BigData*, vol. 2023, 5997–6004. IEEE.
- Wang, N., M. Chen, K. P. Subbalakshmi, et al. 2021. "Explainable CNN-Attention Networks (C-Attention Network) for Automated Detection of Alzheimer's Disease." *arXiv Preprint arXiv: 2066.14135*.
- Wang, W., C. Lesner, A. Ran, et al. 2020. "Using Small Business Banking Data for Explainable Credit Risk Scoring." In *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 8: 13396–13401.
- Weber, P., K. V. Carl, O. Hinz, et al. 2024. "Applications of Explainable Artificial Intelligence in Finance—A Systematic Review of Finance, Information Systems, and Computer Science Literature." *Management Review Quarterly* 74, no. 2: 867–907.
- Wiegrefe, S., and Y. Pinter. 2019. "Attention Is Not Not Explanation." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on*

Natural Language Processing (EMNLP-IJCNLP), 11–20. Association for Computational Linguistics.

Wood, S. N. 2017. *Generalized Additive Models: An Introduction With R*. Chapman and Hall/CRC.

Wu, M., M. Hughes, S. Parbhoo, et al. 2018. “Beyond Sparsity: Tree Regularization of Deep Models for Interpretability.” In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

Yang, W., H. Le, T. Laud, et al. 2022. “Omnixai: A Library for Explainable AI.” *arXiv Preprint arXiv:2206.01612*.

Yankol-Schalck, M. 2022. “The Value of Cross-Data Set Analysis for Automobile Insurance Fraud Detection.” *Research in International Business and Finance* 63: 101769.

Yilmaz, D., and I. Esra Buyuktahtakın. 2024. “An Wxpandable Machine Learning-Optimization Framework to Sequential Decision-Making.” *European Journal of Operational Research* 314, no. 1: 280–296.

Yoneda, A., R. Shimizu, S. Sakurai, et al. 2024. “Effectiveness Verification Framework for Coupon Distribution Marketing Measure Considering Users’ Potential Purchase Intentions.” *Cogent Engineering* 11, no. 1: 2307718.

Yuan, R., R. M. Gower, A. Lazaric, et al. 2022. “A General Sample Complexity Analysis of Vanilla Policy Gradient.” *International Conference on Artificial Intelligence and Statistics* 151: 3332–3380.

Yudkowsky, E. 2013. “Intelligence Explosion Microeconomics.” *Machine Intelligence Research Institute* 23: 2015.

Yue, T., and D. Au. 2023. “Harnessing ChatGPT-4 and Explainable AI for Financial Nowcasting.” *SSRN Electronic Journal* 4633383: 1–33.

Zarifis, A., C. P. Holland, A. Milne, et al. 2023. “Evaluating the Impact of AI on Insurance: The Four Emerging AI- and Data-Driven Business Models.” *Emerald Open Research* 1: 1.

Zhang, J., Z. Li, X. Song, et al. 2021a. “Deep Tobit Networks: A Novel Machine Learning Approach to Microeconometrics.” *Neural Networks* 144: 279–296.

Zhang, Y., F. Xu, J. Zou, et al. 2021b. “XAI Evaluation: Evaluating Black-Box Model Explanations for Prediction.” In *In 2021 II International Conference on Neural Networks and Neurotechnologies (NeuroNT)*, 13–16. IEEE.

Zhang, Y., and Q. Yang. 2022. “A Survey on Multi-Task Learning.” *IEEE Transactions on Knowledge and Data Engineering* 34, no. 12: 5586–5609.

Zhao, H., H. Chen, F. Yang, et al. 2024. “Explainability for Large Language Models: A Survey.” *ACM Transactions on Intelligent Systems and Technology* 15: 2.

Zilke, J. R., E. Loza Mencia, F. Janssen, et al. 2016. “Deepred–Rule Extraction From Deep Neural Networks.” In *Discovery Science: 19th International Conference, DS 2016*, vol. 19, 457–473. Springer.