IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 November 2023, accepted 16 December 2023, date of publication 25 December 2023,
date of current version 3 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347028

SURVEY

Explainable Artificial Intelligence (XAI): A Systematic Literature Review on Taxonomies and Applications in Finance

TIAGO MARTINS[1], ANA MARIA DE ALMEIDA[1,2,3], (Senior Member, IEEE), ELSA CARDOSO[1,4], (Member, IEEE), AND LUÍS NUNES[1,2]

[1]Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisbon, Portugal
[2]Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, 1649-026 Lisbon, Portuga
[3]Center for Informatics and Systems of the University of Coimbra (CISUC), 3030-290 Coimbra, Portugal
[4]CIES-ISCTE—Centro de Investigação e Estudos de Sociologia, 1649-026 Lisbon, Portugal

Corresponding author: Ana Maria de Almeida (Ana.Almeida@iscte-iul.pt)

This work was supported in part by Fundação para a Ciência e a Tecnologia, I. P. (FCT) under Grant UIDB/03126/2020, Grant UIDB/04466/2020, Grant UIDP/04466/2020; and in part by POAT-01-6177-FEDER-000059.

ABSTRACT Explainable Artificial Intelligence (XAI) is a growing area of research that aims to improve the interpretability of the not-so-informative black-box models. However, it is currently difficult to categorize an existing method in terms of its intrinsic characteristics and explainability. We provide a new unified yet simple taxonomy for the categorization of XAI methods and present the explainability methods currently being applied in finance applications. For both purposes, we present two separate systematic literature reviews: an anthological search for surveys on XAI methods in order to present a unified taxonomy, followed by an exposition of the XAI methods currently in use that have been found. We also concisely define the existing explainability methods using the proposed categories based on the ones most commonly addressed in the reviewed literature and pinpoint specific XAI methods being used in practical applications in Finance.

INDEX TERMS AI, artificial intelligence, financial applications, explainable machine learning, systematic literature review, XAI.

I. INTRODUCTION

Explainable Artificial Intelligence, XAI,[1] is an area that aims to improve the interpretation and explanation of Machine Learning (ML) algorithms and their results. Due to the growing relevance of ML algorithms in recent decades, mainly through black-box approaches such as neural networks or random forests, interest in the ability to interpret and explain these approaches has increased in several application areas, with emphasis on areas related to Health and Finance [1]. The most complex models that learn from examples, that is, supervised learning using neural networks or randomization, where one is expected to input the characteristics of the example and its output, exhibit no or limited transparency. Such models, high in performance yet low in comprehension, need to be explained so that users can understand the (reasons for the) outputs of these models and, consequently, informed decisions can be supported.

Although no universal definition of explainability exists, numerous studies related to XAI, with different purposes and levels of detail for explainability, enable the definition of the main objectives of this area. The purpose of an XAI technique is to understand the behavior of an ML model and its output, as mentioned by M.Turek [2]: "...XAI aims to help users understand, appropriately trust, and effectively manage the emerging generation of Artificial Intelligence (AI) systems." The Assessment List for Trustworthy Artificial

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Wei Tsai.

[1]Acronym popularized by the Defense Advanced Research Projects Agency (DARPA), in 2016, when an announcement was made to potentially fund research proposals in ML towards Explainable Artificial Intelligence (https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf)

Intelligence (ALTAI) defines explainability as a "feature of an AI system that is intelligible to non-experts. An AI system is intelligible if its functionality and operations can be explained non-technically to a person not skilled in the art." Besides these definitions, experts are also highly interested in understanding what is happening inside a model, which can be defined as the interpretability of ML models. Christoph Molnar proposes to define Interpretable ML as the methods and models that make the behavior and predictions of machine learning systems understandable to humans [22]. In general, explanations are meant for humans to trust black-box methods, and explainability mainly focuses on models that can summarize the reasons for the model's results or give insights about the causes of the decisions that have been made and be auditable [63]. Other relevant works in the attempt to define explainability and interpretability can be found in [10], [64], [65].

While the primary purpose of this area of study is to help understand ML models, there is also a legal motivation to help further this area, namely the General Data Protection Regulation (2016/679, GDPR), which is a privacy and data protection regulation.[2] Projects involving personal data in the European Union and Economic Area must comply with this regulation and the possible legal repercussions [3]. GDPR is the European Union's effort to serve the interests of its citizens regarding how their personal data is used by third parties, as well as define the obligations of the parties and establish citizens' rights. Among these, in Article 17, we can find the "right to forget," where the data subject, typically the citizen, can ask the data holder to erase his/her personal data, or, according to Article 21, the right to object to the processing of his/her personal data. While the phrasing of these articles is open to interpretation, some pave the way for XAI as an obligation rather than an optional feature. The GDPR clearly defines that personal data should be processed "…lawfully, fairly and in a transparent manner in relation to the data subject…," as seen in Article 5. While there might be some doubt regarding the applicability of this article, there is a more detailed definition of transparency applied to ML models in the right for a data subject to have the information regarding "…the existence of automated decision-making…" as well as the process, importance, and consequences behind such decision-making, according to Article 14, paragraph II.g). The need for an explanation of ML models becomes even more apparent because it implies that the prediction and the logic behind it should be made available to the user.

The purpose of this paper is two-fold: firstly, to integrate current knowledge regarding XAI techniques and methods, specifically for tabular data; secondly, based on the results of a systematic literature review, introduce the specific XAI methods and techniques that have been applied in the financial domain. The paper is organized as follows: Section II describes the methodology used for the systematic search of articles; Section III presents a quantitative analysis of the search results; in Section IV, a qualitative analysis of the reviewed surveys is made, and a more concise taxonomy is proposed; in Section V, the analysis is employed to understand what are the XAI methods that are currently being applied in the financial sector; finally, in Section VI, conclusions are drawn along with a critical discussion of this work's contributions, as well as the limitations of this study.

[2]2016/679 GDPR Regulation: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504

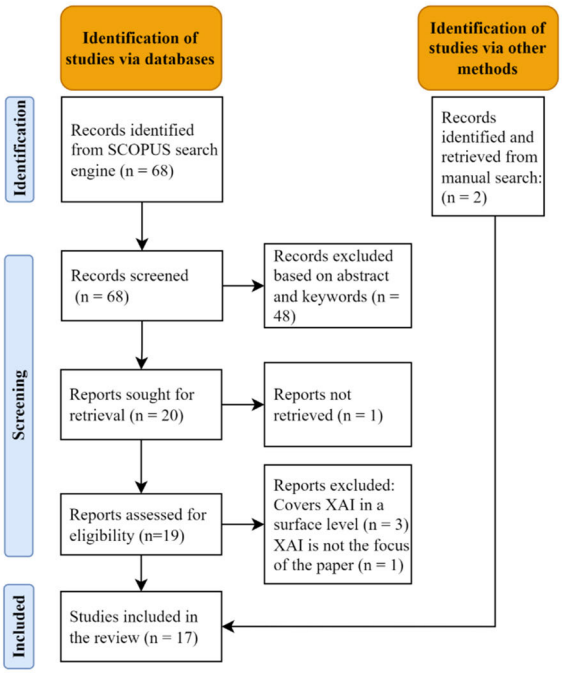

FIGURE 1. PRISMA methodology for surveys.

II. METHODOLOGY

The search for relevant scientific papers follows the PRISMA methodology for systematic literature reviews [4]. The general methodology was adapted for this study's aims, exclusion criteria, and search engines used. Two distinct searches were performed: the first sought a definition of XAI and to understand the different characteristics and implications of this ML area. The second is a systematic search for practical finance applications of XAI methods to bring to light current trends in XAI methods within the financial sector.

A. SEARCH FOR EXISTING LITERATURE SURVEYS ON XAI

The SCOPUS citation database was chosen in the search for surveys and literature reviews, as it is more restrictive than Google Scholar or other engines that do not have a validation component. The query used breaks down into two search elements: first, the definition of the area of study; second, the filter for surveys or literature reviews:

TITLE ("Explainable Artificial Intelligence" OR "Explainable AI" OR "XAI" OR "Interpretable Artificial Intelligence") AND TITLE ("Systematic Review" OR "Review" OR "Survey")

This search was performed without specifying the domain of applications. This fact is not considered relevant since this search aims to obtain an overview of the definition of XAI and a clear specification of the categories of methods. As such, the results seen in Fig. 1 reflect works that are either generic or explicitly applied for a type of data (i.e., tabular data).

After obtaining a batch of 68 original results, an exclusion filter was applied to keywords and abstracts to include only papers focusing on XAI and without any specificity regarding subject areas; this resulted in 20 papers for revision, of which one was considered inaccessible. Finally, two criteria were established to exclude further papers in case XAI was not covered in depth or if it was not the paper's focus, resulting in a final count of 15 documents to be reviewed. Two additional papers were retrieved from a manual search, which increased the total number of surveys to seventeen.

The final list of documents contains 17 surveys whose core concept relates to XAI. These results have been used in Section IV to sustain the proposal of a taxonomy.

B. SEARCH FOR PRACTICAL IMPLEMENTATIONS OF XAI METHODS

As in the previous search, the SCOPUS citation database was chosen as the data source. We needed to define the essential terms for searching for papers on XAI while differentiating between more generic and domain-free approaches and specific applications to finance. Building on previous knowledge, notably of the XAI concept, as well as works describing specific implementations [5], the following domain-free query was constructed, comprising two parts - explainable artificial intelligence and based and generic or specific implementations of XAI methods:

*(TITLE-ABS-KEY("Explainable Artificial Intelligence" OR "Explainable AI" OR xai)) AND (TITLE-ABS-KEY(counterfactual OR *explanation* OR lime OR "Local Surrogate" OR anchors OR "Individual Conditional Expectation" OR ice OR "Accumulated Local Effects" OR ale OR clear OR "Counterfactual Local Explanations for any classifier" OR dice OR permuteattack OR lore OR "Local Rule-Based Explanations" OR dale OR "Differential Accumulated Local Effects" OR pdp OR "Partial Dependence Plot" OR intrees OR treeexplainer OR shap OR "shapley additive explanation" OR "difference net"))*

The 1984 papers obtained show that XAI has gained some traction over the current years. The following search term was added to the query to filter out all papers not related to Finance:

AND (TITLE-ABS-KEY(financ OR loan OR market OR credit))*

As shown in Fig. 2, from the first batch of 1984 papers, 1855 are not subject-specific and were excluded by the automatic filter based on the financial domain keywords. For the remaining 129 papers, a manual analysis of title, abstract, and keywords was made only to include papers on the subject area of Finance and with a focus on XAI, excluding 60 papers.

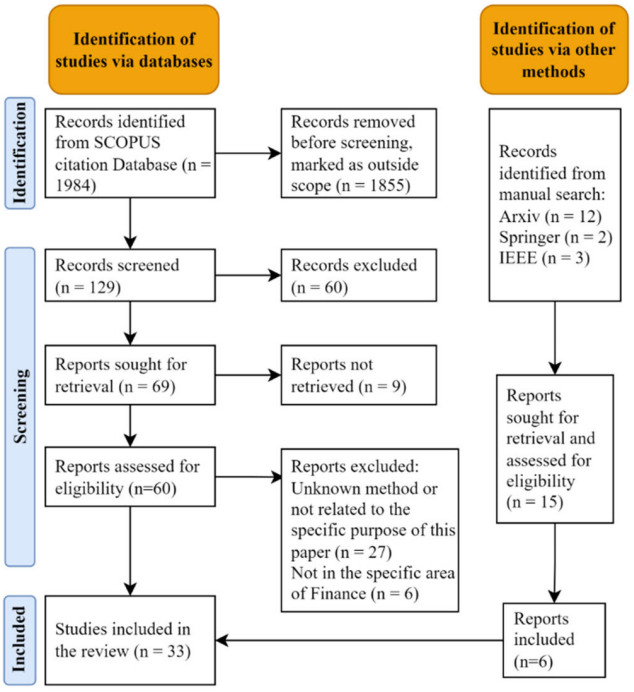

FIGURE 2. PRISMA methodology for practical applications.

This exclusion rendered 69 documents, from which nine were directly excluded due to their unavailability. After analyzing the contents of the 60 accessible documents, another filter was applied. This filter excluded documents that did not specify the XAI method used or were unrelated to a practical application of XAI. A final manual analysis concluded if a paper was unrelated to or not in the area of Finance, namely relevant for credit risk and business failure prediction, which acted as a final criterion for exclusion. Only 27 papers were found to obey the inclusion criteria and were selected for deeper analysis. These final papers mainly focus on practical applications of XAI methods in the financial domain. However, some did not quite fit into this category as they addressed the legal domain. This domain has gained traction in recent years, notably with the broader adoption of GDPR, resulting in these papers being considered important and thus included and mentioned in Section I.

Note that some documents were obtained manually. Fifteen of those were obtained from multiple sources, such as ArXiv,[3] Springer,[4] and IEEE Explorer[5] databases, with an emphasis given to ArXiv due to its characteristic of hosting very recent studies, which allows it to be on par with the current state of specific implementations of XAI methods. XAI is an area with an increasing and recent trend in interest [10], and with studies being published rapidly, ArXiv enables us to know what investigators are working on without waiting for the peer review process. Finally, analyzing the respective papers based on their content resulted in a final list of six out of 15 documents.

[3]https://arxiv.org/
[4]https://link.springer.com/
[5]https://ieeexplore.ieee.org/Xplore/home.jsp

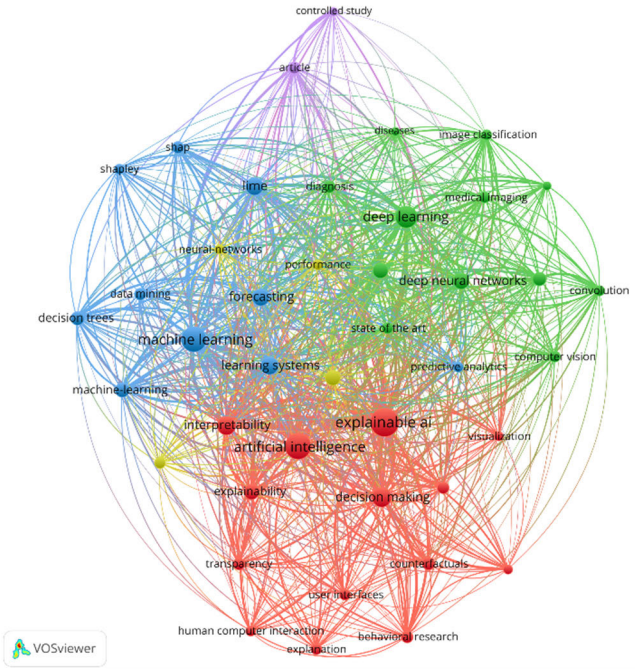

FIGURE 3. Co-occurrence of keywords.

The final list of papers contains 33 documents: 27 from the systematic search and six from the manual search.

III. SEARCH RESULTS

The results of both searches totaled 2069 papers. This section presents an analysis to characterize the rising popularity of XAI as a field of study. We start with a visual analysis produced with the help of the VOSViewer[6] software. The list of results of both the search for practical applications and surveys in the SCOPUS database were combined and passed through VOSViewer by filtering out the most frequent and distinct keywords. This resulted in a co-occurrence network of keywords displayed in Fig. 3. In the graph visualization, the expected closeness between XAI and Artificial Intelligence, interpretability, and decision-making can be easily perceived. Furthermore, it is possible to identify the connections between these research areas and the different implementations of XAI methods, such as SHAP, LIME, Decision Trees, and counterfactual methods. As for application areas, the Health domain is highlighted, with connections to nodes such as medical imaging, diagnosis, and diseases.

Table 1 presents chronological information regarding the publishing years of the papers, showing a definite rise in popularity in recent years. Until 2018, only ten papers regarding XAI were found, as many as the ones that had already been published over the first five days of January 2023. Since then, there has been a stable increase in the number of published papers related to XAI, with each new year approximately doubling the number of publications from the previous year. Interestingly, one of the surveys, Adadi and Berrada's [6], published in 2018, had been cited 999 times at the time of the search (January 5, 2023), which helps point to 2018 as a turning point for the popularity of XAI in general.

TABLE 1. Diachronic overview of papers on XAI.

Year	No. of publications
Pre-2018	10
2018	55
2019	153
2020	318
2021	695
2022	828
2023 (5th January)	10
Total:	2069

TABLE 2. Types of papers published.

Type	Count
Article	782
Book Chapter	44
Conference Paper	1125
Conference Review	41

The document type is significant as, typically, more importance is given to scientific articles than conference papers due to the greater difficulty in publishing the former. Table 2 presents the type of documents obtained. As observed, below half the papers are articles, and the majority are conference papers, which stresses the recent and developing interest in the theme. Still, the number of articles is deemed sufficiently large for this analysis.

To determine the journal's subject area, we used the SCOPUS Journal List[7], which encompassed 43014 journals at the time of the access (November 15, 2022). There were only 527 conference proceedings with corresponding subject areas, so we decided to use only the journal's subject areas. Table 3 presents the subject area and the domains given by the SCOPUS Journal List and displays the counts of papers found by subject area contained in one of the four domains found in the papers: Health Sciences, Life Sciences, Physical Sciences, and Social Sciences. For these subject areas, albeit the fact that (i) this analysis specifically searches for scientific papers and (ii) the focus is on the area of Finance might limit the perspective on other areas, we can still infer that most of the papers arise in the domains of Computer Science, Social Sciences, and Engineering. This is hardly a surprise since these are areas closely related to XAI, especially Computer Science. Other relevant areas are Medicine and Mathematics, where the need for an explanation for any automated decision is most important. While the number of papers classified in these subjects is much less than for the former areas (62 for Medicine and 56 papers for Mathematics), the quantities are still expressive. As for journals in Economics, Accounting, and Finance, a few journals do present papers on this subject (22 papers in total), suggesting that these areas are not yet explored in-depth or, which is common, use AI techniques that are explanatory by default.

[6]https://www.vosviewer.com/

[7]https://www.scopus.com/sources.uri

TABLE 3. Main subject area of the journal the paper is in.

Domain	Subject area	Count
Health Sciences	Medicine	62
	Health Professions	10
	Dentistry	1
Life Sciences	Biochemistry, Genetics, and Molecular Biology	25
	Immunology and Microbiology	3
	Neuroscience	11
	Agricultural and Biological Sciences	10
Physical Sciences	Computer Science	235
	Engineering	127
	Materials Science	21
	Physics and Astronomy	19
	Chemical Engineering	12
	Mathematics	56
	Chemistry	13
	Environmental Science	21
	Energy	15
	Earth and Planetary Sciences	15
Social Sciences	Social Sciences	157
	Arts and Humanities	18
	Psychology	10
	Decision Sciences	24
	Business, Management, and Accounting	15
	Economics, Econometrics, and Finance	7
	General	8
	Total:	895

IV. LITERATURE REVIEW AND ANALYSIS OF EXISTING SURVEYS

Adadi et al. pointed out the need for XAI for several reasons: the need for ML models to comply with existing legislation to better understand the systems developed, which in turn allows for a better perception of the flaws or vulnerabilities of these systems [6]. The authors also propose explainability to make model improvements easier because of the explicit need for an explanation since it helps to extract knowledge. Five domains of application are highlighted, including Finance. The paper also presents a detailed taxonomy for characterizing XAI methods, concluding that this area still needs further research.

A historical perspective of XAI is the focus of Angelov et al., which also detail several XAI methods categorized based on their taxonomy proposal [7]. The authors also describe several key applications for XAI, ranging from the criminal justice system to fraud detection. They conclude with three main points: the importance of the area, how to fill the gap between Deep Learning and Neuroscience with XAI, and finally, future directions for work.

Islam et al. present a systematic review that identifies specific domains and applications of XAI methods based on 137 reviewed papers [8]. From these, only three are found to be in the financial domain. The authors conclude with the proposal of a taxonomy for XAI techniques that, albeit new, is largely influenced by the work of other authors. A similar study, which also classified papers in specific domains and applications, analyzed and classified 350 papers based on these authors' taxonomy proposal, which was created based on the analysis of literature and previously proposed classification systems [9]. Linardatos et al. also propose a taxonomy built upon previous work, emphasizing the application of XAI methods to specific areas of AI and reviewing several techniques, some specifically for Deep Learning, while others with a more general approach, including white-box XAI methods [10].

Minh et al. focus on a review of the theoretical background for XAI [11]. Each paper is categorized in terms of the type of explanation provided, and the advantages and disadvantages of each of the XAI approaches described are discussed. The authors also propose a taxonomy to classify the papers with independent categories [11].

In [12], the authors explore an in-depth review of specific implementations and respective categorization along with some practical applications based on the justifications raised previously in [6]. Finally, the authors discuss the practical applications of XAI per domain, current limitations, and future work in this area.

Lin et al. introduce a hierarchical taxonomy, focusing on XAI approaches emphasizing Deep Learning. The authors also raise some issues, namely the trade-off between model interpretability and performance when using Deep Learning [13].

The definition of a taxonomy for XAI methods, primarily adapted from other papers and with several categories that include but are not limited to the domain of application of the method, is presented in [14]. After reviewing previous work, another paper focusing on the definition of a proper taxonomy of XAI methods concludes with a proposal for a taxonomy trying to adapt the taxonomies found in their review [15].

Darias et al. [16] analyze XAI methods libraries and compare each one of the approaches found. The authors focus on how each XAI method generates explanations, not how they fit in a taxonomy, hence its exclusion from Table 4. In a systematic literature review, the authors systematically analyze papers looking for ways to tackle the problem of cognitive bias or the "systematic error in judgment and decision-making common to all human beings" (as defined in [21]) that has been found in XAI methods used in decision-making systems [17]. While the authors do not provide a taxonomy for XAI methods, it is a relevant paper that helps us understand how we use and trust XAI methods. An exploration of the ethical principles of XAI can be found in [18], with a focus on reviewing current methods used in the area and providing a taxonomy for these based on previous works. In another survey, Stepin et al. discuss contrastive and counterfactual explanations and propose a taxonomy for these methods [19]. Finally, Lopes et al. created a taxonomy not for XAI methods but rather for evaluating such methods [20].

TABLE 4. Support for the proposed taxonomy.

Category for XAI methods	Works that support the category definition
Stage	[6], [8]–[11], [13], [15], [22], [23]
Model	[6], [8], [10], [12], [13], [15], [22], [23]
Scope	[6], [8], [10], [14], [15], [22]

Based on the reviewed literature, we can conclude that no standard categorization of XAI methods still exists. This opinion is supported by Vilone and Longo, who, in 2020, based on an extensive search, concluded that there is no adequate definition of what an explanation is in ML and that the task of formalizing XAI is complex due to the XAI's cross-domain applicability [9]. This disagreement in achieving a unified taxonomy comes from comparing the approaches of Islam et al. [8] and Molnar [22], where the former proposes four main categories, while the latter suggests only three. Nevertheless, two of the categories considered are shared in both approaches.

A. AN INTEGRATED CATEGORIZATION OF XAI METHODS

In this sub-section, we will analyze the findings in the literature directly related to a categorization of XAI methods in terms of supporting an integrative taxonomy. Considering only the most relevant and frequent categories found, we propose three main categories: Stage, Model, and Scope. Table 4 shows these categories along with the works that fully support this division, thus excluding, for instance, the approach found in [18], where the authors contemplate only model-agnostic methods and not model-specific ones. The summary table, Table 4, helps to strengthen the argument for a more straightforward and concise taxonomy.

A ''post-hoc'' XAI method is named after the fact that it acts after predictions are made, not knowing how the predictor model made its decisions (e.g., LIME ([24]). It is a surrogate model since it tries to simplify the function of the black-box model by sampling, perturbing data, and weighing the distance between instances to generate an approximation of the black-box model. ''ante-hoc'' techniques, such as Decision Trees and more specifically, the CART technique ([25]) as used in ML, derive their explainability from their clear approach and logic: a tree where an internal node (attribute) is split based on a specific condition. While the complexity of such a model can become large, thus suffering in terms of interpretability by displaying many nodes and depth, it is always possible to inspect the first levels where the most relevant decisions are made.

These findings suggest our first category, Stage, that indicates if the method is used after the prediction is made - post-hoc - or if the XAI model is intrinsically explainable - ante-hoc. We find evidence for this category in [6], [8], [9], [10], [11], [13], [15], [22], [23].

Some works do not make this distinction clearly, as is the case with the approach found in [18], where intrinsically explainable methods are detailed, such as Linear Regression or kNN, a technique initially proposed by Hodges and Fix [26] and since then widely used in Machine Learning, but post-hoc methods are not presented in the same capacity. The authors conclude that Linear Regression and kNN methods can be applied to complex problems but are inadequate for understanding ML models [26]. Barredo Arrieta et al. [23] define a taxonomy based on the reviewed literature. Contrary to taxonomies on previously mentioned works, where no general order of importance is mentioned, this work presents a hierarchical structure. The first level of the taxonomy tree, with ante-hoc models being referred to as ''Transparent Models'' and post-hoc models as ''Post-Hoc Explainability,'' can be encompassed into the Stage category.

The second category we propose is Model, referring to whether an XAI method is defined for a single or restricted group of models, that is, if it is model-specific, or if the method can be applied generally to any predictive model, that is, is model-agnostic. Evidence for this category can be found in [6], [8], [10], [12], [13], [15], [22], [23].

Model-specific techniques tend to be the most well-known and established models, like in the case of Decision Trees. The intrinsic explainability of this model is one of its disadvantages when compared to the performance of a neural network. While model-specific methods can be great as they have the unique ability to access the predictive model's internals, they suffer greatly in terms of interoperability due to their lack of adaptation for a more general usage [12].

Model-agnostic methods, such as LIME [24], are the opposite. Its general purpose makes it suitable for any predictive model, as shown by the authors, that present examples of explanations of predictive models, such as SVM (as defined in [27]) for text classification. We can find evidence for Model as a category in [9] and [15], where this categorization is proposed as a subcategory of the type post-hoc category. However, for Linardatos et al. [10], this category is named ''Model Specific vs. Model Agnostic'' and is presented in a non-hierarchical taxonomy. The same is seen in the work of Molnar [22] and Sahakyan [12], named ''Model-specific or Model-agnostic.'' On a different approach, the authors of [18] only explored model-agnostic approaches and not model-specific ones.

The final proposal for a category for XAI methods is Scope, intending to separate XAI methods on whether they are used to help understand the general behavior of the model, that is, if these techniques provide global interpretability or if they try to explain singular or a limited group of instances of data, that is, local interpretability [6]. This category is largely accepted within the reviewed literature, where it is found as a main category for classifying XAI methods [6], [8], [10], [14], [15], [22].

Local interpretability encapsulates methods, such as LIME, that introduce explainability by choosing relevant features, along with the features' respective importance, for a subset of the data to help understand singular instances of data. Global interpretability techniques focus on explaining the behavior of the model. One such example is SHAP ([28]), which returns a graphical importance of the

used features [22]. In some of the works found only local explanations are mentioned, like in the example of [23], or where an XAI taxonomy is explicitly stated and Scope is considered as being a sub-class of the model-agnostic class [11], [13], [18].

The three previous categories - Stage, Model, and Scope - were presented based on what the relevant literature shows as most generally used for the reviewed taxonomies to categorize XAI methods. Nonetheless, there are a couple more relevant categories to discuss, as they might be studied more in-depth by other authors, thus gaining the relevancy necessary to become a main category in the near future.

Molnar [22] points out "Result" as a category where importance is given to how the output of the XAI method is categorized. The author points out several possible subclasses, from feature summary statistics and feature importance to data points. This category is a contender for relevancy when defining a taxonomy since other authors support this category, even if under different names [15]. Another work favoring the categorization of results is [14], although this category is named "Presentation Format," showing two sub-classes on whether the generated explanation is textual (when explanations are generated using natural language techniques) or visual, focusing on providing a visual explanation, for example, via graphs or images. We can also find "Result" among other categories mentioned in [23].

Some authors consider "Output Format" a proper category for XAI methods. This classification is somewhat similar to the Result category, but in [15], we can find a difference between these two: while the Result class categorizes the explanation about the type of result provided, the Output Format looks at whether the explanation is of a particular type of data, such as numeric, textual, visual, among others. Such a difference is deemed relevant to define the purpose of the explanation for the different stakeholders, i.e., to whom the explanation is intended [8], [9].

In [10], the category "Purposes of Interpretability" is defined as "the purpose that these methods were created to serve and the ways through which they accomplish this purpose." The authors propose four subcategories for Purposes of Interpretability. Two of these categories, intrinsic and post-hoc, serve as references to the category Stage as previously stated in this section. However, in this case, these categories are inserted as sub-classes in the 'Purpose' category to explain complex black-box models, or post-hoc purpose, and to create white-box models, following an ante-hoc or intrinsic purpose. However, another author separates this purpose into two sub-classes, one explaining how something works and another explaining why something happened [14].

Other categories try to include stakeholders, i.e., to whom the explanation will serve. Hu et al. mention three types of users: developers, the ones who build the algorithm; observers, typically those who examine the system in place; and finally, end-users, people who are affected by the systems' results [14]. Another category proposed by the same authors is "Domain," which defines the subject area or domain for which XAI explanations are generated. Another category, 'Functioning,' is referred to by the authors of [15] to categorize how information is extracted from ML models. For instance, some XAI methods focus on perturbations of the data to gain insights for their explanatory process. In contrast, others focus on leveraging structures, which tend to result in feature importance attributes, among other sub-classes.

One last emerging category is "Type of Problem," which defines for what purposes the XAI method is useful to cover (classification or regression problems) and can be found in [9], [15].

V. LITERATURE REVIEW AND ANALYSIS OF PRACTICAL APPLICATIONS

As depicted in Table 1, XAI methods have gained much traction over the past few years. This section will explore findings related to applications of XAI restricted to the financial sector, with particular emphasis on credit-related problems and fraud detection. However, the latter is significantly less explored, as remarked earlier in Section III. The following section presents the specific applications of XAI methods in the financial domain found in our search, starting with a brief description and presenting a table summarizing the XAI method with examples of applications.

In general, SHAP tends to be one of the most widely used XAI methods for this domain. SHAP is a model-agnostic technique that can provide explanations both on a local and global scope. However, we can find slight differences in how it is implemented, with a mixture of studying feature importance with clustering and decision trees [29] or a simple application of the method on predictions [30], [31]. Some works follow a more complex approach, with a detailed procedure on how data treatment is made and the phases related to the prediction/explanation, culminating in explanations given by a sequence of steps, like feature selection followed by clustering [32]. One approach combines counterfactual explanations with SHAP [41]. The feature importance provided by SHAP is used to present counterfactual explanations in a localized region of the data, resulting in a more detailed explanation than simply using either method independently. This method is model-agnostic and works on the local scope.

SHAP is not the only popular method used, with LIME also being a popular choice. Both methods differ in the Scope category, as SHAP is mostly used globally, while LIME tends to be used locally. Overall, the value in the explanations of SHAP and LIME comes in the form of feature importance, where calculations are made to determine the weight in the contribution that features bear for the prediction process. Some articles mentioned employing both these XAI methods to explain the models used [37], [38]. In summary, LIME is a model-agnostic approach that presents explanations on a local scope.

While SHAP and LIME employ explanations in the form of feature importance, counterfactual methods create explanations for predictive models through the generation of what-if examples where certain feature values are changed to alter

the predicted result [22]. Regarding counterfactual methods, PermuteAttack was found in the manual search for practical applications [5]. This method uses a genetic algorithm that perturbs data by changing randomly selected features and goes through an optimization process to find an instance with the least number of permuted features, resulting in a counterfactual explanation. Another counterfactual method was found in reference [54], where a genetic algorithm is also implemented to produce explanations. As for the optimization process, it works only with features showing a correlation with the target, and for each iteration, the distance between the counterfactual example and the original instance is constrained. The explanation comes in visual explanations, showing what features needed changes to alter the prediction. PermuteAttack is a model-agnostic approach and provides explanations on a local scope.

One widely used method for explainability is Partial Dependence Plots or PDP [34], which helps interpret how one feature affects another. This aids in the explanation for the target feature, where the visual representation of this plot makes this relationship more understandable. PDP can be implemented regardless of the predictive model used and provides explanations in a global scope.

PASTLE [49] and CASTLE [50] are two other methods created by the same authors. The first method introduces explainability by reducing the sample space into pivots or points. In contrast, the second identifies clusters in the data with common behavior and classification, finalizing with the extraction of rule-based explanations. Both methods are model-agnostic and provide explanations on a local basis.

Anchors [51] is a model-agnostic method that provides explanations on a local scope by calculating the predicates or rules most relevant for the predictive outcome. It is an iterative process, starting with a general approach and finalizing in a filtered set of the most relevant rules presented as if-then clauses.

Specifically for deep-learning methods, MANE [52] works by processing features to extract cross features and linear regression is then applied to approximate the nonlinear decision boundary or the curve that separates two classes of data. This aids in understanding behavioral patterns of the data instances, resulting in a model-agnostic method on a local scope.

There are two model-specific approaches, LTreeX [56] and inTrees [57]. LTreeX is a local method that creates a surrogate model directly from the Random Forest, presenting rules that explain the outcome of any given instance. inTrees, on the other hand, is a global method that provides explanations by extracting rules from tree ensembles such as Random Forests or boosted trees.

DALE [58] is an XAI method that makes the calculations made by Area of Local Effects or feasible through an approximation of ALE. Similarly to the explanations provided by PDP, DALE's explanations come in the form of plots where it is possible to see the effect a feature has on the target. DALE is a model-agnostic technique and presents explanations in a global scope.

The final method for XAI found in the literature review is a model-agnostic approach where TREPAN trees are combined with neural networks to explain localized instances [46]. After clustering the data using a neural network, TREPAN is applied to build decision trees on a cluster level, resulting in explanations of the target feature by sets of rules defined by the trees. This hybrid model works on any predictive model and locally in terms of its scope.

Table 5 summarizes the XAI methods found and their respective categorization based on the taxonomy defined in Section IV. An obvious conclusion is that all the methods used in the financial domain are post-hoc, with their explanations formed after the predictions have been made. However, it is important to stress this distinction, as there are XAI methods that do not work on a post-hoc basis, such as the Decision Trees, where the method is not only explanatory in the way decisions are made for the prediction process but the method itself predicts the outcome in question. Therefore, these methods are intrinsically explanatory and, thus, are ante-hoc methods, and our searches only targeted post-hoc explanability.

Next, we describe the practical applications found in the literature review.

Hastie et al. [33] introduced explainability for the prediction of financial distress through XAI methods such as SHAP, PDP [34], and Counterfactuals [22]. In another work, using a dataset containing data from Chinese companies, Zhang et al. introduce Counterfactuals on the three most important features, analyzed via SHAP, where the specific instance of data has its feature values changed. Through a cyclical prediction process, a check is made on the variation prediction to see if its result has changed [35]. Some other works focus on explainability by combining LIME and SHAP applied to predictive models, such as Random Forests and XGBoost. Mandeep et al. worked with a dataset from Yahoo Finance companies' shares, filtered for the most relevant companies [36]. The authors combined the excellent predictive performance with intuitive explanations from LIME and SHAP to support the prediction results. Park et al. [39] investigated reliable prediction explanations for the predictive model built using XGBoost applied to a Korean company's dataset containing 110 features. For evaluating the reliability of LIME, they analyzed, instance by instance, the number of features present for the top ten most important instances when LIME was applied globally to the entire dataset. Another application using LIME to explain the predictions made by a Multi-Layer Perceptron on a transactions dataset can be found in [40].

In Watson's work, the Rational Shapley Values were introduced [41]. This hybrid method uses Shapley values and Counterfactuals, built to reap the benefit from both methods. The process was tested using the German Credit dataset.

On a different note, Hadash et al. focused on improving current implementations of LIME and SHAP methods [42]

TABLE 5. Categorization of XAI methods.

XAI Method	Author	Stage	Model	Scope
SHAP	[28]	Post-hoc	Agnostic	Global/Local
LIME	[24]	Post-hoc	Agnostic	Local
Counterfactuals	[5], [54]	Post-hoc	Agnostic	Local
PDP	[34]	Post-hoc	Agnostic	Global
PASTLE	[49]	Post-hoc	Agnostic	Local
CASTLE	[50]	Post-hoc	Agnostic	Local
Anchors	[51]	Post-hoc	Agnostic	Local
MANE	[52]	Post-hoc	Agnostic	Local
LTreeX	[56]	Post-hoc	Specific	Local
inTrees	[57]	Post-hoc	Specific	Global
DALE	[58]	Post-hoc	Agnostic	Global
Rational Shapley Values	[41]	Post-hoc	Agnostic	Local
TREPAN/Hidden-layer-clustering	[46]	Post-hoc	Agnostic	Local

with an experiment performed on a credit dataset where they used 133 users to evaluate the transformations. The improvements focused primarily on semantic changes to make the explanations given by these methods more understandable [42].

Another application proposes the implementation of 2DCNN (Convolutional Neural Networks), typically used for image-related problems, to tabular data. This process partitions the German Credit Dataset into bins, which are then used to create images. Based on these images, the model made its predictions. Subsequently, LIME and SHAP were used to explain such predictions, where the authors determined that SHAP performance was superior to the one obtained with LIME [43].

Analyzing some more general applications of XAI to Finance, we can find an approach that applied SHAP for the explainability of the model to determine the most used features and using loan data that was reviewed using NLP (Natural Language Processing) techniques [44].

De et al. proposed the combination of TREPAN [45] and hidden-layer clustering to explain predictions made using a credit dataset and for the predictive goal of determining a default in payment. This method was compared with LIME, and the authors concluded that the TREPAN model outperforms LIME [46].

Huynh et al. focused on implementing a framework to answer questions mainly motivated by legal regulations such as GDPR [47]. One of the inquiries relates to the fact that the final decision is ''reached solely via automated means,'' which helps determine whether Article 22 of the GDPR is applicable. The authors worked on a loan scenario, concluding that their developed framework successfully answered eight of the 13 questions that explain the decisions in the loan scenario, encompassing individual concerns or the individual data subject, and institutional concerns or the data controller. While the selection process of features is clear and explained, one of the limitations of this paper is that the ML algorithm itself is not explained [47].

Chromik implements SHAP onto the predictive model XGBoost to create an interface for personal loan applications [48]. This experiment shows mixed results when tested through user queries, with the users finding the interface overwhelming due to the presentation of several types of explanations calculated through SHAP. However, complementing the experiment through several different elements made it possible to determine that the explanations were detailed enough to understand the system's behavior in a prediction scenario.

A novel XAI method, PASTLE, was introduced by Gatta et al. and used to decrease the dataset to the points representing regions where the predictive model behaves differently. As for the data used, many experiments were performed, including the use of a financial dataset [49]. The same authors also developed another XAI method, CASTLE, whose main difference is what is used to decrease the number of instances used. While the first method uses pivots, the new method utilizes clustering [50]. Compared to Anchors [51], the authors found it less taxing on computational resources.

While the applications seen so far are primarily model-agnostic, the authors of [52] propose an XAI method specifically for deep learning models called MANE. Using a dataset of private transactions, they evaluated the proposed method against LIME, concluding that the performance was similar for both, albeit with a slight difference when compared with the proposed approach. When testing the fidelity, i.e., the degree of correctness of selected features of MANE, only five features were used to create the explanations contrasting with LIME, which needed 25 features for the same goal [52].

Lesser-known approaches use Feature Importance and Partial Dependence Plots to improve the interpretability of the predictive model, like in the case of XGBoost [53]. In another approach, the authors utilize an XAI method they developed to create counterfactual explanations, resulting in a low number of features needed to change the given outcome [54]. In [55], using the Home Equity Line of Credit (HELOC) dataset, the authors extended Shapley Values to mixed features without assuming them to be independent, concluding that no model outperformed the others.

Dedja et al. implemented another method, LTreeX, testing it over several datasets, although none was described as a financial dataset [56]. Nonetheless, this very recent approach deserves to be evaluated for possible implementation in the financial domain since the value of the explanation comes from the summarization of Random Forests, a common technique employed in modeling. In this regard, Deng explains Random Forests and Boosted Trees by expanding on known methods such as Area of Local Effects ([58]), even if not for the specific area of Finance [57]. Within the related literature, we can also encounter different implementations of counterfactuals [59], [60], [61] and a similar approach presenting a combination of Linear Regression and Neural Networks in order to explain the predictions [62].

TABLE 6. Recent applications of XAI methods.

XAI Method	Works that make use of the method
SHAP	[29-33, 35-38, 41, 43, 48, 55]
LIME	[36-40, 43, 46, 52]
Counterfactuals	[33, 35, 41, 59-61]
Hybrid models	[41, 45]
CERTIFAI	[53, 54]
Others	[33, 49, 50-52, 56, 57]

The summary table below (Table 6) describes the most predominant XAI methods emerging from this literature review, ordered by descending popularity. SHAP is the most popular method, referred to in most of the works reviewed here. The novel approaches, such as the LTreeX defined in [56], are placed in the category 'Others', encompassing several more recent and thus less used methods.

A. DATASETS

Finally, when reviewing related work, it is essential to discuss the datasets used. All the datasets found are from the financial domain, but only a handful are publicly available. One of the publicly available datasets is the German Credit[8] dataset, which contains 21 features and 1000 instances. This dataset encompasses clients' financial information and is used to predict the risk posed when credit is granted [32], [41], [43], [54], [57], [59], [61]. Another public dataset is the Default of Credit Card Clients in Taiwan,[9] containing 30,000 samples (customers) and information on 25 variables related to credit, default, billing, payments, and demographic factors [46], [60], [62]. Three more datasets that were publicly available were found. The first dataset,[10] used in [59], [61], aims to predict whether a person makes over $50,000 a year, containing 14 features and 48,842 records. The second[11] is an anonymized credit card transactions dataset where the target is to determine whether a transaction is legitimate. This second dataset, used in [37], [40], has 31 features and 284,807 transactions. Finally, the third dataset was found only in [62] and aims to determine the probability that a person will experience financial distress in the next two years. This dataset contains 11 features and totals 251,503 records.

VI. CONCLUSION

This paper reviews existing literature on applying XAI methods with a focus on works pertaining to the financial domain. First, a search was made exclusively for surveys relating to XAI. A second search was performed to discover practical applications of XAI specifically for finances. Identifying the lack of standardized knowledge in the area was also possible, with different authors proposing differing categorizations for XAI methods. From the data obtained with both searches, we could point out the major categories for XAI methods. This study highlights the main method's characteristics in the existing literature and proposes a unified yet simple taxonomy for XAI methods.

In a second contribution, we present the methods that are being used to achieve explainability for models applied in the financial sector. The existing literature favors SHAP and LIME as the preferred explainability methods. The applications found show that it is possible to use different methods simultaneously. This is due to the fact that most of the XAI techniques analyzed are applied post-hoc, allowing them to work independently and be used together. Though the popularity of LIME and SHAP in this domain seems to prevail, numerous new approaches are being proposed, broadening the spectrum of XAI methods available, from counterfactual explanations to partial dependence plots or more novel approaches that repurpose techniques used in image classification for tabular data. This work reflects the current understanding of the state-of-the-art regarding XAI methods in financial applications and presents a solid proposal for categorizing the existing XAI methods. Nevertheless, due to the recent rise in the investigation of explainable methods for artificial intelligence applications, it is expected that new developments will arise in the near future, paving the way for new anthological descriptive research to emerge.

A. CHALLENGES AND LIMITATIONS

There are limitations to this study, mainly because the survey about practical applications focused on the area of Finance. This ultimately means that the study is customized for this area of knowledge, and although it is possible to take advantage of it in similar areas, the same cannot be said for other areas (e.g., Health and Genetics). Another limitation is the number of studies reviewed here. The specific filters used in Section II implied that only 50 out of the 2069 initial papers were indeed read and analyzed. Most of these papers concern practical applications and focus on a specific subject. Although appropriate for the purposes of the study, the application of the specific filters may have left out some methods.

The main challenge encountered in this work was the lack of consensus regarding the general taxonomy of XAI methods. While some authors presented an in-depth categorization of XAI techniques, others presented a more simplified taxonomy. This can have a negative impact on the research field, as authors proposing new XAI methods may find it challenging to categorize them properly, making them more obscure to an XAI researcher.

Although it has been possible to overcome the lack of consensus on a general taxonomy by providing a categorization that uses part of each of the studies, it will only be with the advancement of the research area that the taxonomy will be put to the test, and whether or not it is inclusive enough for the categorization of newer XAI methods. Further, due to the limitations presented previously, only with the progress of XAI as a research area will the taxonomy be tested regarding its applicability in differing knowledge areas.

[8]https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[9]https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
[10]https://archive.ics.uci.edu/dataset/2/adult
[11]https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

REFERENCES

[1] M. Attaran and P. Deb, "Machine learning: The new 'big thing' for competitive advantage," *Int. J. Knowl. Eng. Data Mining*, vol. 5, no. 1, p. 1, 2018, doi: 10.1504/IJKEDM.2018.10015621.

[2] M. Turek. *Explainable Artificial Intelligence (XAI)*. Accessed: Jul. 3, 2023. [Online]. Available: https://www.darpa.mil/program/explainable-artificial-intelligence

[3] European Union, Official Journal of the European Union. (2016). *Regulation 2016/679, The Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*. Accessed: Dec. 16, 2022. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[4] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, no. 7, Mar. 2021, doi: 10.1136/bmj.n71.

[5] M. Hashemi and A. Fathi, "PermuteAttack: Counterfactual explanation of machine learning credit scorecards," Aug. 2020, *arXiv:2008.10138*.

[6] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[7] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: An analytical review," *WIREs Data Mining Knowl. Discovery*, vol. 11, no. 5, Sep. 2021, Art. no. e1424, doi: 10.1002/widm.1424.

[8] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks," *Appl. Sci.*, vol. 12, no. 3, p. 1353, Jan. 2022, doi: 10.3390/app12031353.

[9] G. Vilone and L. Longo, "Explainable artificial intelligence: A systematic review," May 2020, *arXiv:2006.00093*.

[10] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.

[11] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: A comprehensive review," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3503–3568, Jun. 2022, doi: 10.1007/s10462-021-10088-y.

[12] M. Sahakyan, Z. Aung, and T. Rahwan, "Explainable artificial intelligence for tabular data: A survey," *IEEE Access*, vol. 9, pp. 135392–135422, 2021, doi: 10.1109/ACCESS.2021.3116481.

[13] K.-Y. Lin, Y. Liu, L. Li, and R. Dou, "A review of explainable artificial intelligence," in *Proc. IFIP Int. Conf. Adv. Prod. Manag. Syst.*, 2021, pp. 574–584, doi: 10.1007/978-3-030-85910-7_61.

[14] Z. F. Hu, T. Kuflik, I. G. Mocanu, S. Najafian, and A. Shulner Tal, "Recent studies of XAI—Review," in *Proc. 29th ACM Conf. User Model., Adaptation Personalization*, Jun. 2021, pp. 421–431, doi: 10.1145/3450614.3463354.

[15] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Jun. 2022, pp. 2239–2250, doi: 10.1145/3531146.3534639.

[16] J. M. Darias, B. Díaz-Agudo, and J. A. Recio-Garcia, "A systematic review on model-agnostic XAI libraries," in *Proc. Workshops 29th Int. Conf. Case-Based Reasoning (ICCBR-WS)*, Sep. 2021, pp. 28–29, doi: 10.5281/zenodo.5838263.

[17] A. Bertrand, R. Belloum, J. R. Eagan, and W. Maxwell, "How cognitive biases affect XAI-assisted decision-making: A systematic review," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2022, pp. 78–91, doi: 10.1145/3514094.3534164.

[18] A. Hanif, X. Zhang, and S. Wood, "A survey on explainable artificial intelligence techniques and challenges," in *Proc. IEEE 25th Int. Enterprise Distrib. Object Comput. Workshop (EDOCW)*, Oct. 2021, pp. 81–89, doi: 10.1109/EDOCW52865.2021.00036.

[19] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence," *IEEE Access*, vol. 9, pp. 11974–12001, 2021, doi: 10.1109/ACCESS.2021.3051315.

[20] P. Lopes, E. Silva, C. Braga, T. Oliveira, and L. Rosado, "XAI systems evaluation: A review of human and computer-centred methods," *Appl. Sci.*, vol. 12, no. 19, p. 9423, Sep. 2022, doi: 10.3390/app12199423.

[21] D. Kahneman, P. Slovic, and A. Tversky, Eds., *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, U.K.: Cambridge Univ. Press, 1982, doi: 10.1017/CBO9780511809477.

[22] C. Molnar, *Interpretable Machine Learning*, 2nd ed., Feb. 2022. Accessed: Jan. 8, 2023. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[23] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[25] D. H. Moore, "Classification and regression trees, by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Brooks/Cole Publishing, Monterey, 1984,358 pages, $27.95," *Cytometry*, vol. 8, no. 5, pp. 534–535, Sep. 1987, doi: 10.1002/cyto.990080516.

[26] J. L. Hodges and E. Fix, "Discriminatory analysis—Nonparametric discrimination: Consistency properties," USAF School Aviation Med., Randolph Field, TX, USA, Tech. Rep. 4, Feb. 1951. [Online]. Available: https://apps.dtic.mil/sti/pdfs/ADA800276.pdf

[27] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[28] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4766–4776. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[29] C. Maree, J. E. Modal, and C. W. Omlin, "Towards responsible AI for financial transactions," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2020, pp. 16–21, doi: 10.1109/SSCI47803.2020.9308456.

[30] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," *Comput. Econ.*, vol. 57, no. 1, pp. 203–216, Jan. 2021, doi: 10.1007/s10614-020-10042-0.

[31] S. Kim and J. Woo, "Explainable AI framework for the financial rating models," in *Proc. 10th Int. Conf. Comput. Pattern Recognit.*, Oct. 2021, pp. 252–255, doi: 10.1145/3497623.3497664.

[32] J. Chaquet-Ulldemolins, F.-J. Gimeno-Blanes, S. Moral-Rubio, S. Muñoz-Romero, and J.-L. Rojo-Álvarez, "On the black-box challenge for fraud detection using machine learning (II): Nonlinear analysis through interpretable autoencoders," *Appl. Sci.*, vol. 12, no. 8, p. 3856, Apr. 2022, doi: 10.3390/app12083856.

[33] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.

[34] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[35] Z. Zhang, C. Wu, S. Qu, and X. Chen, "An explainable artificial intelligence approach for financial distress prediction," *Inf. Process. Manage.*, vol. 59, no. 4, Jul. 2022, Art. no. 102988, doi: 10.1016/j.ipm.2022.102988.

[36] Mandeep, A. Agarwal, A. Bhatia, A. Malhi, P. Kaler, and H. S. Pannu, "Machine learning based explainable financial forecasting," in *Proc. 4th Int. Conf. Comput. Commun. Internet (ICCCI)*, Jul. 2022, pp. 34–38, doi: 10.1109/ICCCI55554.2022.9850272.

[37] I. Ullah, A. Rios, V. Gala, and S. Mckeever, "Explaining deep learning models for tabular data using layer-wise relevance propagation," *Appl. Sci.*, vol. 12, no. 1, p. 136, Dec. 2021, doi: 10.3390/app12010136.

[38] S. Tyagi, "Analyzing machine learning models for credit scoring with explainable AI and optimizing investment decisions," Sep. 2022, *arXiv:2209.09362*.

[39] M. S. Park, H. Son, C. Hyun, and H. J. Hwang, "Explainability of machine learning models for bankruptcy prediction," *IEEE Access*, vol. 9, pp. 124887–124899, 2021, doi: 10.1109/ACCESS.2021.3110270.

[40] T.-Y. Wu and Y.-T. Wang, "Locally interpretable one-class anomaly detection for credit card fraud detection," in *Proc. Int. Conf. Technol. Appl. Artif. Intell. (TAAI)*, Nov. 2021, pp. 25–30, doi: 10.1109/TAAI54685.2021.00014.

[41] D. Watson, "Rational Shapley values," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Jun. 2022, pp. 1083–1094, doi: 10.1145/3531146.3533170.

[42] S. Hadash, M. C. Willemsen, C. Snijders, and W. A. IJsselsteijn, "Improving understandability of feature contributions in model-agnostic explainable AI tools," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2022, pp. 1–9, doi: 10.1145/3491102.3517650.

[43] X. Dastile and T. Celik, "Making deep learning-based predictions for credit scoring explainable," *IEEE Access*, vol. 9, pp. 50426–50440, 2021, doi: 10.1109/ACCESS.2021.3068854.
[44] A. Stevens, P. Deruyck, Z. V. Veldhoven, and J. Vanthienen, "Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2020, pp. 1241–1248, doi: 10.1109/SSCI47803.2020.9308371.
[45] M. Craven and J. Shavlik, "Extracting tree-structured representations of trained networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 8, 1995, pp. 24–30. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1995/file/45f31d16b1058d586fc3be7207b58053-Paper.pdf
[46] T. De, P. Giri, A. Mevawala, R. Nemani, and A. Deo, "Explainable AI: A hybrid approach to generate human-interpretable explanation for deep learning prediction," *Proc. Comput. Sci.*, vol. 168, pp. 40–48, Jan. 2020, doi: 10.1016/j.procs.2020.02.255.
[47] T. D. Huynh, N. Tsakalakis, A. Helal, S. Stalla-Bourdillon, and L. Moreau, "Addressing regulatory requirements on explanations for automated decisions with provenance—A case study," *Digit. Government, Res. Pract.*, vol. 2, no. 2, pp. 1–14, Apr. 2021, doi: 10.1145/3436897.
[48] M. Chromik, "Making SHAP Rap: Bridging local and global insights through interaction and narratives," in *Proc. IFIP Conf. Hum.-Comput. Interact.*, 2021, pp. 641–651, doi: 10.1007/978-3-030-85616-8_37.
[49] V. La Gatta, V. Moscato, M. Postiglione, and G. Sperlì, "PASTLE: Pivot-aided space transformation for local explanations," *Pattern Recognit. Lett.*, vol. 149, pp. 67–74, Sep. 2021, doi: 10.1016/j.patrec.2021.05.018.
[50] V. La Gatta, V. Moscato, M. Postiglione, and G. Sperlì, "CASTLE: Cluster-aided space transformation for local explanations," *Expert Syst. Appl.*, vol. 179, Oct. 2021, Art. no. 115045, doi: 10.1016/j.eswa.2021.115045.
[51] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1527–1535, doi: 10.1609/aaai.v32i1.11491.
[52] Y. Tian and G. Liu, "MANE: Model-agnostic non-linear explanations for deep learning model," in *Proc. IEEE World Congr. Services (SERVICES)*, Oct. 2020, pp. 33–36, doi: 10.1109/SERVICES48979.2020.00021.
[53] Y. Zou, C. Gao, and H. Gao, "Business failure prediction based on a cost-sensitive extreme gradient boosting machine," *IEEE Access*, vol. 10, pp. 42623–42639, 2022, doi: 10.1109/ACCESS.2022.3168857.
[54] X. Dastile, T. Celik, and H. Vandierendonck, "Model-agnostic counterfactual explanations in credit scoring," *IEEE Access*, vol. 10, pp. 69543–69554, 2022, doi: 10.1109/ACCESS.2022.3177783.
[55] A. Redelmeier, M. Jullum, and K. Aas, "Explaining predictive models with mixed features using Shapley values and conditional inference trees," in *Proc. Int. Cross-Domain Conf. Mach. Learn. Knowl. Extraction*, 2020, pp. 117–137, doi: 10.1007/978-3-030-57321-8_7.
[56] K. Dedja, F. K. Nakano, K. Pliakos, and C. Vens, "BELLATREX: Building explanations through a locally accurate rule extractor," Mar. 2022, *arXiv:2203.15511*.
[57] H. Deng, "Interpreting tree ensembles with inTrees," *Int. J. Data Sci. Anal.*, vol. 7, no. 4, pp. 277–287, Jun. 2019, doi: 10.1007/s41060-018-0144-8.
[58] V. Gkolemis, T. Dalamagas, and C. Diou, "DALE: Differential accumulated local effects for efficient and accurate global explanations," Oct. 2022, *arXiv:2210.04542*.
[59] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and counterfactual explanations for black box decision making," *IEEE Intell. Syst.*, vol. 34, no. 6, pp. 14–23, Nov. 2019, doi: 10.1109/MIS.2019.2957223.
[60] A. White and A. D. Garcez, "Measurable counterfactual local explanations for any classifier," Aug. 2019, *arXiv:1908.03020*.
[61] E. Ç. Mutlu, N. Yousefi, and O. Ozmen Garibay, "Contrastive counterfactual fairness in algorithmic decision-making," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2022, pp. 499–507, doi: 10.1145/3514094.3534143.
[62] D. Chen, W. Ye, and J. Ye, "Interpretable selective learning in credit risk," Sep. 2022, *arXiv:2209.10127*.
[63] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Turin, Italy, Oct. 2018, pp. 80–89, doi: 10.1109/DSAA.2018.00018.
[64] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, May/Jun. 2018, doi: 10.1145/3236386.3241340.
[65] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," Mar. 2017, *arXiv:1702.08608*.

TIAGO MARTINS received the B.Sc. degree in computer science and the M.Sc. degree in data science from Instituto Universitário de Lisboa (ISCTE), Lisbon, Portugal, in 2021 and 2023, respectively.

ANA MARIA DE ALMEIDA (Senior Member, IEEE) received the Ph.D. degree. She is currently an Associate Professor with the Information Science and Technologies Department, Instituto Universitário de Lisboa (ISCTE), and a Researcher with ISTAR-IUL—Information Sciences, Technologies and Architecture Research Center, where she coordinates the Software Systems Engineering Research Group, and the Cognitive and Media Systems Research Group, Center for Informatics and Systems of the University of Coimbra (CISUC). Her current research interests include the areas of algorithmic, complexity, machine learning and pattern recognition, data science, evolutionary computation, and ethics for AI and research. She has a particular interest in the development of self-adjusting predictive and reactive models for real applications, and in evolutionary strategies for tackling multicriteria combinatorial problems. She is a member of ACM.

ELSA CARDOSO (Member, IEEE) received the B.Sc. degree in electrical and computer engineering and the M.Sc. degree in computer science and engineering from Técnico, Universidade de Lisboa, in 1995 and 2003, respectively, and the Ph.D. (European) degree in information sciences and technologies from Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal. She is currently an Assistant (tenured) Professor with the Information Science and Technologies Department, ISCTE-IUL. She is also a Researcher with Centro de Investigação e Estudos de Sociologia (CIES-ISCTE) and the Information and Decision Support Systems Group, INESC-ID Lisbon, Portugal. She has participated in several national and international research projects.

LUÍS NUNES was born in Lisbon, Portugal. He received the B.Sc. degree from Universidade de Lisboa, Lisbon, in 1993, the M.Sc. degree from Instituto Superior Técnico, Lisbon, in 1997, and the Ph.D. degree from Faculdade de Engenharia da Universidade do Porto, in 2006. He was a Researcher with INESC (1992–1997), Adetti (1997–2001), LIACC (2001–2011), IT (2010–2021), and ISTAR (since 2015). He joined Departamento de Ciências e Tecnologias da Computação, ISCTE, as a Teaching Assistant, in 1997, teaching mainly programming and machine learning courses. He is currently an Associate Professor with Instituto Universitário de Lisboa (ISCTE), Lisbon. His published work [journals (18), conferences (37) and book-chapters (seven)] and projects (seven) are mainly in the area of machine learning and its applications.