Computers in Industry 155 (2024) 104044

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.sciencedirect.com/journal/computers-in-industry

A methodological and theoretical framework for implementing explainable artificial intelligence (XAI) in business applications

Dieudonné Tchuente[a,*], Jerry Lonlac[b], Bernard Kamsu-Foguem[c]

[a] TBS Business School, Department of Information, Operations and Management Sciences, 1 Place Alphonse Jourdain - CS 66810, 31068 TOULOUSE Cedex 7, France
[b] IMT Nord Europe, University of Lille, Center for Digital Systems, Cité scientifique, Rue Guglielmo Marconi, BP 20145, 59653 Villeneuve d'Ascq Cedex, France
[c] Université de Toulouse, Laboratoire de Génie de Production (LGP), EA 1905, 47 Avenue d'Azereix, BP 1629, 65016 Tarbes Cedex, France

ARTICLE INFO

Keywords:
Explainable artificial intelligence
XAI
Interpretable artificial intelligence
Interpretable machine learning
Deep learning
Business applications
Framework
Management
TCCM

ABSTRACT

Artificial Intelligence (AI) is becoming fundamental in almost all activity sectors in our society. However, most of the modern AI techniques (e.g., Machine Learning – ML) have a black box nature, which hinder their adoption by practitioners in many application fields. This issue raises a recent emergence of a new research area in AI called Explainable artificial intelligence (XAI), aiming at providing AI-based decision-making processes and outcomes to be easily understood, interpreted, and justified by humans. Since 2018, there has been an exponential growth of research studies on XAI, which has justified some review studies. However, these reviews currently focus on proposing taxonomies of XAI methods. Yet, XAI is by nature a highly applicative research field, and beyond XAI methods, it is also very important to investigate how XAI is concretely used in industries, and consequently derive the best practices to follow for better implementations and adoptions. There is a lack of studies on this latter point. To fill this research gap, we first propose a holistic review of business applications of XAI, by following the Theory, Context, Characteristics, and Methodology (TCCM) protocol. Based on the findings of this review, we secondly propose a methodological and theoretical framework in six steps that can be followed by all practitioners or stakeholders for improving the implementation and adoption of XAI in their business applications. We particularly highlight the need to rely on domain field and analytical theories to explain the whole analytical process, from the relevance of the business question to the robustness checking and the validation of explanations provided by XAI methods. Finally, we propose seven important future research avenues.

1. Introduction

Artificial intelligence (AI) is often used to refer to machines or computer systems able to exhibit human-like intelligence (Turing, 1950). Nowadays, AI becomes practically unavoidable in almost all areas of our society, particularly with the emergence of modern data-oriented AI approaches (e.g., machine learning – ML, deep learning – DL, reinforcement learning – RL) (Dalzochio et al., 2020; Kamm et al., 2023; Wamba et al., 2021); boosted by the explosion of data volumes and computer processing capacities with the advent of the big data era and associated technologies such as cloud computing and the Internet of Things.

However, despite the very high interest in AI technologies in recent years, the black-box nature of most data-oriented AI approaches hinders their adoption by practitioners of many application fields (Onchis and Gillich, 2021; Schwalbe and Finzel, 2023; Souza et al., 2023). These brakes are particularly motivated by the fact that black-box ML, DL, or RL approaches confer a lack of interpretability on their output, which leads to many potential problems when they are used for sensitive decision support systems that affect human lives in many application fields (e.g., healthcare, law, defense, banking). For example, in February 2019, the Polish government added an amendment to a banking law that gives a customer the right to receive an explanation in case of a negative credit decision. This is one of the direct consequences of implementing the General Data Protection Regulation (GDPR) in the European Union (EU) (European Banking Authority, 2020; European Commission, 2016). This means that a bank must be able to explain why a loan was not granted if the decision process was automatic using methods such as ML. Another example of a context where interpretability is fundamental is healthcare, where accurate and transparent decision making is crucial (e.g., medical diagnosis). It promotes trust, facilitates clinical justifications, ensures accountability, addresses ethical concerns, supports continuous

[*] Corresponding author.
E-mail address: d.tchuente@tbs-education.fr (D. Tchuente).

https://doi.org/10.1016/j.compind.2023.104044
Received 26 August 2023; Received in revised form 14 October 2023; Accepted 7 November 2023
Available online 17 November 2023

learning, and engages patients in their own healthcare journey (Bharati et al., 2023). More globally, the right to explanations of algorithmic decision making is becoming a major issue in our society (European Commission, 2016; T. W. Kim and Routledge, 2022).

To face this issue, Explainable AI (XAI) is a very recent development of AI aiming at providing explanations for AI models (Adadi and Berrada, 2018; Arrieta et al., 2020). As illustrated in Fig. 1 (the result of a search with the terms “explainable AI” or “XAI” or “explainable artificial intelligence” in November 2022 in the scientific Scopus database), scientific publications related to XAI really started to emerge in 2018 (44 documents) and are increasing at an exponential rate (e.g., 1618 documents in 2022).

The goal of XAI is to bridge the gap between the inherent complexity of certain AI algorithms, often referred to as "black-box" models, and the need for accountability, responsibility, transparency, and trust in AI systems (Adadi and Berrada, 2018). Accountability usually refers to the means to explain and justify one’s decisions and actions to the partners, users, and others with whom the system interacts. Responsibility usually refers to the role of people themselves and to the capability of AI systems to answer for one’s decision and identify errors and unexpected results. Transparency usually refers to the need to describe, inspect, and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment and the governance of used data. Trust usually refers to the confidence and belief that individuals or organizations place in the abilities, reliability, and ethical behavior of AI systems. Other initiatives focus on additional considerations, such as fairness, confidence, informativeness, causality, and transferability, for setting goals of XAI (Arrieta et al., 2020).

Given the high interest in XAI in recent years, some literature reviews have been proposed (Angelov et al., 2021; Arrieta et al., 2020; Das and Rad, 2020; Minh et al., 2022; Schwalbe and Finzel, 2023; Vilone and Longo, 2021) (see Table 1). However, these reviews mainly focus on providing clear definitions of terms or proposing taxonomies of methods with several characteristics (e.g., model-specific vs model-agnostic methods, intrinsic vs post hoc explanations, global vs local explanations, explanations on structured data such as tabular data vs explanations of unstructured data such as texts or images, surrogate methods, visualisation methods) (see Table 2 in Section 3 for more details and references about characteristics of XAI methods). In general, an XAI method is model agnostic when it can explain the predictions of any ML model (e.g., SHAP – Shapley Additive exPlanations, LIME – Local Interpretable Model-agnostic Explanations) (Lundberg and Lee, 2017; Ribeiro et al., 2016); otherwise, it is model specific (only explains the predictions of a specific ML model) (e.g., feature importance embedded in Random Forest implementation) (Breiman, 2001). An XAI technique is local when it explains a single prediction (e.g., Shapley values, LIME); it is also global when the entire model can be explained (e.g., features importance methods, SHAP) (Fisher et al., 2019; Friedman, 2001). An XAI method is post hoc when the explanation is provided only after the ML model is built; it is intrinsic when explanations are constructed during the model building (e.g., white-box ML models, such as decision trees). An XAI method can be a surrogate model when it constructs an interpretable model (e.g., a linear model) to approximate the predictions of the complex black-box model (e.g., an artificial neural network), so that we can provide conclusions about the black-box model by interpreting its surrogate. An XAI method can also be a visualisation helping to explore the patterns inside the model (e.g., partial dependence plots – PDP, plots of a neural unit in a deep neural network) or dedicated to unstructured data, such as images (e.g., Grad-CAM) (Selvaraju et al., 2017).

Focusing on XAI methods is important, but XAI is a highly applicative research area, and it is fundamental to also investigate how XAI methods are concretely used in real-world business applications, and consequently derive the best practices for better implementations or adoptions. There is a lack of review studies on the latter point. Even if some very recent review studies focus on the application of XAI in specific applications fields (e.g., Arashpour, 2023; Javed et al., 2023).

To resume, Table 1 shows most of the existing reviews articles focusing on XAI with their objective, scope (method-oriented or application-oriented) and their main findings or proposals. We can see that most of those reviews are method-oriented proposing definition of related concepts and taxonomies. Even if some reviews are application-oriented, they are focusing on only specific application fields (e.g., healthcare, smart cities, environmental management, biomedical imaging). To the best of our knowledge, there is currently no review study that investigates XAI usages with a holistic scope and a broader vision beyond a specific application field. The goal of this paper is to fill this research gap, the final objective being to propose a generic, practical, and theoretical framework that can be used for by all practitioners and stakeholders for a better implementation and adoption of XAI in business applications.

Thus, the main research question addressed in this study is (**RQ**): what are the best practices to follow for a suitable implementation and adoption of XAI in business applications?

To be able to effectively answer this research question, we will initially rely on understanding current practices in the state of the art by addressing the following four sub-research questions (**SRQ**):

SRQ 1. What are the main field theories and analytical theories used in applicative XAI studies?

SRQ 2. What are the main application contexts of XAI in companies?

SRQ 3. What are the characteristics of the most frequently used XAI methods in this context?

SRQ 4. What are the methodologies used in this context?

To answer these sub-research questions, we relied on best practices

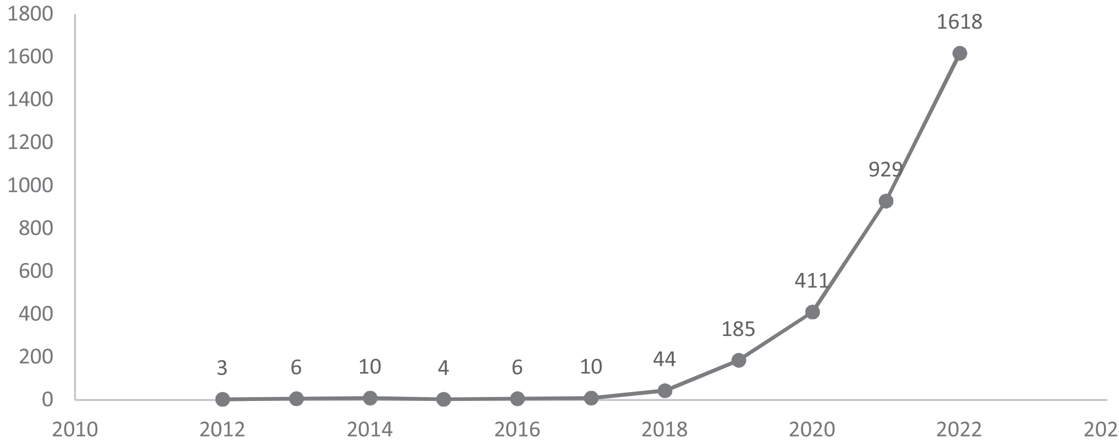


Fig. 1. Publication trends on XAI from 2012 to November 2022 (from Scopus database).

Table 1
Overview of existing reviews on XAI and the specificity of this study.

Review study	Main objective	Method oriented	Application oriented	Application fields	Main findings or proposal
(Adadi and Berrada, 2018)	Exploring XAI concepts, definitions, and methods	Yes	No	Holistic	Proposition of a taxonomy of methods. Depiction of major research trajectories.
(Arrieta et al., 2020)	Exploring XAI concepts, definitions, and methods	Yes	No	Holistic	Proposition of a taxonomy of methods and audiences' profiles. Depiction of XAI challenges and their impacts on the principles for Responsible AI
(Angelov et al., 2021)	Exploring XAI concepts, definitions, and methods	Yes	No	Holistic	Proposition of a taxonomy of methods.
(Schwalbe and Finzel, 2023)	Exploring existing surveys related to XAI methods' taxonomies	Yes	No	Holistic	Unification of existing taxonomies into a global taxonomy
(Das and Rad, 2020)	Exploring deep learning based XAI concepts, definitions, and methods	Yes	No	Holistic	Proposition of a taxonomy of methods. Timeline of seminal works towards XAI and explanation maps generated by eight XAI algorithms on image data
(Minh et al., 2022)	Exploring XAI concepts, definitions, and methods with a focus on deep learning explanations	Yes	No	Holistic	Proposition of a taxonomy of methods. Focus on the trade-off between the performance and the explainability, evaluation methods, security, and policy.
(Carvalho et al., 2019)	Exploring XAI methods and metrics to assess the quality of explanations	Yes	No	Holistic	Synthesis of the literature methods and metrics used to evaluate the quality of explanation provided by XAI.
(Vilone and Longo, 2021)	Clustering scientific studies related to XAI via a hierarchical system that classifies theories and notions related to the concept of explainability and the evaluation approaches for XAI methods.	Yes	No	Holistic	Two main approaches to assess the quality of explanations provided by XAI: human-centered evaluations and objective-metrics based evaluations.
(Nimmy et al., 2022)	Exploring the use of XAI in supply chain operational risk management	Yes	Yes	Supply chain operational risk management	Gaps in the existing AI approaches used in supply chain operational risk management in meeting the features of XAI.
(Arashpour, 2023)	Exploring the use of XAI in environmental management	Yes	Yes	Environmental management	An explainability framework with triadic structure to focus on input, AI model, and output.
(Javed et al., 2023)	Discussion of the concept of XAI for smart cities applications	Yes	Yes	Smart cities	XAI technology use cases, challenges, applications, possible alternative solutions, and current and future research enhancements
(Nazir et al., 2023)	Exploring the use of XAI for biomedical imaging diagnostics	Yes	Yes	Biomedical imaging diagnostics	Taxonomy of XAI techniques. Open challenges and future research directions
(Bharati et al., 2023)	Exploring XAI aspects and challenges in the healthcare domain	Yes	Yes	Healthcare	Taxonomy of XAI methods. Challenges of interpretability in healthcare
Our study	**Holistic investigation of how XAI is used in business applications and provide best practices to follow for a better implementation and adoption in industries**	**Yes**	**Yes**	**Holistic**	**Identification of main theories, contexts, characteristics of methods, and methodologies in the existing literature. Proposition of a holistic methodological and theoretical framework to be followed for a better implementation.**

for conducting systematic literature reviews (Grant and Booth, 2009; Page et al., 2021; Paul et al., 2021). We specifically screened 99 eligible articles on the subject from high-quality journals, and 37 relevant articles have been finally included for the analysis. We followed a recent and relevant framework-based review protocol, the Theory, Context, Characteristics, Methodology (TCCM) protocol (Paul and Rosado-Serrano, 2019) for the analysis. From the most important findings of this literature review, we draw a global picture of the current implementations which allowed us to propose a holistic theoretical and methodological framework providing best practices to follow for a suitable implementation and adoption of XAI in industries. Considering the proposed framework, we also identify the main gaps in the current literature that can be addressed in future research.

The main findings of the literature review show a very wide domain field applicative area, with more emphasis on economy and finance (25% of articles) and marketing (23% of articles). Most of the studies relied on post hoc methods (76% of articles) compared to intrinsic explanations provided by white-box or intrinsic methods (24% of articles). Supervised learning applications (classifications or regressions) are the main explained systems (92% of articles), while only a few articles (8% of articles) addressed specific predictive recommender systems. In terms of the explained data types, structured tabular data were the most frequently used (77% of articles), compared to unstructured textual data (13% of articles) or image data (10% of articles). From the methodological perspective, we found that most of the articles only focus on presenting sample outputs of explanations provided by XAI post-hoc methods (e.g., 44% of articles using SHAP, 17% of articles using LIME, 17% of articles using PDP), and very few articles also address the explanation of other analytical phases (e.g., business question importance, data collection, feature engineering), or address the robustness checking and the validation of the relevance of provided explanations by business users. From theoretical perspective, we found that only few articles relied on theories for providing better explanations of studied phenomena, particularly for analytical phases such as data collection and feature engineering. Thus, to provide a holistic view of best practices to be handled for a better implementation or adoption of XAI, we propose a synthesis into a global methodological and theoretical framework that could be followed by all practitioners for explaining the whole analytical process including six steps: business question explanation, data collection explanation, feature engineering explanation, modelling and predictive capacity explanations, models output (predictions) explanations, robustness checking and validation of explanations. For each of these steps, we present examples of theories or regulations used in some relevant studies in the existing literature. Finally, we propose seven relevant future research avenues.

The rest of this paper is structured as follows: Section 2 elaborates on the methodology used for our systematic review. Section 3 presents the most important findings that answer the four sub-research questions. In light of these results, Section 4 presents and discusses the proposed theoretical and methodological framework to answer the main research question. This section also presents the implications for research, the implications for practice, and the limitations and future research directions of our study. Finally, Section 5 concludes this paper.

2. Methodology of the research

This study first relies on a systematic literature reviews (SLR) approach. In general, SLR differ from traditional narrative reviews by adopting methodological accuracy, systematization, exhaustiveness, and reproducibility (Mengist et al., 2020). Fig. 2 shows an overview of the methodology adopted for the SLR in this research through a chosen global framework (SALSA), a formalism used to represent some steps in this framework (PRISMA), and a protocol used to structure the analysis of selected articles (TCCM).

For conducting SLR studies, many frameworks have been proposed in the literature, and among them, most SLRs follow the simple and robust Search, Appraisal, Synthesis, and Analysis (SALSA) framework (Grant and Booth, 2009). In this framework, the search step consists of clearly delimitating the scope of the study, defining the search strategy, and searching the documents through search databases. The appraisal step consists of selecting relevant studies with quality assessment criteria. The synthesis step consists of extracting and categorizing the data for further analysis. The analysis step consists of analyzing the selected papers (e.g., quantitative, qualitative, or narrative analysis), and identify and discuss the most important results. In this study, we represent the first three steps of the SALSA framework through the PRISMA formalism, and we use the TCCM protocol for the analysis step. The findings from the analysis step allow us to provide answers to the four sub-research questions which provide insights for the proposal of a holistic theoretical and methodological framework for the implementation of XAI applications.

2.1. Search, appraisal, and synthesis through PRISMA

The scope of our study is defined through the research questions elaborated in the introduction. To make fully transparent the first three steps of the SALSA framework, we considered using the well-known Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) formalism (Page et al., 2021) through the “identification”, “screening” and “included” phases (Fig. 3).

For the search, we started by selecting a suitable database. We chose Scopus, which is the largest database of scholarly articles (Norris and Oppenheim, 2007), with 60% more coverage than the Web of Science (Comerio and Strozzi, 2019). To elaborate the search string, we relied on the different terms identified in previous reviews papers related to XAI (e.g., explainable AI, XAI, transparent AI, responsible AI, trustworthy AI, intelligible AI) (e.g., Adadi and Berrada, 2018; Arrieta et al., 2020). As many authors directly refer to “machine learning” techniques in the context of AI, we also used ML or machine learning in the search string. Thus, we used the search string presented in Fig. 4.

This initial search was performed on 4 November 2022 and returned 6582 documents. Because our study is restricted to business applications of XAI, we only retained documents from the corresponding “business, management and accounting” subject area in Scopus (292 documents). It is important to note that document classification in Scopus subject areas is not exclusive. For instance, the selected subject area also contains documents from other subject areas (e.g., computer science, engineering, decision sciences, and social sciences) if they are business oriented or applied research. This corresponds to our need for this paper, as we are focusing on business applications of XAI. Next, we excluded non-English documents (two documents), as usually performed in SLR studies, and we only kept journal article documents, excluding conference papers, book chapters, conference reviews, reviews, books, editorials, notes, and letters to editors. Thus, the identification (search) step ended up with 175 articles selected.

For the quality assessment in the screening (appraisal) step, we chose to rely on well-known and renowned international journal quality rankings (e.g., Donthu et al., 2021): AJG ranking (Academic Journal Ranking, previously ABS) provided by the Chartered Association of Business Schools, and ABDC ranking (Australian Business Deans Council). From the highest to the lowest journal qualities, AJG ranks journals in five classes (4 *, 4, 3, 2, 1) and ABDC ranks in four classes (A*, A, B, C). High-quality journals from AJG are ranked 4 * , 4, or 3, while high-quality journals from ABDC are ranked A* or A. Thus, from the previous 175 articles, we only kept high-quality articles ranked in high-quality journals from AJG or ABDC (97 articles): (ranked 4 * or 4, or 3 in AJG) or (ranked A* or A in ABDC). To make sure that those articles follow the requirements of our studies (XAI business application with real-world data), two AI expert researchers exclusively performed a manual checking by reading the abstract or the full content of these 99 articles. They jointly agreed to exclude 62 articles (articles focusing on defining XAI methods without concrete real-world applications and articles using simulated data). Thus, 37 articles were finally selected for the qualitative analysis of our study.

A large majority (26 articles) were published in 2022, nine articles in 2021, and one article in 2018. Most of these articles come from *Knowledge-Based Systems* (7), *Technological Forecasting and Social Change* (5), *Decision Support Systems* (4), *Journal of Cleaner Production* (3), and *Management Science* (2). Many other journals are also present with one article each: *Computers in Industry*, *Information Systems Research, International Journal of Production Economics, Journal of Marketing Research, International Journal of Production Research, Journal of the Operational Research Society, International Journal of Accounting Information Systems, Production, and Operations Management, Real Estate Economics, International Journal of Human Resource Management, IEEE Transactions on Engineering Management, Journal of Construction Engineering and Management, Journal of Retailing and Consumer Services, Industrial Management and Data Systems, Journal of Advanced Transportation, Journal of Economic Behavior and Organization*. To analyze those articles, many categorizations will be performed by using the TCCM protocol.

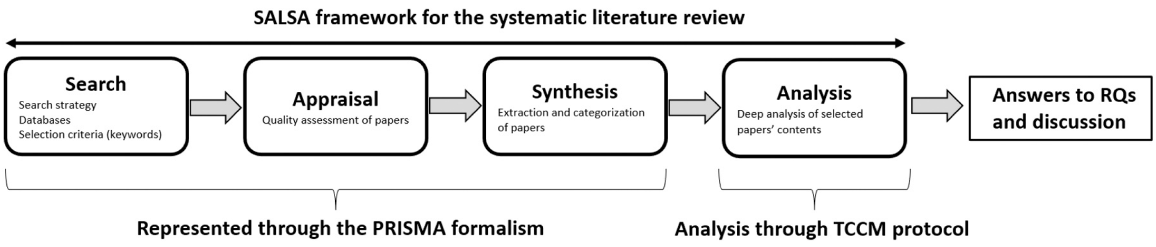

SALSA framework for the systematic literature review
Search
Search strategy
Databases
Selection criteria (keywords)
Appraisal
Quality assessment of papers
Synthesis
Extraction and categorization of papers
Analysis
Deep analysis of selected papers' contents
Answers to RQs and discussion
Represented through the PRISMA formalism
Analysis through TCCM protocol

Fig. 2. Overview of the methodology of the research.

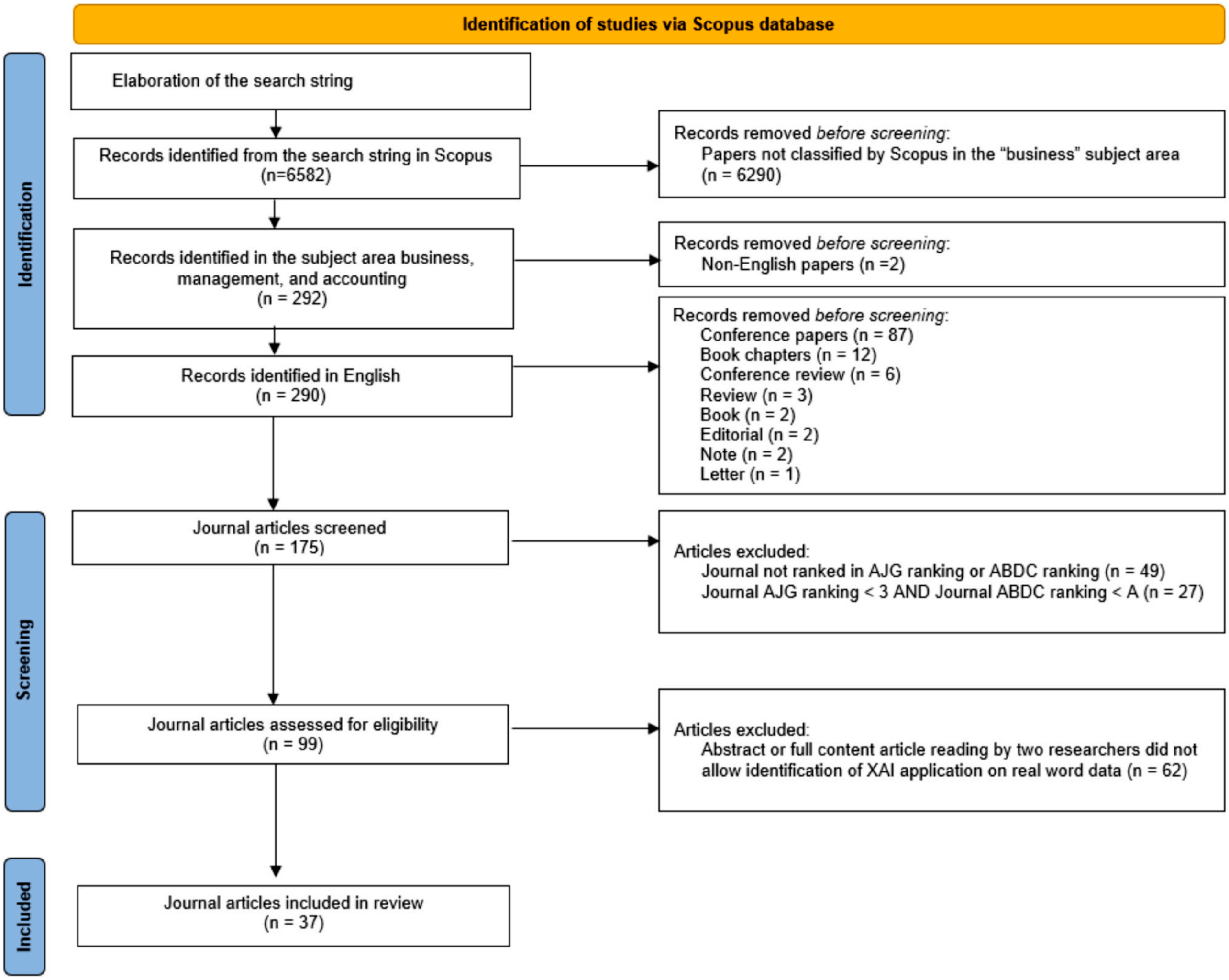
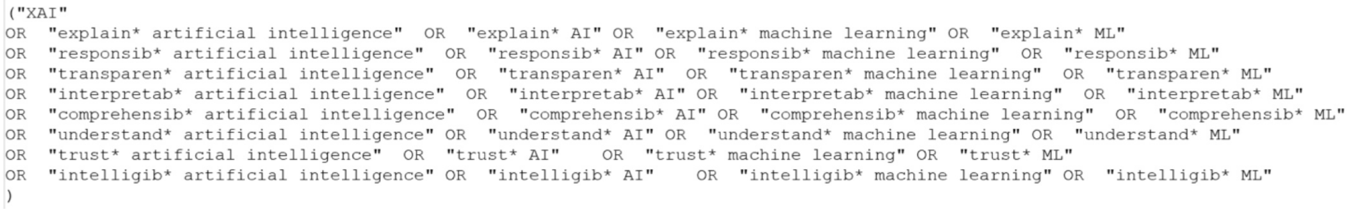

Identification of studies via Scopus database
Identification
Elaboration of the search string
Records identified from the search string in Scopus
(n=6582)
Records removed before screening:
Papers not classified by Scopus in the "business" subject area
(n = 6290)
Records identified in the subject area business, management, and accounting
(n = 292)
Records removed before screening:
Non-English papers (n =2)
Records identified in English
(n = 290)
Records removed before screening:
Conference papers (n = 87)
Book chapters (n = 12)
Conference review (n = 6)
Review (n = 3)
Book (n = 2)
Editorial (n = 2)
Note (n = 2)
Letter (n = 1)
Screening
Journal articles screened
(n = 175)
Articles excluded:
Journal not ranked in AJG ranking or ABDC ranking (n = 49)
Journal AJG ranking < 3 AND Journal ABDC ranking < A (n = 27)
Journal articles assessed for eligibility
(n = 99)
Articles excluded:
Abstract or full content article reading by two researchers did not allow identification of XAI application on real word data (n = 62)
Included
Journal articles included in review
(n = 37)

Fig. 3. PRISMA representation for the search, appraisal, and synthesis steps.



```
("XAI"
OR  "explain* artificial intelligence"  OR  "explain* AI" OR  "explain* machine learning" OR  "explain* ML"
OR  "responsib* artificial intelligence"  OR  "responsib* AI" OR  "responsib* machine learning"  OR  "responsib* ML"
OR  "transparen* artificial intelligence"  OR  "transparen* AI"  OR  "transparen* machine learning"  OR  "transparen* ML"
OR  "interpretab* artificial intelligence"  OR  "interpretab* AI" OR  "interpretab* machine learning"  OR  "interpretab* ML"
OR  "comprehensib* artificial intelligence"  OR  "comprehensib* AI" OR  "comprehensib* machine learning"  OR  "comprehensib* ML"
OR  "understand* artificial intelligence" OR  "understand* AI" OR  "understand* machine learning" OR  "understand* ML"
OR  "trust* artificial intelligence"  OR  "trust* AI"    OR  "trust* machine learning" OR  "trust* ML"
OR  "intelligib* artificial intelligence" OR  "intelligib* AI"    OR  "intelligib* machine learning" OR  "intelligib* ML"
)
```

Fig. 4. Search string used for selecting records in Scopus.

2.2. *Analysis with TCCM protocol*

For the analysis of the articles in this review, we opted to use the TCCM protocol (Paul and Rosado-Serrano, 2019) for two main reasons. First, to answer our research questions, we needed to analyze both theoretical and practical aspects of the applications of XAI in industries. In general, framework-based review protocols such as the TCCM framework shed light on both theoretical and empirical aspects of a specific research domain, thus overcoming the limitations of narrower domain-based (e.g., Arashpour, 2023; Nimmy et al., 2022), theory-based (e.g., Branstad and Solem, 2020), or method-based (Adadi and Berrada, 2018; Arrieta et al., 2020) literature reviews (Y. Chen et al., 2021). Second, framework-based review protocols such as TCCM are generally more impactful than other types of reviews (Paul et al., 2021).

In the next section, the results of the analysis of the selected articles will be presented in terms of (i) theories used through all the steps of processes building business XAI applications; (ii) contexts presenting the most common application fields or industries using XAI along with the main stakeholders; (iii) characteristics such as the most commonly used methods, type of applicative systems, data types, or scope of explanations; (iv) the most commonly used methodologies for design. These four complementary views will allow us to answer our research questions and discuss our proposed new methodological and theoretical framework for the implementation of the XAI in industries.

3. Findings

3.1. Main field theories and analytical theories used in applicative XAI studies

More globally, a scientific theory represents a well-founded and widely accepted statement, hypothesis, or explanation that has withstood rigorous testing and scrutiny (Rudner, 1968). Thus, it might be interesting to study XAI business applications from the perspective of theories. From the articles analyzed in this review, we observed that XAI business applications are often based on various theoretical frameworks and paradigms to explain relevant antecedents, parameters, or outcomes of the underlying system. We can divide the theories used in these articles into two sets: analytical-based theories (e.g., statistical learning theory, formal argumentation theory, game cooperation theory) and domain field theories (e.g., quality management theory, behavior change theories, emergency management framework theory, resource dependence theory, resource-based view theory, information diagnosticity theory, self-reference theory). In the context of this paper, we also extend theory perspectives to regulation perspectives (e.g., standards, norms) specific to some organizations or industries. In the next sub-sections, we present a short common definition of each used theory (or regulation), and how it is used in the context of XAI business application.

3.1.1. Statistical learning theory

The main goal of statistical learning theory is to provide a framework for studying the problem of inference, that is, of gaining knowledge, making predictions, making decisions, or constructing models from a set of data (Bousquet et al., 2004). This theory provides a formal definition of the most important concepts behind predictive ML algorithms, such as the building of learning functions from historical data, generalization, overfitting, and performance of predictive capacity for designing better algorithms. Most supervised machine-learning algorithms are built following this theory (Vapnik, 1999).

Given the definition of statistical learning theory, we can consider that most predictive ML models built in the context of our study rely first on this theory for building either black-box or white-box models, even if only one article (Bodendorf et al., 2021) explicitly refers to this theory. This article refers to this theory as a second artefact for building black-box ML or DL models in an XAI framework, which also includes four other artefacts: (1) features selection, (3) cost estimations, (4) model explanations, and (5) a multi-agent system.

3.1.2. Game cooperation theory

Cooperative game theory is a branch of game theory that studies how people can cooperate to achieve mutual benefits in strategic situations where the outcomes depend on the choices of multiple individuals or groups.

In a cooperative game, players can form coalitions, or groups, and work together to achieve a common goal or to divide the gains from cooperation. One of the main research questions in cooperative game theory is how to distribute the gains from cooperation among the members of a coalition fairly and efficiently. Cooperative game theory assumes that players can communicate and make binding agreements, which enables them to coordinate their actions and achieve better outcomes than they can by acting independently. It also assumes that players are rational and seek to maximize their own payoffs but may be willing to make concessions or cooperate to achieve a better outcome for the group as a whole. Examples of cooperative games include business partnerships, negotiations, and international alliances. Cooperative game theory is used in various fields, such as economics, political science, psychology, and computer science, to study situations where cooperation is essential for achieving optimal outcomes.

The underlying technical basis for SHAP is rooted in the calculation of Shapley values originating from cooperative game theory (Lundberg and Lee, 2017). Within the realm of XAI, Shapley values offer a technique to illustrate the proportional influence of every individual feature (or variable) under assessment on the final output of the ML model. This is achieved by contrasting the varying effects of inputs against their average impact. Most of the current applications of XAI in business rely on the SHAP framework through the computation of Shapley values (see Table 2).

3.1.3. Formal argumentation theory

Formal argumentation is a way of logical inference based on constructing and evaluating arguments, each of which provides reasons for a particular claim. Compared with previous formal methods for non-monotonic and common-sense reasoning, formal argumentation has the advantage that it reflects particular forms of human reasoning, which creates opportunities for XAI.

The process of formal argumentation involves a sequence of stages (Cerutti et al., 2014). Initially, facts are translated into a logical structure and stored within a knowledge repository. Subsequently, arguments emerge from the knowledge repository according to the rules of the formalism being used. Following this, attacks among the arguments are generated. Afterward, the validity of each argument is assessed, and groups of arguments that are collectively acceptable (in accordance with meaning) are pinpointed. Lastly, the deductions drawn from arguments within these acceptable groups are linked to potentially valid inferences, as per the argumentation framework (Doumbouya et al., 2018; Prakken and Vreeswijk, 2002).

For an explainable fake news detection model on social media (Twitter in particular), (Chi and Liao, 2022) proposed a Quantitative Argumentation-based Automated eXplainable Decision-making System (QA-AXDS). Their process can be categorized into three main steps: (1) building a quantitative argument tree (a dialog or conversion tree) for a tweet and all its reply tweets, (2) analyzing node contents in each link in the tree to assign a polarity (support or attack) and correlation measurements, (3) providing a reasoning machine (inference mechanism with o-QuAD algorithm) to provide a decision from the tree analysis, and (4) providing an explanation model in the form of natural language to explain the decision. It is important to note that ML models are also used in this approach, particularly in polarity (stance) detection from the text in each node of the argument tree.

3.1.4. Information diagnosticity theory

Information processing theories, especially the accessibility-diagnosticity theory by Feldman and Lynch (Feldman and Lynch, 1988), claim that the likelihood of using a piece of information for making a choice depends both on its accessibility and its diagnosticity. Therefore, this theory might suggest that consumers are more likely to use more accessible or visible online reviews to make a choice (Susan and David, 2010). More generally, (Jiang and Benbasat, 2004) defined information diagnosticity as consumers' perceptions of the ability of a website to convey relevant product information that can assist them in understanding and evaluating the quality and performance of products sold online.

In his XAI framework for explaining a prediction of the success or failure of a consumer complaint in RegTech applications, (Siering, 2022) relied on information diagnosticity theory for building relevant and understandable text features to be used as input of black-box ML models. This study emphasized under researched theory-based feature engineering by deriving features from information diagnosticity theory to explain why they influence complaint success. Thereafter, other well-established technologies for post hoc explanation could be applied.

3.1.5. Resource dependence theory

Resource Dependence Theory (Pfeffer and Salancik, 2003) explains how organizational behavior is affected by external resources. In contrast to the resource-based view, which is more internally focused, resource dependence theory focuses on the survival and improvement of an organization and focuses exclusively on complementary, externally sourced resources (Barringer and Harrison, 2000). In purchasing and supply management, the resource dependence theory focuses on collaboration between buyers and suppliers and on cooperation between supply chain partners to create mutual benefits (Paulraj and Chen, 2007).

Inspired by resource dependence theory, (Bodendorf et al., 2022) introduced a holistic cost estimation framework to collaboratively manage product costs with suppliers under information asymmetry as well as to support the business-to-business (B2B) cost and price negotiation process from a manufacturer's perspective. They relied on ML and DL models coupled with a multi-agent system. To foster acceptance for both suppliers and buyers, they used a combination of model agnostic post hoc XAI approaches.

3.1.6. Resource-based view theory

The resource-based view (RBV) argues that a firm's sustained competitive advantage is based on its valuable, rare, inimitable, and non-substitutable internal resources (Barney, 1991). The capability of firms to create or acquire these resources affects their performance and competitiveness over their competitors.

In the human resource management context, (Chowdhury et al., 2022) presented a conceptual review of AI algorithmic transparency and then discussed its significance to sustain competitive advantage by using the principles of resource-based view theory. In practice, they demonstrated the capability of the LIME technique to intuitively explain to HR managers the employee turnover predictions generated by AI-based models.

3.1.7. Behavior changes theories

Behavior change theories are psychological models that explain how people change their behavior over time. These theories provide a framework for understanding the factors that influence behavior change and the different stages that people go through as they adopt new behaviors (Zimbardo and Ebbesen, 1970). Some of the most prevalent are the Social Cognitive Theory, the Theory of Planned Behavior, and the Transtheoretical Model.

To understand shifts in consumer behavior towards organic products, (Taghikhah et al., 2021) relied on Stern's buying theory to classify decisions as planned, impulsive, and unplanned. To collect data in their experimentation, they integrated five behavioral theories in the survey design: the theory of planned behavior to account for factors driving planned decisions, the theory of interpersonal behavior to integrate the influence of emotions, impulsive buying theory to capture factors driving impulsive purchasing, alphabet theory to integrate the role of habits, and goal framing theory to account for the variety of goals. To establish a causal link between the act of buying organic wine and to choose the classification algorithm that exhibits superior accuracy, efficiency, and predictive capability, they conducted a thorough assessment of multiple supervised ML algorithms. For the explanations of the results, they mostly relied on the relative variable importance provided by the Random Forest method.

3.1.8. Emergency management theory

The emergency management framework theory describes in detail the most important components to successfully manage any kind of disaster (e.g., floods, fires, earthquakes, hurricanes, sanitary crises, epidemics). These components should include activities related to mitigation, preparedness, response, and recovery (McLoughlin, 1985).

(Johnson et al., 2021) proposed an improvement of the emergency management framework by developing an explainable AI solution that identifies opioid overdose (OD) trends and determines the significant factors for improving survival rates. The proposed AI-based solution contains three stages. The first stage creates a dataset from various open data sources, while the second phase trains three AI algorithms: RF, ANN, and SVM. The third phase utilizes SHAP to make the models more transparent and to generate insights into the significant factors affecting OD survival rates. This explainable AI solution shows an example of how AI can particularly improve the mitigation and preparation phases.

3.1.9. Quality management theory

Quality management theory is a set of principles and practices aimed at improving the quality of products, services, and processes within an organization. It involves a systematic approach to identifying and addressing potential quality problems as well as implementing processes and procedures to prevent future quality issues. To that end, quality management theory suggests identifying and eliminating sources of quality variation (Schmenner and Swink, 1998; Taguchi, 1986).

(Senoner et al., 2022) proposes a decision-making model driven by data to enhance the quality of manufacturing processes. Their model for decision-making was bifurcated into a pair of stages: the initial one focused on arranging processes in order of importance for quality enhancement, followed by the subsequent selection of appropriate strategies for improvement. They modified the SHAP method, adapting it to the realm of quality management. Thus, they introduced an original gauge for the importance of processes, grounded in quality management theory. Their gauge for process importance estimates the degree to which production parameters of a given process contribute to fluctuations in overall process quality. This supports the effective allocation of improvement efforts.

3.1.10. Self-reference theory

The self-reference theory or self-reference effect refers to people's tendency to better remember information when that information has been linked to the self than when it has not been linked to the self. In marketing for instance, according to self-reference theory, consumers' ability to relate the brand more easily to their own personal experiences in brand selfies can generate higher levels of cognitive elaboration and mental simulation of brand consumption (Escalas, 2007), which has been linked to higher levels of brand engagement (Elder and Krishna, 2012).

(Hartmann et al., 2021) rely on self-reference theory to collect different types of consumers-selfies images from social media, to classify social media brand imagery and explain user response (purchase intention).

3.1.11. Regulations, standards, or norms

A regulation standard is a set of rules, guidelines, or requirements established by a regulatory agency or governing body to ensure that products, services, or processes meet certain safety, quality, or performance criteria. These standards are often legally enforceable and may apply to a wide range of industries and activities, including manufacturing, healthcare, finance, transportation, and more.

Given the increasing use of AI models in businesses, many regulation standards are adapting to avoid the biases that these models may suffer from, especially when they are not transparent. As an illustration, existing guidelines governing audit documentation and audit evidence (e.g., PCAOB AS 1105; AS 1215) suggest that if auditors are unable to elucidate and document the internal mechanisms or outcomes of an AI model, their ability to rely on such tools is constrained (AICPA, 2020; CPAB, 2021). In this context, (Zhang et al., 2022) explored how diverse XAI approaches can align with the prerequisites set by audit documentation and evidence standards. They showcased the significance of XAI methods, particularly LIME and SHAP, employing an audit task involving the evaluation of material misstatement risk. In another context, (Bücker et al., 2022) motivate their proposed framework for transparency, auditability and explainability of credit scoring models, by referring to new requirements of the European Banking Authority and the European commission (European Banking Authority, 2020; European Commission, 2016).

3.2. Main application contexts of XAI in companies

In this study, we assimilated the context to the application field of each article. Fig. 5 shows the number of articles per application field among our selected set of articles that apply XAI tools to explain the reasons behind the decisions made by the AI models. Details about each article in each application field are provided in Table 2. This classification per application field was performed manually (after reading all the articles), and some articles were classified in more than one application field. We found twelve main application fields, which are described in the following sub-sections.

3.3. Characteristics of the most frequently used XAI methods

For the characteristics, we evaluated the use of white-box and post-hoc explanations, and other characteristics such as the scope of provided explanations or data types.

3.3.1. White-box vs post hoc explanations

Among the studied articles, a large majority (76%) rely on post hoc explanations of black-box models, while 24% rely on white-box models, which are de facto explainable (Fig. 6).

a. **Explanations with white-box models**

Because many ML algorithms by design provide explainable outputs (e.g., coefficient of linear regressions, decision trees, if-then rules), some authors directly use them when providing acceptable predictive capacity in their business context. This is, for instance, the case of (Gue et al., 2022), who rely on the rough set theory (Pawlak, 1982) with an RSML (Rough Set-based Machine Learning) model for predicting cities' waste management performance with an accuracy of up to 91%. The RSML method can produce if-then rules for explaining the results. Similarly, (Svoboda and Minner, 2022) rely on rules from decision trees for providing a multidimensional classification inventory scheme for supply chains. Rules from decision trees are also used by (Irarrázaval et al., 2021) for telecom traffic pumping fraud detection. It is also the case of (Andini et al., 2018), who rely on explainable decision trees for deciding whether a citizen should be eligible for a tax rebate in Italy.

Multiple white-box models have also been combined in some studies. For example, (Naumets and Lu, 2021) trained an M5P model (decision trees with linear regressions on leaf nodes) for predicting the labor cost of steel fabrication or the compressive strength of concrete in curing. Even though they trained many other more efficient black-box models, the M5P model was preferred by experts in the field for its transparency and its facility to be interpreted by humans.

In some contexts, white-box models are preferred to black-box models, even though black-box models trained for the same issue provide better predictive capacities. This is the case of (Pessach et al.,

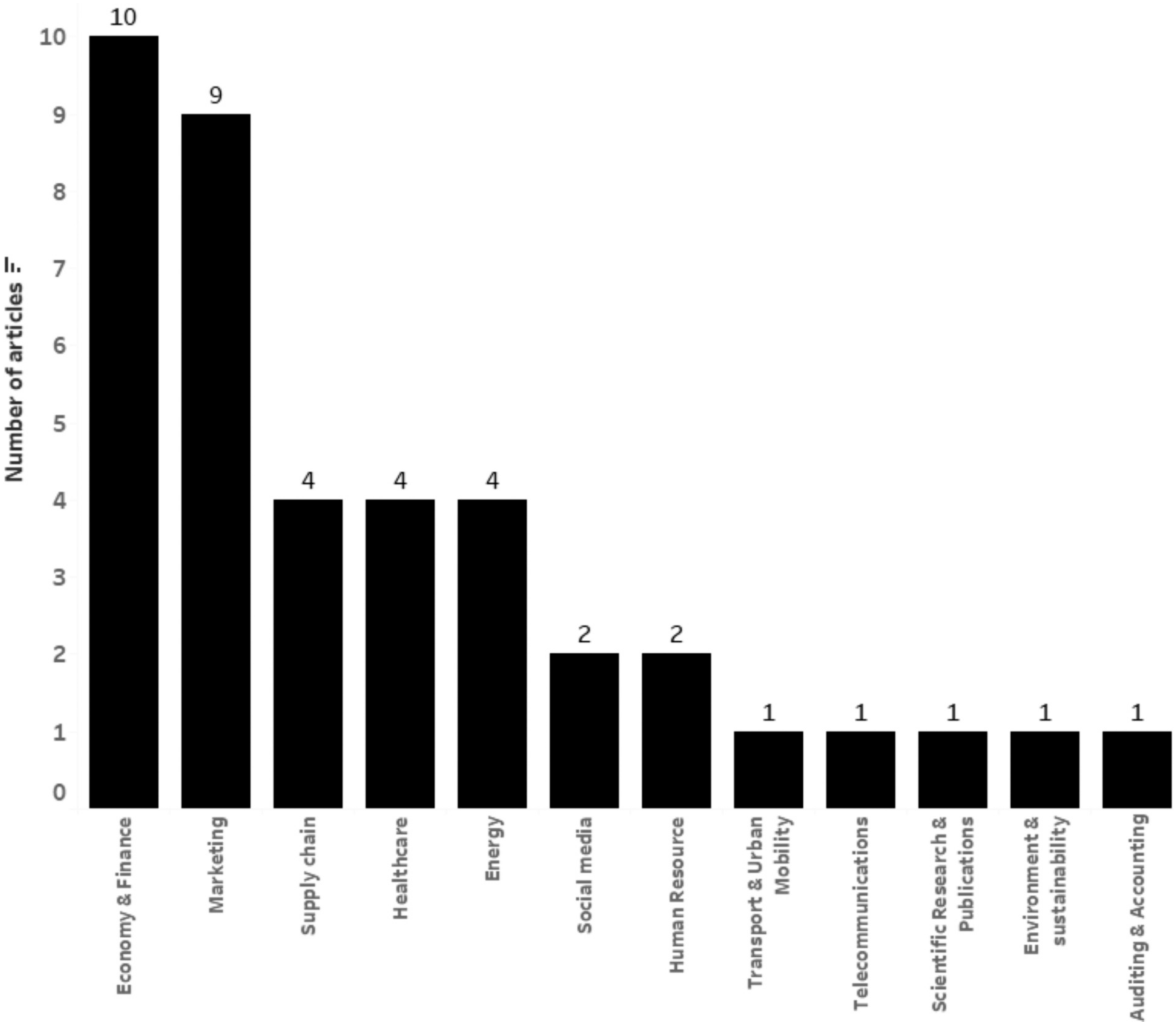


Fig. 5. Number of articles per application field.

Table 2
Articles per application field.

Application field	Articles	Respective predictive goals
Economy and finance	(Bodendorf et al., 2022; Bücker et al., 2022; C. Chen et al., 2022;Ghosh et al., 2022;Jana et al., 2022; J.Kim et al., 2022;Lorenz et al., 2022; Park and Yang 2022;Siering, 2022;Vo et al., 2021)	Predicting material costs or manufacturing costs; Credit scoring; Predicting financial lending; Predicting futures prices of stocks listed on the National Stock Exchange (NSE) in India; Predicting and controlling energy consumption and electronic waste generation in bitcoin mining; Predicting the economic value of technologies; Real estate rental prediction; Predicting economic growth rates and crises of countries; Predicting successful consumer complaints in financial services; Customer churn prediction in financial services
Marketing	(Hartmann et al., 2021; Liu et al., 2021; Park et al., 2022; Siering, 2022; Taghikhah et al., 2021; Vo et al., 2021; L. Wang et al., 2022; T. Wang et al., 2022; Zanon et al., 2022)	Brand images classification; Predicting the Click Through Rate (CTR) of a search engine; Predicting customers' choices from online reviews; Predicting successful consumer complaints; Predicting purchasing decisions; Customer churn prediction; Predicting venue popularity on location-based services; Predicting malls' customer traffic; Movie and music recommendations
Energy	(Boulmaiz et al., 2022; Jana et al., 2022; Tsoka et al., 2022; Zhao et al., 2021)	Recommending action plans to improve the energy comfort in buildings; Predicting and controlling energy consumption and electronic waste generation in bitcoin mining; Classifying a building's Energy Performance Certificate (EPC) label; Predicting hydrogen production of biomass
Healthcare	(Ganeshkumar et al., 2021; Gozzi et al., 2022; Johnson et al., 2021; Raza et al., 2022)	Multi-label classification of ECG (Electrocardiography) signals; Electromyography (EMG) hand gesture classification; Predicting survival rates from the consumption of opioid OD drugs; Classifying different arrhythmias from electrocardiography (ECG) signals
Supply chain	(Bodendorf et al., 2022; Naumets and Lu, 2021; Senoner et al., 2022; Svoboda and Minner, 2022)	Predicting material cost or manufacturing costs in a supply chain; Predicting the labor cost of steel fabrication; Predicting process quality in semiconductor manufacturing); Inventory classification in supply chains
Environment and sustainability	(Gue et al., 2022; Jana et al., 2022; Lee et al., 2022)	Predicting waste management system performance from city and country attributes; Predicting and controlling energy consumption and electronic waste generation in bitcoin mining; Predicting harmful algal blooms – concentration of chlorophyll in water
Social media	(Chi and Liao, 2022; Hartmann et al., 2021)	Fake news detection from social media; Brand image classification from social media
Human resources	(Andini et al., 2018; Chowdhury et al., 2022)	Predicting the successful placement of a candidate in a specific position; Predicting employee turnover
Auditing and accounting	(Zhang et al., 2022)	Predicting risk of material misstatement assessment for auditing tasks
Transport and mobility	(E.-J.Kim, 2021)	Predicting travel mode choices
Telecommunication	(Irarrázaval et al., 2021)	Traffic pumping fraud detection
Scientific research and publications	(Ha, 2022)	Predicting the number of citations of a scientific paper

2020), who used an explainable Variable-Order Bayesian Network (VOBN) for predicting the successful placement of a candidate in a specific position at a pre-hire stage. This model was preferred by experts, even though a black-box gradient boosting machine model provided better predictive metrics.

White-box models are also used for more complex problems. Some authors propose new hybrid models that can be partially explainable by design for some specific interesting features, and non-explainable for other features. For instance, (Wang et al., 2022) developed an innovative interpretable Generalized Additive Neural Network Model (GANNM). This model consists of two parts: an interpretable component (Generalized Additive Model, a generalized linear model) and a black-box component (Neural Network). Both components can be trained together as a single component for predicting malls' customer traffic, given a set of features related to marketing campaigns and budget allocations. The authors showed the relevance of their new method, along with explanations, compared to the use of post hoc explanations on many black-box models. In a different context, for predicting financial lending decisions, (C. Chen et al., 2022) combined in a transparent way two-layer additive risk model, akin to a dual-layer neural network but disassembled into distinct subscales. Within this framework, each node within the initial (hidden) layer embodies a significant subscale model, with the non-linear elements being clear and comprehensible. In addition, they proposed a visualization tool for exploring this model to easily justify a prediction.

Finally, some white-box models or processes are used to provide explainable results for recommender systems. This is the case of (Liu et al., 2021), who relied on explainable topic modelling (through the Poisson factorization method) and explainable recommendation (through the Bayesian inference method) for predicting search volumes and click-through rates (CTR) in a search engine.

b. **Explanations with black-box models**

The distribution of post hoc explanation methods on black-box models is presented in Fig. 7. For the description of each method in this figure, please refer to Table 3. These methods are also the most represented in the taxonomies developed in several reviews studies on XAI (e.g., Adadi and Berrada, 2018; Arrieta et al., 2020; Das and Rad, 2020; Minh et al., 2022).

In the studied articles, these explanation methods are used over common ML black-box models, such as DL models (e.g., LSTM, CNN), Artificial Neural Networks, XGBoost, Gradient Boosting, Random Forest, or Support Vector Machines.

The most popular explanation framework in the studies is the SHAP Framework (16 articles, 44%). Many of those studies use SHAP as the single explanation framework (e.g., Johnson et al., 2021; Lee et al., 2022; Park and Yang 2022) while others compare or complement it with

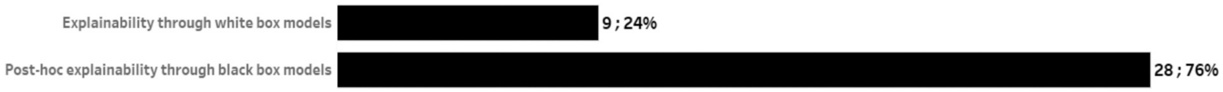


Fig. 6. Distribution between white-box and post hoc explanations in the study.

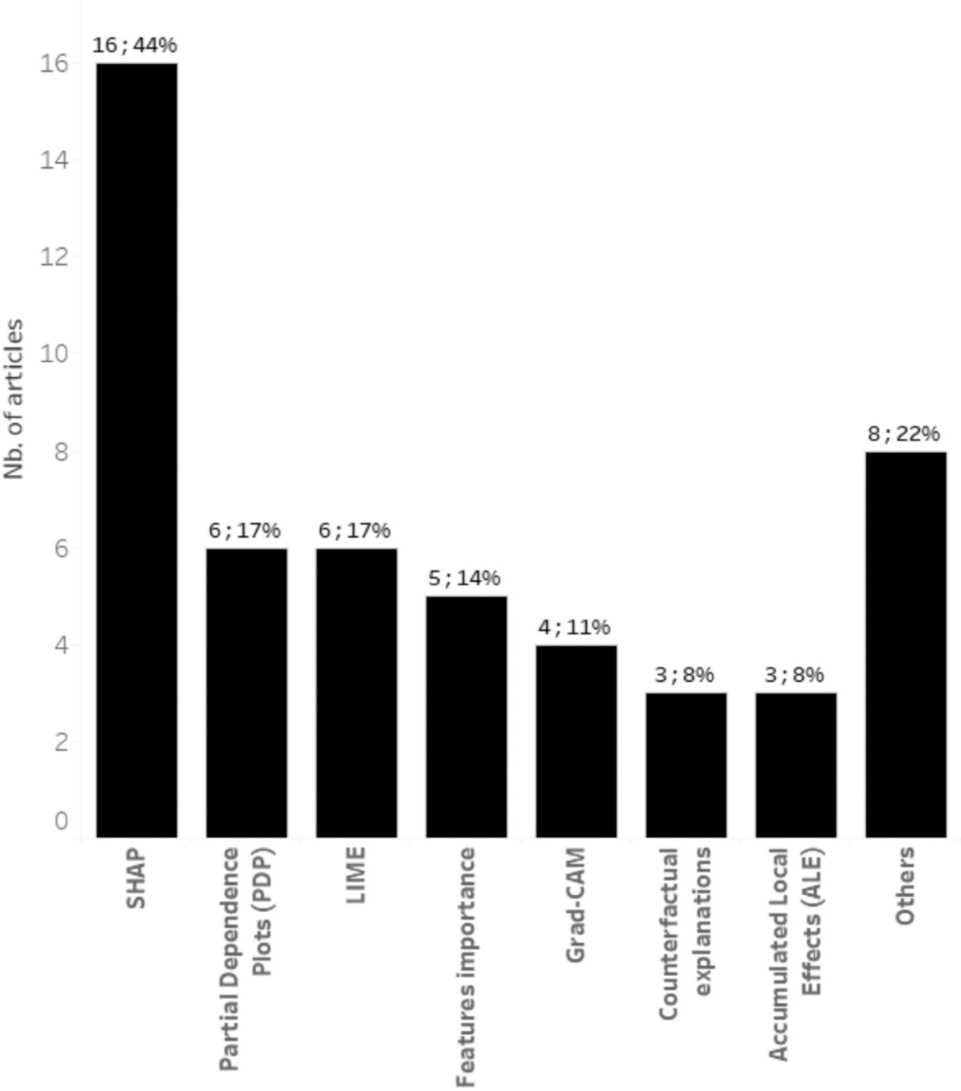


Fig. 7. Distribution between post hoc explanation methods in the study.

other explanations methods, such as LIME, PDP, ALE, Grad-CAM, features importance (e.g., Bücker et al., 2022; Ghosh et al., 2022; Gozzi et al., 2022).

With six articles (17%) each, the second most popular explanation methods are LIME (e.g., Chowdhury et al., 2022; Tsoka et al., 2022; Zhang et al., 2022) and Partial Dependence Plots (PDP) (e.g., Bodendorf et al., 2022; Bücker et al., 2022; Zhao et al., 2021).

The third most popular method (five articles, 14%) relies on several studies to compute variable importance from feature permutation methods. This could be feature importance embedded in some tree-based algorithms, such as XGBoost (Lorenz et al., 2022) or Random Forest (Taghikhah et al., 2021). Other studies rely on feature importance computation method derived from (Friedman, 2001) (see Zhao et al., 2021) or from (Fisher et al., 2019) (see E.-J. Kim, 2021; Lorenz et al., 2022; Zhang et al., 2022).

The fourth most popular method is Grad-CAM (Gradient-weighted Class Activation Mapping) (four articles, 11%). Grad-CAM is particularly designed for visual explanations of image classification, such as electrocardiography signals (ECG) (Ganeshkumar et al., 2021; Raza et al., 2022), residual neuromuscular activity signals (EMG) (Gozzi et al., 2022), or brand images from social media (Hartmann et al., 2021). For instance, (Gozzi et al., 2022) used Grad-CAM for visual explanations of the classification of EMG hand movement to improve forearm electrode placement and configuration. They also relied on SHAP explanations, but to perform feature selection and reduce the number of features of the model without dropping the classification metric. In another context, (Hartmann et al., 2021) also used Grad-CAM to explain brand image classification (brand selfie vs consumer selfie vs packshot) from social media. To explain purchase intention in the same context from image-associated textual data, they relied on LIME.

With three articles (8%) each, the fifth most popular methods are Accumulated Local Effects (ALE) (Andini et al., 2018; Lorenz et al., 2022; Zhang et al., 2022) and counterfactual explanations (Bodendorf et al., 2022; Boulmaiz et al., 2022; Zhang et al., 2022). In those studies,

Table 3
Most common post hoc explainable methods in the selected studies.

Methods	Description	Advantages	Disadvantages	Studies
SHAP	SHAP is based on Shapley values derived from game theory, for providing local or global explanations (features importance) of any ML model (Lundberg and Lee, 2017; Molnar 2020). The Shapley value is defined as the marginal contribution of variable value to prediction across all conceivable "coalitions" or subsets of features (Lundberg and Lee, 2017).	• Allows contrastive explanations • Has solid backup theory • Can deliver a full explanation that is fairly distributed among the feature values	• Requires a large amount of computing power	(Ha, 2022;Johnson et al., 2021; J.Kim et al., 2022;Lee et al., 2022;Park et al., 2022; Park and Yang 2022;Senoner et al., 2022; L.Wang et al., 2022; Bodendorf et al., 2022;Bücker et al., 2022;Ghosh et al., 2022;Gozzi et al., 2022;Jana et al., 2022;Tsoka et al., 2022;Vo et al., 2021;Zhang et al., 2022).
LIME	LIME uses a local surrogate model that can interpret individual predictions (local interpretation) (Ribeiro, Singh, and Guestrin, 2016). Local surrogate models are interpretable models (e.g., linear regression, decision trees) that are used to explain individual predictions of black-box ML models (Molnar 2020).	• Has local fidelity (i.e., can be used to interpret individual predictions). • Explanations are human friendly	• The choice of surrogate model is subjective • The definition of the neighborhood of an instance is vague • Results subject to the choice of neighbors	(Chowdhury et al., 2022; Ghosh et al., 2022; Hartmann et al., 2021; Jana et al., 2022; Tsoka et al., 2022; Zhang et al., 2022)
Partial Dependence Plots (PDP)	PDP shows the global marginal effect of one or two features on the predicted outcome of a machine-learning model (Friedman, 2001)	• Intuitive • Easy to implement	• The maximum number of features in a PDP is two, due to human perception limitation • The assumption of independence between features is always violated in practice	(Bodendorf et al., 2022; Bücker et al., 2022; E.-J.Kim, 2021;Lorenz et al., 2022;Zhang et al., 2022;Zhao et al., 2021)
Permutation Feature Importance (PFI)	PFI is a technique to assess the global importance of features (variables or predictors) on the predicted value. The basic idea behind PFI is to shuffle the values of a single feature in the dataset and measure the resulting impact on the model's performance. Several approaches exist; they can be model specific (e.g., for random forest) (Breiman, 2001) or model agnostic (Fisher, Rudin, and Dominici, 2019; Friedman, 2001).	• Easy of interpretation • Does not require retraining the model	• Prone to the randomness added to the feature value permutation process • The feature value permutation process may generate unrealistic data points	(Taghikhah et al., 2021; Zhao et al., 2021; E.-J.Kim, 2021;Lorenz et al., 2022;Zhang et al., 2022)
Grad-CAM (Gradient-weighted Class Activation Mapping)	Grad-CAM is an explainable technique to visualize and understand the regions of an input image that are important for the local predictions made by a large class of Convolutional Neural Networks (CNN) (Selvaraju et al., 2017). This helps to explain why the model makes certain predictions by highlighting the regions that contribute most to the decision.	• Ease of interpretation with visualization • Possibility of assessing counterfactual explanations	• Does not provide measures of uncertainty or confidence in its visualizations. • Manipulating or perturbing input images in subtle ways can mislead the visualisation and potentially generate misleading interpretations of the model's behavior.	(Ganeshkumar et al., 2021; Raza et al., 2022; Gozzi et al., 2022; Hartmann et al., 2021)
Accumulated Local Effects (ALE) Plot	Accumulated local effects describe how features globally influence the prediction of a machine learning model on average (Apley and Zhu 2020; Molnar 2020). ALE plots are a faster and more unbiased alternative to partial dependence plots (PDPs).	• Less prone to correlated features compared to PDP • Faster computation time than PDP	• Less intuitive to understand compared to PDP • The plot can be unstable when the number of intervals increases	(Andini et al., 2018; Lorenz et al., 2022; Zhang et al., 2022)
Counterfactual explanations	Counterfactual explanations are a type of interpretability technique for explaining the local predictions of a model by providing alternative scenarios or counterfactuals. They help answer questions like, "What would have happened if a certain input feature had a different value?" (Molnar 2020; Wachter, Mittelstadt, and Russell 2017)	• Can be used to assess the robustness or sensitivity of a model • Can help provide actionable insights into the causes and effects	• May require a large amount of computing power to find suitable counterfactual instances • The counterfactual instances may not fully capture the complexity of a real-world system • Does not always provide measures of uncertainty or confidence in the results	(Bodendorf et al., 2022; Boulmaiz et al., 2022; Zhang et al., 2022)

ALE are always used for improving PDP analyses. For counterfactual explanations, different approaches are used. For instance, in the context of supplier-buyer cost negotiations, (Bodendorf et al., 2022) generated several “what-if” questions from their model (e.g., what happens if we want the supplier to decrease the total cost by 6%), and infer corresponding actions to be done, by using custom genetic algorithms. Other techniques for counterfactual explanations rely on two steps: (i) for the instance to explain, find its most similar instances measured by a chosen distance metric, but that has an opposite prediction; and (ii) compare the instance to be explained with its counterpart to spot differences, which are the counterfactual explanations. This approach is used by (Zhang et al., 2022) using Google’s What-If Tool (WIT).[1] This approach is also used by (Boulmaiz et al., 2022) but with a proposed custom method.

Finally, eight other post hoc explainable methods (22%) were identified in the study (each in a single article). These methods can be divided into two groups. First, well-known existing methods were applied in a specific context: global surrogate model, Individual

[1] https://pair-code.github.io/what-if-tool/

Conditional Expectation (ICE), and Scoped Rules. A global surrogate model is a simplified and interpretable model (e.g., linear regression) to approximate a built complex black-box model (e.g., artificial neural network). ICE is like PDP in that it plots the relationship between the predicted outcome and a feature of interest. However, ICE plots are used for explanation at the local (instance) level, whereas PDP plots are used at the global level (Goldstein et al., 2015). Scoped rules explain the behavior of a complex black-box model with high-precision rules called anchors (Ribeiro et al., 2018). In terms of output, scoped rules can be considered a variant of LIME using rules. These three techniques are experimented with by (Zhang et al., 2022) in the auditing context.

Second, there are new custom-specific methods proposed and experimented with by some authors: SHAP-MRMR, combining case-based reasoning and Bayesian network, knowledge graph and semantic profile, quantitative argument trees, and interpretable features. For instance, for explaining customer churn predictions, (Vo et al., 2021) proposed SHAP-MRMR, a combination of SHAP and the Minimum Redundancy Maximum Relevance (MRMR) method. MRMR is considered more powerful than maximum relevance feature selection and can select features that are mutually far away from each other but still have a high correlation to the classification variables. To recommend actions to improve the comfort of building occupants, (Boulmaiz et al., 2022) proposed a new approach to explanations that merges case-based reasoning with Bayesian networks, offering three types of explanation: features-based explanations, counterfactual explanations, and causal-based explanations. For another recommender system for movies and music recommendations, (Zanon et al., 2022) relied on a new method using linked open data knowledge graphs and semantic profiles to explain the output of a black-box collaborative filtering algorithm. Another new approach relying on formal logic with quantitative argumentation trees was proposed by (Chi and Liao, 2022) for explaining fake news detection on social media. Rather than explaining the outputs of black-box models, other authors chose to focus on building interpretable and transparent features from unstructured textual data and theory-based feature engineering (Siering, 2022).

3.3.2. *Other characteristics*

Fig. 8 shows other characteristics with the repartition of studied articles by data types, type of predictive systems, and scope of explanations.

Concerning data types, most of articles relied on explaining models from structured tabular data (30 articles, 83%). Only five articles used textual data (14%) (e.g., text of users' complaints in social media, online reviews) (e.g., Siering, 2022; Park et al., 2022). Finally, only four articles used image data (11%) (e.g., electrocardiogram images) (e.g., Ganeshkumar et al., 2021; Raza et al., 2022).

Concerning the types of predictive systems, most of the use of XAI is related to supervised ML, with either classifications (57%) or regressions (35%). Beyond the explanation of traditional classification or regression tasks, we can observe that three articles (8%) focus particularly on explaining the output of recommender systems (Boulmaiz et al., 2022; Liu et al., 2021; Zanon et al., 2022). A recommender system is a type of algorithmic system that predicts or recommends items or products to users based on their specific needs, past behaviors, preferences, or

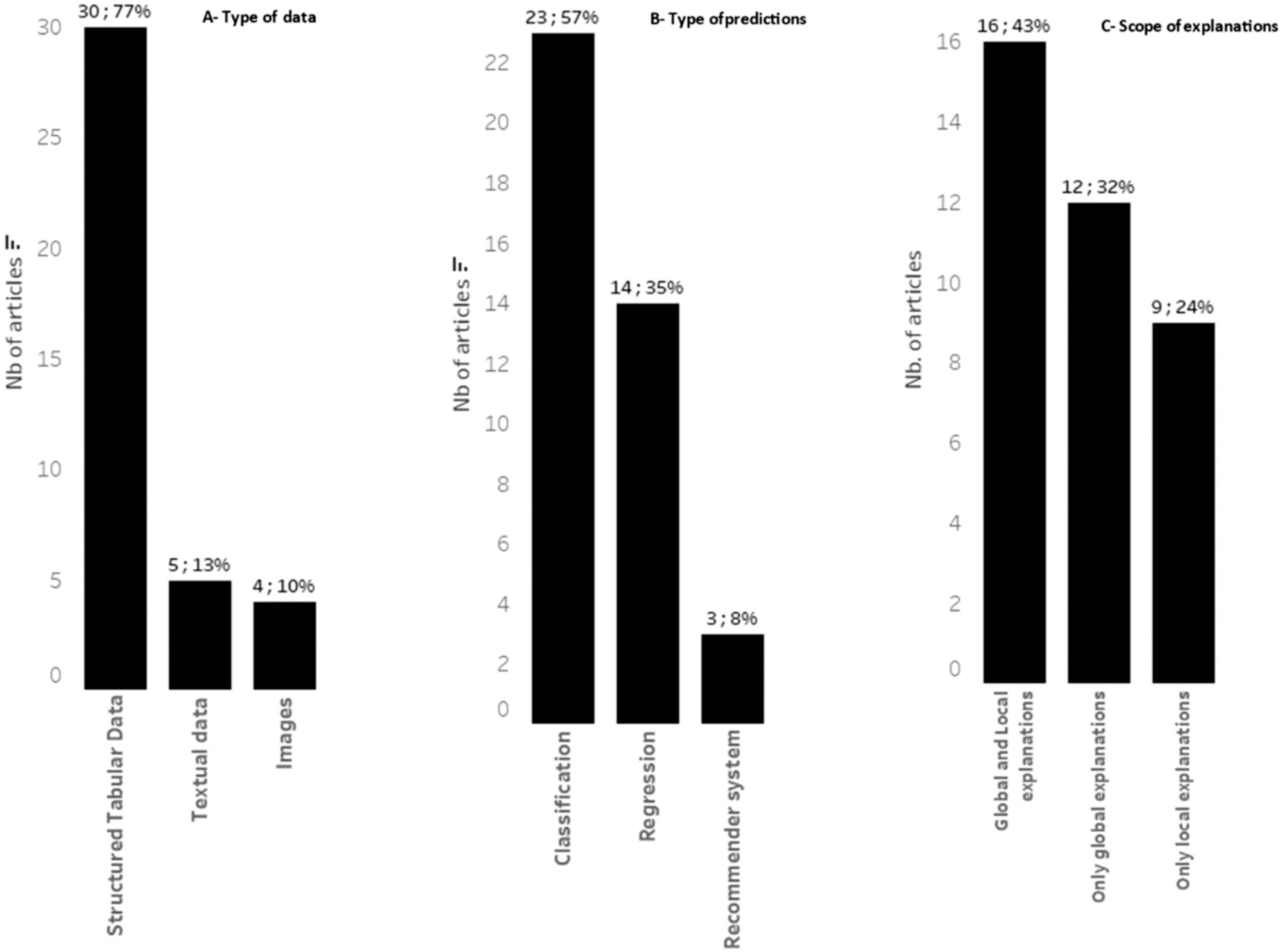


Fig. 8. Repartition by type of data (A), type of predictions (B) and scope of explanations (C).

interests. Building a recommender system can embed several techniques, including ML and other complex heuristics that usually make them very difficult to explain (Vultureanu-Albişi and Bădică, 2021). To tackle this issue, (Zanon et al., 2022) introduced a multi-domain item reordering system, for instance, in the context of movie and music recommendations. They used a traditional recommendation engine, improved by a reordering algorithm using a knowledge graph (semantic profile extraction), for reordering recommendations that are better explainable. (Liu et al., 2021) developed transparent steps for a flexible content-based search method that links the content preferences of search engine users to query search volume and click-through rates, while allowing content preferences to vary systematically based on the context of the search. (Boulmaiz et al., 2022) proposed three algorithms to explain the reasoning process behind a recommender system designed to recommend action plans to households for improving energy saving and energy efficiency in buildings.

Concerning the scope of explanations, most of the studies provided both global and local explanations (16 articles, 43%). Only global explanations were provided by 12 articles (33%), and 9 articles (24%) provided only local explanations.

3.4. Common methodologies used in applicative XAI studies

All the studied articles share a common methodological structure. This common structure can be represented by five major steps: (1) the definition of the business question and the target variable to be explained through ML models; (2) data collection: The data used can be structured, unstructured, or a combination of both; (3) data preprocessing: When necessary, data cleaning and feature engineering are used to derive more interpretable features to be used as input for ML algorithms. (4) selection and execution of one or more ML models. When multiple models are executed, the most performant models in terms of predictive capacity (e.g., using metrics such as precision, recall, MAE, R2) are commonly used in the next step for explanations; (5) generation of explanations of the outputs of the model: global explanations, local explanations, or both. This can be done in two ways, depending on the chosen models from the previous step. If the model is a white-box model, explanations are inherent in the model's outputs (e.g., regression coefficients, decision tree). Sometimes, multiple white-box models' explanations are combined (e.g., Naumets and Lu, 2021). If the model is a black-box model, one or many explainable post hoc methods are used (e.g., SHAP, LIME, PDP). Sometimes, new explainable methods are proposed for some specific context (e.g., SHAP-MRMR, GANMN, post hoc explanations of recommender systems).

Most of the authors end their study after step 5 to show that explanations can be useful in their studied field. However, some authors go beyond this step, and provide extra steps to validate the robustness and the relevance of the provided explanations (e.g., by domain experts) (e.g., Chi and Liao, 2022; Hartmann et al., 2021).

4. Discussion

4.1. Proposal of a framework

Considering all the dimensions analyzed in the articles of this literature review, we propose a methodological and theoretical framework for XAI (Fig. 9). In order to make transparent and explainable any decision-making process based on ML models, this framework is divided into six main steps for explaining the whole analytical process: business question importance, data collection, feature engineering, ML modelling and evaluation, models outputs (predictions), robustness checking and validation of explanations by relevant stakeholders. Unlike many existing works that focus mainly on the explanations of models' predictions, this framework can be distinguished at three main levels: (i) explicability is present at all stages of the process (transparency of the process); (ii) an additional stage of evaluation of the robustness of the explanations as well as their validation, rarely addressed in the literature, is explicitly added to the process; (iii) a theoretical support of the different stages is highlighted and provides a fundamental and extensible basis for better explanations.

The objective of the framework is to provide guidance for future studies for a better handling of XAI in business applications. For a better understanding of the current literature, Table 4 shows our representation of the reviewed studies in the light of this framework. The column numbers of this table represent the different steps described in the proposed framework (Fig. 9). Three main observations can be derived from this representation: (i) the current literature mostly address explanations for the modelling step (4a and 4b) and for the prediction step (5); (ii) only few studies deeply address explanations for the two three steps (business question importance, data collection); (ii) The third step

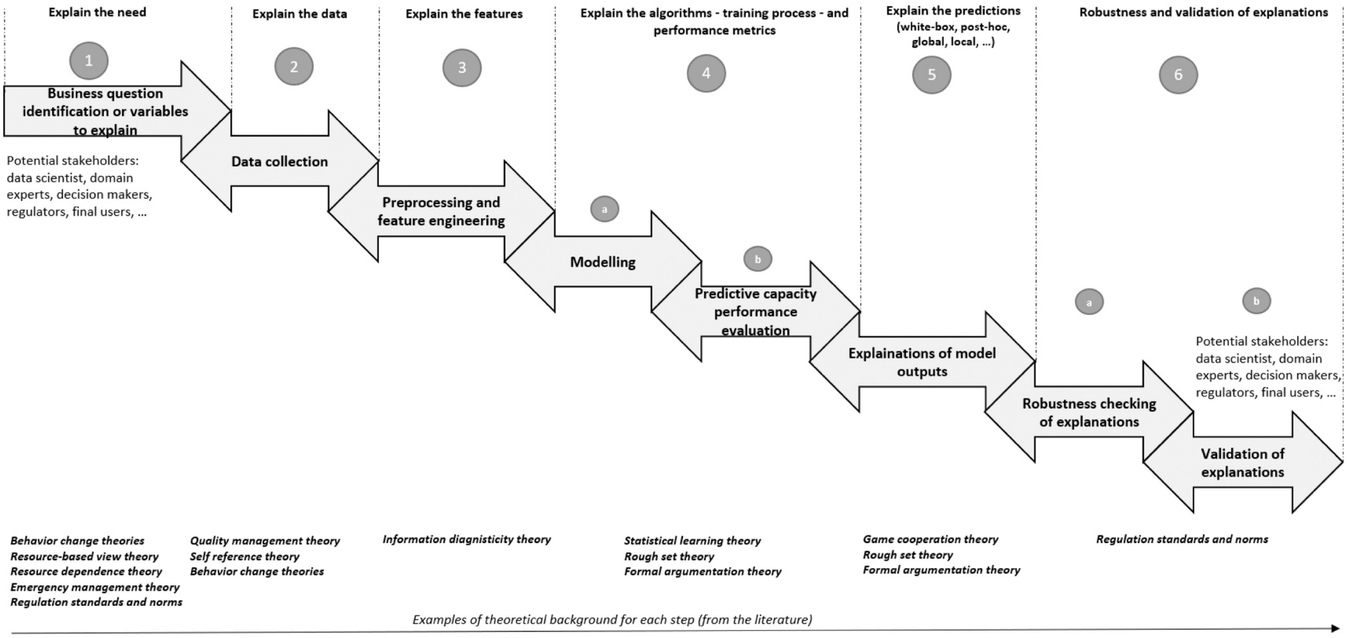

Explain the need
Explain the data
Explain the features
Explain the algorithms - training process - and performance metrics
Explain the predictions (white-box, post-hoc, global, local, ...)
Robustness and validation of explanations
1
2
3
4
5
6
Business question identification or variables to explain
Potential stakeholders: data scientist, domain experts, decision makers, regulators, final users, ...
Data collection
Preprocessing and feature engineering
a
Modelling
b
Predictive capacity performance evaluation
Explainations of model outputs
a
b
Robustness checking of explanations
Validation of explanations
Potential stakeholders: data scientist, domain experts, decision makers, regulators, final users, ...
Behavior change theories
Resource-based view theory
Resource dependence theory
Emergency management theory
Regulation standards and norms
Quality management theory
Self reference theory
Behavior change theories
Information diagnisticity theory
Statistical learning theory
Rough set theory
Formal argumentation theory
Game cooperation theory
Rough set theory
Formal argumentation theory
Regulation standards and norms
Examples of theoretical background for each step (from the literature)

Fig. 9. Methodological and theoretical framework for ML model explanations.

Table 4
Representation of studies into the proposed framework.

Article/Steps in the proposed framework	1	2	3	4a	4b	5	6a	6b
(Lee et al., 2022)	✓	✓	✓	✓	✓	✓✓		
(Ha, 2022)	✓	✓✓	✓	✓	✓	✓✓		
(J.Kim et al., 2022)	✓	✓	✓	✓	✓✓	✓✓		
(Zanon et al., 2022)	✓	✓	✓	✓✓	✓✓	✓		
(Gue et al., 2022)	✓	✓	✓	✓✓	✓	✓✓		
(Boulmaiz et al., 2022)	✓	✓	✓	✓✓	✓	✓✓		
(Senoner et al., 2022)	✓✓	✓✓	✓	✓	✓	✓✓	✓✓	✓
(Ghosh et al., 2022)	✓	✓	✓✓	✓✓	✓✓	✓✓		
(Park et al., 2022)	✓	✓	✓	✓	✓	✓✓		
(Park and Yang 2022)	✓	✓	✓	✓✓	✓	✓✓		
(Zhang et al., 2022)	✓✓	✓	✓	✓	✓	✓✓		
(L.Wang et al., 2022)	✓	✓	✓	✓	✓✓	✓✓		
(Siering, 2022)	✓	✓✓	✓✓	✓✓	✓✓	✓		
(T.Wang et al., 2022)	✓	✓	✓	✓✓	✓✓	✓✓	✓	✓
(Tsoka et al., 2022)	✓	✓	✓	✓	✓✓	✓✓		
(Jana et al., 2022)	✓	✓	✓	✓✓	✓✓	✓✓		
(Chi and Liao, 2022)	✓	✓	✓	✓✓	✓✓	✓✓	✓	
(Gozzi et al., 2022)	✓	✓	✓	✓✓	✓✓	✓✓		✓
(Bodendorf et al., 2022)	✓✓	✓	✓	✓✓	✓✓	✓✓		
(Raza et al., 2022)	✓	✓	✓	✓	✓✓	✓		
(Lorenz et al., 2022)	✓	✓	✓	✓	✓	✓✓		
(Chowdhury et al., 2022)	✓✓	✓	✓	✓	✓	✓		
(C.Chen et al., 2022)	✓	✓	✓	✓✓	✓✓	✓✓		
(Svoboda and Minner, 2022)	✓	✓	✓	✓✓	✓✓	✓✓		
(Bücker et al., 2022)	✓✓	✓	✓	✓✓	✓✓	✓✓		
(Hartmann et al., 2021)	✓	✓✓	✓	✓	✓	✓✓		✓
(Irarrázaval et al., 2021)	✓	✓	✓	✓✓	✓	✓		✓
(Liu et al., 2021)	✓	✓	✓	✓✓	✓✓	✓✓		
(Zhao et al., 2021)	✓	✓	✓	✓	✓✓	✓		
(Naumets and Lu, 2021)	✓	✓	✓	✓✓	✓✓	✓✓		✓
(Taghikhah et al., 2021)	✓✓	✓✓	✓	✓	✓	✓		
(Vo et al., 2021)	✓	✓	✓	✓✓	✓✓	✓✓		
(Johnson et al., 2021)	✓✓	✓	✓	✓	✓✓	✓✓		
(Ganeshkumar et al., 2021)	✓	✓	✓	✓✓	✓✓	✓		
(E.-J.Kim, 2021)	✓	✓	✓	✓	✓✓	✓✓		
(Pessach et al., 2020)	✓	✓	✓	✓✓	✓	✓✓		
(Andini et al., 2018)	✓	✓	✓	✓	✓	✓✓		

✓: The step is addressed in the study
✓✓: The step is deeply addressed in the study or relies on clear identified theories

(feature engineering) and the last step (robustness checking and validation of explanation) are particularly very little addressed in the literature. We consider that the biggest gap in the current literature is related on this last point. More detailed information about each step of the framework is presented in the next sub-sections.

4.1.1. *Explain the need*

In general, a business analytics process begins by defining the business question to answer. In our context, the dependent variables are specifically to be explained. The aim here is to sufficiently motivate the objectives and relevance of the project. To make this step as comprehensible as possible, it would be wise to rely on well-established theoretical or regulatory foundations. In the articles examined for this study, the authors, for example, drew upon theories such as change behavior theories to justify a classification of purchasing decisions in a marketing context (Taghikhah et al., 2021); resource-based view theory to justify employees turnover predictions (Chowdhury et al., 2022); resource dependence theory to justify cost predictions in negotiations between suppliers and buyers (Bodendorf et al., 2022); emergency management theory to justify the prediction of survival rates from the consumption of opioid drugs (Johnson et al., 2021). From a practical standpoint, the need for model interpretability can also be driven by existing or new standards and norms, such as those in the field of auditing and accounting (Zhang et al., 2022). New standards and norms requiring the explanations of the use of ML models are becoming increasingly common in many other sectors of activity (European Commission, 2016; Goodman and Flaxman, 2017; T. W. Kim and Routledge, 2022), such as banking or real estate (European Banking Authority, 2020; Watch, 2021).

4.1.2. *Explain the data*

Data are usually the raw material for business analytics processes. Thus, identifying suitable data sources for analysis is fundamental. Beyond structured data, models can be built and explained on unstructured data, such as texts or images. Relying on suitable theories can also help the collection and management of good-quality data. For instance, (Taghikhah et al., 2021) integrated five behavioral theories for collecting data through a survey. (Senoner et al., 2022) preferred the quality management theory for improving process quality in manufacturing.

4.1.3. *Explain the features*

Data preparation (e.g., data cleaning, features engineering) usually consists of preparing, cleaning, or transforming collected raw data into suitable features to be used as input for ML algorithms. In practice, it has generally been found that data preparation is a very time-consuming task and can account for approximately 80% of the total predictive modelling effort (Zhang et al., 2003). Thus, explaining data preparation for building ML models is essential for transparency with simulatability or reproducible studies (Chakraborty et al., 2017). In addition, explaining the outputs of a model will naturally be very complicated if the input features used to train that model are not explicable. To tackle this issue in predictive modelling processes, some authors concentrate their efforts on building transparent and understandable features, for instance (Siering, 2022), who relied on information diagnosticity theory for building relevant and understandable features from textual data (users' complaints) to explain why they influence complaint success. In another context, (Ghosh et al., 2022) specifically explain an ensemble feature selection process based on theoretical model for forecasting stocks futures prices.

4.1.4. *Explain algorithms, training process, and performance metrics*

The algorithms, training process, and predictive capacity performance metrics used to build and validate ML models are more technical tasks that should also be well explained to make the overall process more transparent (Nyawa et al., 2023). It could be relevant to first evaluate the predictive capacity of both white-box and black-box models before choosing which one could provide the best explanations to end users (Pessach et al., 2020). It is very common for many authors to systematically use black-box models for problems considered complex; however, some studies show that white-box models (e.g., RSML based on rough set theory) (Gue et al., 2022) or the combination of multiple white-box models into new models (Naumets and Lu, 2021; T. Wang et al., 2022) could also be relevant in such contexts. Even if most white-box models are nowadays data-based or statistical-based AI approaches, approaches relying on knowledge-based symbolic AI (e.g., using formal logic) could also be considered. A good illustration is (Chi and Liao, 2022), who relied on the formal argumentation theory.

When black-box models are used, explaining the assumptions behind the implementation of those models could be important for greater transparency (Chakraborty et al., 2017). The hyperparameters used to train the models should also be well explained (X. Wang et al., 2021; Watson et al., 2022). In many cases, multiple black-box models are evaluated (e.g., through cross-validations), and those presenting the best predictive capacity performances are retained for the explanation step. However, depending on each context, it is important to identify and explain the right metrics to be used to assess the predictive performances among all the possible options (e.g., precision, recall, F1-score, AUC for classifications or MAE, RMSE, MSLE, R2 for regressions) (Dessain, 2022). Overall, best practices from statistical learning theory can be used (Bousquet et al., 2004; Hastie et al., 2009).

4.1.5. Explain the predictions

White-box models are intrinsically designed to propose explainable predictions. In this study, we found that 24% of the articles relied on white-box model explanations. Depending on the context, explanations from white-box models are commonly used when they present good and acceptable predictive capacities (e.g., Andini et al., 2018; Gue et al., 2022; Svoboda and Minner, 2022). Some experts also prefer these models even when black-box models show better predictive capacities (Naumets and Lu, 2021). However, most studies (76%) relied on post hoc explanations of black-box models. The most frequently used post hoc methods are presented in Fig. 7 (SHAP, LIME, Permutation Feature Importance methods, Partial Dependence Plots, Grad-CAM, ALE, and counterfactual explanations). The most-used method (SHAP) is the only method that is built with a strong theoretical background (cooperative game theory), but it only works with structured and tabular data. For unstructured data, such as images, Grad-CAM is the most popular method. For unstructured textual data, most of the studies relied on extracting relevant features from text to construct structured data to be used for explanations. Beyond these frequent post hoc methods, some authors proposed custom methods, such as specific methods for recommender systems (Boulmaiz et al., 2022; Zanon et al., 2022) or the extension of existing methods (e.g., SHAP-MRMR) (Vo et al., 2021). Whether with white-box or post hoc explanations, most of the studies provided both global and local explanations. However, a vast majority of those studies just provide illustrations of the use of explanations of predictions, and very few studies deepened the analysis by checking the robustness of explanations or validating the usefulness of explanations with domain experts or final users (Doshi-Velez and Kim, 2017).

4.1.6. Robustness and validation of explanations

The most common approaches used, particularly for post hoc model explanations, can suffer from a lack of robustness in their outputs. For example, methods relying on input perturbations such as SHAP and LIME can provide inconsistent results depending on the nature of the underlying classifier (Slack et al., 2020). Similarly, counterfactual explanations can also provide misleading outputs when the automatically generated instances used to provide the explanations do not reflect ground-truth data (Laugel et al., 2019). More globally, post hoc explanation methods can suffer from faithfulness, stability, or explicitness (Alvarez Melis and Jaakkola, 2018; Rudin, 2019). Despite these important potential issues, only a fraction of the articles studied in this paper focuses on evaluating the robustness or quality of their explanations when relying on post hoc methods (Chi and Liao, 2022; Irarrázaval et al., 2021; Senoner et al., 2022; T. Wang et al., 2022). We, therefore, include this additional step in our proposed framework to emphasise the need to evaluate explanations in two sub-steps: robustness checking of explanations (which can be done automatically) and validation of explanations (made by stakeholders, such as final users, domain experts, regulators, or decision makers). These two sub-steps can be assimilated to the concepts of objective evaluations and human-centered evaluations highlighted by (Vilone and Longo, 2021) in their review.

The robustness checking of explanations can consist of comparing the explanations (global or local) provided by different methods (Senoner et al., 2022). Other authors rely on other approaches based on metrics such as fidelity (the degree to which the explanation model can mimic the original model's decision) or accuracy (the extent to which the explanation model can accurately make the right predictions) (Chi and Liao, 2022). In contexts sensitive to personal data, the robustness of explanations can also be assessed with the evaluation of important potential model bias properties such as fairness (e.g., impacts or importance of specific variables such as gender, race, age, or other socio-demographic characteristics) (Siering, 2022; Zanon et al., 2022).

The validation of explanations should rather be performed by human experts in the field (e.g., final users, domain experts, regulators, and decision makers). This validation is important because explanations are usually used as a decision support tool; thus, it is fundamental to get quality feedback from the decision makers about the provided explanations. Currently, only a few studies push the evaluation process until this final stage with domain experts or regulators (Naumets and Lu, 2021; Senoner et al., 2022; T. Wang et al., 2022).

4.2. Implications for research

XAI is a highly emerging research field in recent years. Many review articles have been proposed to elaborate shared definitions of key terms (e.g., explainable AI, transparent AI, trustworthy AI, responsible AI, accountability, fairness) or taxonomies of methods (e.g., white-box, post hoc, global, local, visualisations, surrogates) (e.g., Angelov et al., 2021; Arrieta et al., 2020; Das and Rad, 2020; Minh et al., 2022; Schwalbe and Finzel, 2023; Vilone and Longo, 2021). More recently, a survey of surveys about the definitions and taxonomies of XAI has even been proposed (Schwalbe and Finzel, 2023). As the XAI concept is very applicative, there is a crucial need for associated applicative research (Belle and Papantonis, 2021). From the business perspective, only a few reviews have been proposed for specific activity sectors, such as supply chain operational risk management (Nimmy et al., 2022), environmental management research (Arashpour, 2023), smart cities (Javed et al., 2023), biomedical sciences (Nazir et al., 2023) or healthcare (Bharati et al., 2023). However, there is currently a lack of studies providing a holistic research analysis of the use of XAI in real-world applications. To the best of our knowledge, our study is the first to review the global applications of XAI in industry. Grounded in a theoretical analysis through the TCCM framework, we highlighted the most important theories used in the applications, and we proposed a new methodological framework that can be easily reused or extended in future studies.

4.3. Implications for practice

This study is mainly oriented towards the practice of XAI in companies. AI is rapidly transforming every aspect of society with exponential adoption in companies this last decade, following the emergence of modern data-oriented AI techniques, such as ML, DL, or RL (Wamba et al., 2021). Nonetheless, the lack of transparency or interpretability of the process and outputs of these techniques slows down the adoption of AI in many application fields (Schwalbe and Finzel, 2023). In recent years, the rise of XAI has brought some solutions to these problems. However, many questions remain regarding the practical implementation of XAI methods (e.g., When, and why to rely on XAI? How should XAI be integrated into business analytics processes? Which methods should be used? How can we evaluate the relevance of the outputs of XAI methods?). With this study, all practitioners can quickly find an overview of the main issues and concepts to consider for a good implementation of XAI in their specific context. The different steps and theoretical background provided in the proposed methodological framework can help establish solid foundations for a roadmap to follow. This concerns many target audiences (Arrieta et al., 2020): domain experts (for trusting the models themselves or gaining scientific knowledge), regulatory entities (for certifying model compliance with the legislation in force, or for audits), final users (for understanding their situation or verifying fair decisions), data scientists or developers (for ensuring/improving product efficiency, research, or new functionalities), managers, and executives (for assessing regulatory compliance or understanding corporate AI applications).

4.4. Limitations and future research directions

Despite efforts to minimize bias, there can still be subjectivity in the methodological process of literature reviews in general. For instance, even if the selected keywords are carefully chosen, the search of articles may not have covered all published papers. Another possible limitation is the use of a single database (Scopus) to perform the keyword search.

However, Scopus is by far the largest database of scholarly articles (Comerio and Strozzi, 2019; Norris and Oppenheim, 2007), and we also opted for a highly qualitative study by only including renowned journals through the use of well-established international rankings. In this context, this gives us confidence in the consistency of our results, knowing that future studies could also investigate with other keywords and include other scientific databases, scopes of study, or search criteria. Beyond these possible methodological limitations, the findings of this study allow us to identify seven relevant research and practical challenges that should be addressed in future studies to improve the use and the adoption of XAI in industry.

The proposed methodological and theoretical framework in this paper is intended to be directly used by practitioners or researchers. However, there is a room for future research studies to expand it globally or adapt it according to specific application contexts. Six other complementary and high potential practical challenges could also be explored. First, future studies must put more emphasis on the explanation of the features engineering step, and the evaluation of the robustness of the explanations, along with the validation of explanations by human stakeholders. As can be observed in the summary in Table 4, these important steps are currently very little explored. In addition, some related issues, such as finding the right presentation of XAI results to users, should be addressed (Riveiro and Thill, 2021). More globally, both design science and behavioral science approaches should be developed in this context (Hevner et al., 2004). Second, even if many application domains appear in our results, experiments in many other important AI application fields are still missing and should be explored (e.g., cybersecurity, automotive, aerospace, telecommunications, chemistry). For instance, the proliferation of AI applications in cybersecurity in recent years also reinforces the necessity of employing XAI in this field. Third, although most XAI applications rely on post-hoc explanations of black-box models, these explanations can only provide an approximation of the internal logic of these highly complex and non-linear models. More efforts should be directed towards the improvement and the use of white-box models, relying, for example, on symbolic AI (e.g., formal argumentation) which is inherently explainable. It could also be relevant to explore the combination of these symbolic AI methods with black-box AI methods. Fourth, our study shows that explanations of supervised learning systems (mostly classifications or regressions) are the most evaluated in business applications. Only a few emphases are dedicated to recommender systems, and these should be strengthened given the major importance of these systems in many application fields (e.g., social media, e-commerce, search engines). The same remark also applies to recent supervised learning techniques, such as deep reinforcement learning (Heuillet et al., 2021). Beyond supervised learning, there should be more focus particularly on unsupervised learning applications (e.g., clustering), which are also widely used in industry, and would be more relevant if they are well explained (Antwarg et al., 2021; Moshkovitz et al., 2020). Five, our study revealed that most XAI applications in business rely on structured tabular data. Knowing that companies are processing more and more unstructured data in the current big data era, future studies should place more emphasis on using XAI with unstructured data (e.g., texts, images, graphs, audio, videos) (e.g., Tiddi and Schlobach, 2022). Finally, because global warming is nowadays a major threat to our planet and our society, explainability and transparency of AI models could go beyond the modelling process or the outputs of models, and also focus on transparently providing environmental metrics (e.g., carbon footprint) generated by the training and the usage of AI models (Delanoë et al., 2023).

5. Conclusion

Unlike existing reviews of XAI, this study focuses on a holistic review of business applications of XAI techniques in 37 rigorously selected articles published in high-quality journals. From the theoretical perspective, we found that only few studies relied on domain field theories for providing deeper explanations (e.g., for data collection and feature engineering). From the methodological perspective, we found that the current literature mostly addresses explanations for the modelling phase (ML algorithms, training of ML models, predictive capacity of ML models), and the outputs of this modelling phase (e.g., explanation of predictions). Only few studies deeply address explanations for the importance of the business question, the data collection, or the feature engineering process. The robustness checking and the validation of explanations by business users are particularly very little addressed in the literature.

From our global analysis using the TCCM protocol, we also propose a generic methodological and theoretical framework that can be used by all practitioners or stakeholders (e.g., managers and executives, domain experts, final users, data scientists or developers, and regulators) for useful implementation of XAI in their business applications. We highlight the need for explaining the whole analytical process usually used for deriving value from data, including six main steps that can be supported with a strong theoretical background. We particularly emphasize the need for the last step (robustness and human validation of explanations), which is very little discussed in the current literature. This can be one of the many identified high-potential future research avenues, among others, such as the use or the extension of the proposed framework, explanations of unsupervised learning applications, unstructured data applications, exploration of new application fields, or the transparency of external environmental metrics such as the carbon footprint generated by the training and the usage of AI models.

CRediT authorship contribution statement

Dieudonné Tchuente: Conceptualization, Methodology, Data curation, Visualization, Validation, Writing – original draft, Writing – review & editing, Supervision. **Jerry Lonlac:** Conceptualization, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Bernard Kamsu-Foguem:** Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

References

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

Alvarez Melis, David, Jaakkola, Tommi, 2018. Towards robust interpretability with self-explaining neural networks. Adv. Neural Inf. Process. Syst. 31.

Andini, Monica, et al., 2018. Targeting with machine learning: an application to a tax rebate program in Italy. J. Econ. Behav. Organ. 156, 86–102.

Angelov, Plamen P., et al., 2021. Explainable artificial intelligence: an analytical review. " *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* 11 (5), e1424.

Antwarg, Liat, Miller, Ronnie Mindlin, Shapira, Bracha, Rokach, Lior, 2021. Explaining anomalies detected by autoencoders using shapley additive explanations. Expert Syst. Appl. 186, 115736.

Arashpour, Mehrdad, 2023. AI explainability framework for environmental management research. J. Environ. Manag. 342, 118149.

Arrieta, Alejandro Barredo, et al., 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

Barney, Jay, 1991. Firm resources and sustained competitive advantage. J. Manag. 17 (1), 99–120.

Barringer, Bruce R., Harrison, Jeffrey S., 2000. Walking a tightrope: creating value through interorganizational relationships. J. Manag. 26 (3), 367–403.

Belle, Vaishak, Papantonis, Ioannis, 2021. Principles and practice of explainable machine learning. *Front. Big Data*.

Bharati, Subrato, Mondal, M.Rubaiyat Hossain, Podder, Prajoy, 2023. A review on explainable artificial intelligence for healthcare: why, how, and when? IEEE Trans. Artif. Intell.

Bodendorf, Frank, Merkl, Philipp, Franke, J.örg, 2021. Artificial neural networks for intelligent cost estimation–a contribution to strategic cost management in the manufacturing supply chain. Int. J. Prod. Res.

Bodendorf, Frank, Xie, Qiao, Merkl, Philipp, Franke, J.örg, 2022. A multi-perspective approach to support collaborative cost management in supplier-buyer dyads. Int. J. Prod. Econ. 245, 108380.

Boulmaiz, Fateh, Reignier, Patrick, Ploix, Stephane, 2022. An occupant-centered approach to improve both his comfort and the energy efficiency of the building. Knowl. -Based Syst. 249, 108970.

Bousquet, Olivier, Stéphane Boucheron, and G.ábor Lugosi. 2004. "Introduction to Statistical Learning Theory." Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2–14, 2003, Tübingen, Germany, August 4–16, 2003, Revised Lectures: 169–207.

Branstad, Are, Solem, Birgit A., 2020. Emerging theories of consumer-driven market innovation, adoption, and diffusion: a selective review of consumer-oriented studies. J. Bus. Res. 116, 561–571.

Breiman, L., 2001. Random Forests. Mach. Learn. 45 (1), 5–32.

Bücker, Michael, Szepannek, Gero, Gosiewska, Alicja, Biecek, Przemyslaw, 2022. Transparency, auditability, and explainability of machine learning models in credit scoring. J. Oper. Res. Soc. 73 (1), 70–90.

Cerutti, Federico, Nava Tintarev, and Nir Oren. 2014. "Formal Argumentation: A Human-Centric Perspective." In *Eleventh International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2014)*,.

Chakraborty, Supriyo et al. 2017. "Interpretability of Deep Learning Models: A Survey of Results." In 2017 IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI), IEEE, 1–6.

Chen, Chaofan, et al., 2022. A holistic approach to interpretability in financial lending: models, visualizations, and summary-explanations. Decis. Support Syst. 152, 113647.

Chen, Yanyan, Mandler, Timo, Meyer-Waarden, Lars, 2021. Three decades of research on loyalty programs: a literature review and future research agenda. J. Bus. Res. 124, 179–197.

Chi, H., Liao, B., 2022. A quantitative argumentation-based automated explainable decision system for fake news detection on social media. Knowl. -Based Syst. 242.

Chowdhury, Soumyadeb, et al., 2022. Embedding transparency in artificial intelligence machine learning models: managerial implications on predicting and explaining employee turnover. Int. J. Hum. Resour. Manag.

Comerio, Niccolò, Strozzi, Fernanda, 2019. Tourism and its economic impact: a literature review using bibliometric tools. Tour. Econ. 25 (1), 109–131.

Dalzochio, Jovani, et al., 2020. Machine learning and reasoning for predictive maintenance in industry 4.0: current status and challenges. Comput. Ind. 123, 103298.

Das, Arun, Rad, Paul, 2020. "Opportunities and challenges in explainable artificial intelligence (Xai): a survey. Artif. Intell. (Xai): A Surv. " *arXiv Prepr. arXiv* 2006, *11371*.

Delanoë, Paul, Tchuente, Dieudonné, Colin, Guillaume, 2023. Method and evaluations of the effective gain of artificial intelligence models for reducing CO_2 emissions. J. Environ. Manag. 331, 117261.

Dessain, Jean, 2022. Machine learning models predicting returns: why most popular performance metrics are misleading and proposal for an efficient metric. Expert Syst. Appl. 199, 116970.

Donthu, Naveen, et al., 2021. Mapping the electronic word-of-mouth (EWOM) research: a systematic review and bibliometric analysis. J. Bus. Res. 135, 758–773.

Doshi-Velez, Finale, Kim, Been, 2017. Towards a rigorous science of interpretable machine learning. *arXiv Prepr. arXiv* 1702, *08608*.

Doumbouya, Mamadou Bilo, Kamsu-Foguem, Bernard, Kenfack, Hugues, Foguem, Clovis, 2018. Argumentation graphs with constraint-based reasoning for collaborative expertise. Future Gener. Comput. Syst. 81, 16–29.

Elder, Ryan S., Krishna, Aradhna, 2012. The 'visual depiction effect' in advertising: facilitating embodied mental simulation through product orientation. J. Consum. Res. 38 (6), 988–1003.

Escalas, Jennifer Edson, 2007. Self-referencing and persuasion: narrative transportation versus analytical elaboration. J. Consum. Res. 33 (4), 421–429.

European Banking AuthorityGuidelines on Loan Origination and Monitoring. 〈https://www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Guidelines/2020/Guidelines%20on%20loan%20origination%20and%20monitoring/884283/EBA%20GL%202020%2006%20Final%20Report%20on%20GL%20on%20loan%20origination%20and%20monitoring.pdf) 2020.

European Commission. 2016. "General Data Protection Regulation." Official Journal of the European Union. 〈https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679〉.

Feldman, Jack M., Lynch, John G., 1988. Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. J. Appl. Psychol. 73 (3), 421.

Fisher, Aaron, Rudin, Cynthia, Dominici, Francesca, 2019. "All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. " *J. Mach. Learn. Res.* 20 (177), 1–81.

Friedman, Jerome H., 2001. Greedy function approximation: a gradient boosting machine. *Ann. Stat.*: 1189–1232.

Ganeshkumar, M. et al. 2021. "Explainable Deep Learning-Based Approach for Multilabel Classification of Electrocardiogram." *IEEE Transactions on Engineering Management*.

Ghosh, Indranil, et al., 2022. A hybrid approach to forecasting futures prices with simultaneous consideration of optimality in ensemble feature selection and advanced artificial intelligence. Technol. Forecast. Soc. Change 181, 121757.

Goldstein, Alex, Kapelner, Adam, Bleich, Justin, Pitkin, Emil, 2015. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J. Comput. Graph. Stat. 24 (1), 44–65.

Goodman, Bryce, Flaxman, Seth, 2017. European union regulations on algorithmic decision-making and a 'right to explanation. AI Mag. 38 (3), 50–57.

Gozzi, Noemi, Malandri, Lorenzo, Mercorio, Fabio, Pedrocchi, Alessandra, 2022. XAI for myo-controlled prosthesis: explaining emg data for hand gesture classification. Knowl. -Based Syst. 240, 108053.

Grant, Maria J., Booth, Andrew, 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Inf. Libr. J. 26 (2), 91–108.

Gue, Ivan Henderson V., et al., 2022. Predicting waste management system performance from city and country attributes. J. Clean. Prod. 366, 132951.

Ha, Taehyun, 2022. An explainable artificial-intelligence-based approach to investigating factors that influence the citation of papers. Technol. Forecast. Soc. Change 184, 121974.

Hartmann, Jochen, Heitmann, Mark, Schamp, Christina, Netzer, Oded, 2021. The power of brand selfies. J. Mark. Res. 58 (6), 1159–1177.

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome H., Friedman, Jerome H., 2009. 2 The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Heuillet, Alexandre, Couthouis, Fabien, Díaz-Rodríguez, Natalia, 2021. Explainability in deep reinforcement learning. Knowl. -Based Syst. 214, 106685.

Hevner, Alan R., March, Salvatore T., Jinsoo Park, Sudha Ram, 2004. Design science in information systems research. MIS Q. 28 (1), 75–105. 〈http://www.jstor.org/stable/25148625〉.

Irarrázaval, María Elisa, Maldonado, Sebastián, Pérez, Juan, Vairetti, Carla, 2021. Telecom traffic pumping analytics via explainable data science. Decis. Support Syst. 150, 113559.

Jana, Rabin K., Ghosh, Indranil, Wallin, Martin W., 2022. Taming energy and electronic waste generation in bitcoin mining: insights from facebook prophet and deep neural network. Technol. Forecast. Soc. Change 178, 121584.

Javed, Abdul Rehman, et al., 2023. A survey of explainable artificial intelligence for smart cities. Electronics 12 (4), 1020.

Jiang, Z., Benbasat, I., 2004. Virtual product experience: effects of visual and functional control of products on perceived diagnosticity and flow in electronic shopping. J. Manag. Inf. Syst. 21 (3), 111–147.

Johnson, Marina, Albizri, Abdullah, Harfouche, Antoine, Tutun, Salih, 2021. Digital transformation to mitigate emergency situations: increasing opioid overdose survival rates through explainable artificial intelligence. Ind. Manag. Data Syst. 123 (1), 324–344.

Kamm, Simon, et al., 2023. A survey on machine learning based analysis of heterogeneous data in industrial automation. Comput. Ind. 149, 103930.

Kim, Eui-Jin, 2021. Analysis of travel mode choice in seoul using an interpretable machine learning approach. J. Adv. Transp. 2021, 1–13.

Kim, Juram, Lee, Gyumin, Lee, Seungbin, Lee, Changyong, 2022. Towards expert–machine collaborations for technology valuation: an interpretable machine learning approach. Technol. Forecast. Soc. Change 183, 121940.

Kim, Tae Wan, Routledge, Bryan R., 2022. Why a right to an explanation of algorithmic decision-making should exist: a trust-based approach. Bus. Ethics Q. 32 (1), 75–102.

Laugel, Thibault, et al., 2019. "The dangers of post-hoc interpretability: unjustified counterfactual explanations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19. AAAI Press,, pp. 2801–2807.

Lee, Donghyun, et al., 2022. Integrated explainable deep learning prediction of harmful algal blooms. Technol. Forecast. Soc. Change 185, 122046. 〈https://www.sciencedirect.com/science/article/pii/S0040162522005674〉.

Liu, J., Toubia, O., Hill, S., 2021. Content-based model of web search behavior: an application to TV show search. Manag. Sci. 67 (10), 6378–6398.

Lorenz, Felix, Willwersch, Jonas, Cajias, Marcelo, Fuerst, Franz, 2022. Interpretable machine learning for real estate market analysis. " *Real. Estate Econ.*

Lundberg, Scott M., Lee, Su-In, 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

McLoughlin, David, 1985. A framework for integrated emergency management. Public Adm. Rev. 45, 165–172.

Mengist, Wondimagegn, Soromessa, Teshome, Legese, Gudina, 2020. Method for conducting systematic literature review and meta-analysis for environmental science research. MethodsX 7, 100777.

Minh, Dang, Xiang Wang, H., Fen Li, Y., Nguyen, Tan N., 2022. Explainable artificial intelligence: a comprehensive review. Artif. Intell. Rev. 1–66.

Moshkovitz, Michal, Dasgupta, Sanjoy, Rashtchian, Cyrus, Frost, Nave, 2020. Explainable K-means and k-medians clustering. In *International Conference on Machine Learning*. PMLR, pp. 7055–7065.

Naumets, S., Lu, M., 2021. Investigation into explainable regression trees for construction engineering applications. J. Constr. Eng. Manag. 147 (8).

Nazir, Sajid, Dickson, Diane M., Akram, Muhammad Usman, 2023. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. Comput. Biol. Med., 106668

Nimmy, Sonia Farhana, et al., 2022. Explainability in supply chain operational risk management: a systematic literature review. Knowl. -Based Syst. 235, 107587.

Norris, Michael, Oppenheim, Charles, 2007. Comparing alternatives to the web of science for coverage of the social sciences' literature. J. Informetr. 1 (2), 161–169.

Nyawa, Serge, Gnekpe, Christian, Tchuente, Dieudonné, 2023. Transparent machine learning models for predicting decisions to undertake energy retrofits in residential buildings. *Ann. Oper. Res.*

Onchis, Darian M., Gillich, Gilbert-Rainer, 2021. Stable and explainable deep learning damage prediction for prismatic cantilever steel beam. Comput. Ind. 125, 103359.

Page, Matthew J., et al., 2021. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 372.

Park, Seyoung, Joung, Junegak, Kim, Harrison, 2022. Spec guidance for engineering design based on data mining and neural networks. Comput. Ind. 144, 103790.

Paul, Justin, Rosado-Serrano, Alexander, 2019. Gradual internationalization vs born-global/international new venture models: a review and research agenda. Int. Mark. Rev. 36 (6), 830–858.

Paul, Justin, Merchant, Altaf, Dwivedi, Yogesh K., Rose, Gregory, 2021. Writing an impactful review article: what do we know and what do we need to know? J. Bus. Res. 133, 337–340.

Paulraj, Antony, Chen, Injazz J., 2007. Strategic buyer–supplier relationships, information technology and external logistics integration. J. Supply Chain Manag. 43 (2), 2–14.

Pawlak, Zdzisław, 1982. Rough sets. Int. J. Comput. Inf. Sci. 11, 341–356.

Pessach, Dana, et al., 2020. Employees recruitment: a prescriptive analytics approach via machine learning and mathematical programming. Decis. Support Syst. 134, 113290.

Pfeffer, Jeffrey, Salancik, Gerald R., 2003. The External Control of Organizations: A Resource Dependence Perspective. Stanford University Press.

Prakken, Henry, Vreeswijk, Gerard, 2002. Logics for defeasible argumentation. Handb. Philos. Log. 219–318.

Raza, Ali, Tran, Kim Phuc, Koehl, Ludovic, Li, Shujun, 2022. Designing ecg monitoring healthcare system with federated transfer learning and explainable Ai. Knowl. -Based Syst. 236, 107763.

Ribeiro, Marco Tulio, Singh, Sameer, Guestrin, Carlos, 2016. Why should i trust you?' Explaining the predictions of any classifier. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 1135–1144.

Ribeiro, Marco Tulio, Singh, Sameer, Guestrin, Carlos, 2018. "Anchors: high-precision model-agnostic explanations. *Proc. AAAI Conf. Artif. Intell.*

Riveiro, Maria, Thill, Serge, 2021. That's (Not) the output i expected!' On the role of end user expectations in creating explanations of AI systems. Artif. Intell. 298, 103507.

Rudin, Cynthia, 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1 (5), 206–215.

Rudner, Richard S., 1968. Philosophy of social science. *Br. J. Philos. Sci.* 18, 4.

Schmenner, Roger W., Swink, Morgan L., 1998. On theory in operations management. J. Oper. Manag. 17 (1), 97–113.

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

Selvaraju, Ramprasaath R., et al., 2017. "Grad-cam: visual explanations from deep networks via gradient-based localization. *Proc. IEEE Int. Conf. Comput. Vis.* 618–626.

Senoner, Julian, Netland, Torbjørn, Feuerriegel, Stefan, 2022. Using explainable artificial intelligence to improve process quality: evidence from semiconductor manufacturing. Manag. Sci. 68 (8), 5704–5723.

Siering, Michael, 2022. Explainability and fairness of regtech for regulatory enforcement: automated monitoring of consumer complaints. Decis. Support Syst. 158, 113782.

Slack, Dylan et al. 2020. "Fooling Lime and Shap: Adversarial Attacks on Post Hoc Explanation Methods." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 180–186.

Souza, Marcos Leandro Hoffmann, da Costa, Cristiano André, de Oliveira Ramos, Gabriel, 2023. A machine-learning based data-oriented pipeline for prognosis and health management systems. Comput. Ind. 148, 103903.

Susan, M.Mudambi, David, Schuff, 2010. What makes a helpful online review? A study of customer reviews on Amazon. Com. MIS Q. 34 (1), 185–200.

Svoboda, Josef, Minner, Stefan, 2022. Tailoring inventory classification to industry applications: the benefits of understandable machine learning. Int. J. Prod. Res. 60 (1), 388–401.

Taghikhah, Firouzeh, Voinov, Alexey, Shukla, Nagesh, Filatova, Tatiana, 2021. Shifts in consumer behavior towards organic products: theory-driven data analytics. J. Retail. Consum. Serv. 61, 102516.

Taguchi, Genichi. 1986. *Introduction to Quality Engineering: Designing Quality into Products and Processes.*

Tiddi, Ilaria, Schlobach, Stefan, 2022. Knowledge graphs as tools for explainable machine learning: a survey. Artif. Intell. 302, 103627.

Tsoka, Thamsanqa, et al., 2022. Explainable artificial intelligence for building energy performance certificate labelling classification. J. Clean. Prod. 355, 131626.

Turing, Alan Mathison, 1950. "Computing machinery and intelligence. Mind 59 (236), 433.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Netw. 10 (5), 988–999.

Vilone, Giulia, Longo, Luca, 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion 76, 89–106.

Vo, Nhi N.Y., Liu, Shaowu, Li, Xitong, Xu, Guandong, 2021. Leveraging unstructured call log data for customer churn prediction. Knowl. -Based Syst. 212, 106586.

Vultureanu-Albişi, A., Bădică, C., 2021. "Recommender systems: an explainable AI perspective. *2021 Int. Conf. Innov. Intell. Syst. Appl. (INISTA)* 1–6.

Wamba, Samuel Fosso, et al., 2021. Are we preparing for a Good AI society? A bibliometric review and research agenda. Technol. Forecast. Soc. Change 164, 120482.

Wang, T., He, C., Jin, F., Hu, Y.J., 2022. Evaluating the effectiveness of marketing campaigns for malls using a novel interpretable machine learning model. Inf. Syst. Res. 33 (2), 659–677.

Wang, Xin, Fan, Shuyi, Kuang, Kun, Zhu, Wenwu, 2021. Explainable automated graph representation learning with hyperparameter importance. In *International Conference on Machine Learning*. PMLR, pp. 10727–10737.

Watch, Algorithm 2021. "In Poland, a Law Made Loan Algorithms Transparent. Implementation Is Nonexistent." ⟨https://algorithmwatch.org/en/poland-credit-loan-transparency/⟩ (May 15, 2023).

Watson, Matthew, Awwad Shiekh Hasan, Bashar, Al Moubayed, Noura, 2022. Using model explanations to guide deep learning models towards consistent explanations for EHR data. Sci. Rep. 12 (1), 19899 https://doi.org/10.1038/s41598-022-24356-6.

Zanon, A.L., Rocha, L.C.D.D., Manzato, M.G., 2022. Balancing the trade-off between accuracy and diversity in recommender systems with personalized explanations based on linked open data. Knowl. -Based Syst. 252.

Zhang, Chanyuan (Abigail, Cho, Soohyun, Vasarhelyi, Miklos, 2022. Explainable Artificial Intelligence (XAI) in auditing. Int. J. Account. Inf. Syst. 46, 100572.

Zhao, Sheng, et al., 2021. Interpretable machine learning for predicting and evaluating hydrogen production via supercritical water gasification of biomass. J. Clean. Prod. 316, 128244.

Zimbardo, Philip, and Ebbe B. Ebbesen. 1970. "Influencing Attitudes and Changing Behavior: A Basic Introduction to Relevant Methodology, Theory, and Applications."