IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 May 2024, accepted 3 June 2024, date of publication 5 June 2024, date of current version 14 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3409843

TOPICAL REVIEW

EXplainable Artificial Intelligence (XAI)—From Theory to Methods and Applications

EVANDRO S. ORTIGOSSA, THALES GONÇALVES, AND LUIS GUSTAVO NONATO, (Member, IEEE)
Institute of Mathematics and Computer Science, University of São Paulo (ICMC-USP), São Carlos 13566-590, Brazil

Corresponding author: Luis Gustavo Nonato (gnonato@icmc.usp.br)

This work was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES)–Finance Code 001, in part by São Paulo Research Foundation (FAPESP), and in part by National Council for Scientific and Technological Development (CNPq). Evandro S. Ortigossa was supported by CAPES. Thales Gonçalves was supported by FAPESP under Grant 2018/24516-0. Luis Gustavo Nonato was supported by FAPESP under Grant 2022/09091-8 and CNPq under Grant 307184/2021-8.

ABSTRACT Intelligent applications supported by Machine Learning have achieved remarkable performance rates for a wide range of tasks in many domains. However, understanding why a trained algorithm makes a particular decision remains problematic. Given the growing interest in the application of learning-based models, some concerns arise in the dealing with sensible environments, which may impact users' lives. The complex nature of those models' decision mechanisms makes them the so-called "black boxes," in which the understanding of the logic behind automated decision-making processes by humans is not trivial. Furthermore, the reasoning that leads a model to provide a specific prediction can be more important than performance metrics, which introduces a trade-off between interpretability and model accuracy. Explaining intelligent computer decisions can be regarded as a way to justify their reliability and establish trust. In this sense, explanations are critical tools that verify predictions to discover errors and biases previously hidden within the models' complex structures, opening up vast possibilities for more responsible applications. In this review, we provide theoretical foundations of Explainable Artificial Intelligence (XAI), clarifying diffuse definitions and identifying research objectives, challenges, and future research lines related to turning opaque machine learning outputs into more transparent decisions. We also present a careful overview of the state-of-the-art explainability approaches, with a particular analysis of methods based on feature importance, such as the well-known LIME and SHAP. As a result, we highlight practical applications of the successful use of XAI.

INDEX TERMS Black-box models, explainability, explainable machine learning, interpretability, interpretable machine learning.

I. INTRODUCTION

Machine Learning models are utilized in our daily lives. The increased attention to intelligent-based applications is owing to the unprecedented performance levels that modern learning models have achieved, solving high-complexity tasks that place Artificial Intelligence at the center of many domains and activities in which technology has been a transforming factor [1]. The concept of Artificial Intelligence is not a novelty, with its roots dating back several decades, among the first steps in computer science, following the old dream of creating "intelligent" machines that are able to make decisions similar to human thought. The adoption of computer models that learn patterns and predict new data is at a remarkable moment of ubiquity that, if properly explored, can produce remarkable results within the many domains where those systems are applied [2]. Therefore, Artificial Intelligence is regarded as a fundamental tool to accelerate future advances in the development of a more algorithmic and digital society [3], [4].

Machine Learning is a branch of Artificial Intelligence that actively uses scientific knowledge, such as mathematics, physics, biology, statistics, linguistics, and psychology, to simulate the cognitive abilities of human intelligence

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

computationally. Machine Learning has attracted considerable attention from the research community because of its ability to accurately predict a wide range of complex phenomena [5]. Many learning-based algorithms have emerged in the last decade, especially after 2012, owing to the significant reduction in data storage costs, thus increasing the amount of information available through large datasets [6], improvements in hardware, especially Graphics Processing Units (GPUs) with high computational power, enabling the processing of large datasets in reasonable times, and new programming languages and high-quality open-source libraries, which have leveraged worldwide programmers for creating prototypes, running and testing models, and developing new optimized algorithms [7].

The sophistication of learning-based algorithms has increased up to the point they have achieved (above) human performances [6], [8], even surpassing human abilities in several computer tasks, including computer vision, image classification, language processing, and pattern recognition [9], [10], [11]. Some intelligent systems require almost no human intervention for tuning or training [2]; consequently, the application of Machine Learning has been transformative, with intelligent models employed in most diverse contexts, from products, text documents, music, movies, and friend recommendations on social networks to decision-making in critical fields such as medicine, financial markets, autonomous cars, government strategic planning, bioinformatics, and criminal systems [2], [12], [13], [14], [15], [16]. On the other hand, such complex models also draw attention to trust-related problems, particularly when the outputs involve sensitive contexts, such as in medical diagnosis. In this case, the reasons behind a decision must be known [4].

A. THE TRANSPARENCY CHALLENGE

As stated by Breiman [17] when defining Random Forests algorithm, ''a forest of trees is impenetrable as far as simple interpretations of its mechanism go.'' Despite the high levels of accuracy, the complex nature in which learning models operate reduces the transparency of their decision processes, turning them into so-called ''black boxes'' [18]. In other words, modern learning algorithms suffer from opacity – describing the degree of impact of each part of the information provided as an input with respect to the corresponding output can be challenging [19]. Delegating critical decisions to systems that cannot be interpreted or do not provide explanations about the logical path of their outputs can be dangerous, especially in sensitive scenarios, such as healthcare, autonomous cars, public security, and counter-terrorism [4], [20]. Therefore, ''interpretability'' and ''explainability'' have emerged as new concepts brought to the surface in the Explainable Artificial Intelligence (XAI) research community.

Interpretable and explainable do not share the same meaning. The word ''interpretable'' can be defined as the ability to present something in an understandable manner [21]. Humans can justify their actions through logically consistent, describable, and understandable choices produced by their ability to ''think'' [10]. Except for the final output, the interpretation of the reasoning behind a complex machine learning model is not easy, thus preventing its results from being fairly understood [9]. However, if a decision cannot be directly interpreted, understandable elements that shed light on the opaque decision-making processes of the models can be provided, thus making them explainable. In this sense, explainability can advance toward more transparency in complex models, providing elements of explanation of the logic behind a prediction and debugging the simple presentation of output data. In the Machine Learning context, explainability can be can be considered a counterpart to the decision-making rationalization of human thought [10].

In general, all initiatives and efforts to reduce the complexity of learning-based models and improve both transparency and understanding of their actions can be considered XAI approaches [2], [22]. XAI is a research area that leverages ideas from the social sciences. It also considers the psychology of explanation to create techniques that make the outputs of machine learning applications more understandable while maintaining a high level of predictive performance, enabling humans to interpret, trust, and manage the next generations exposed to Artificial Intelligence [2], [23]. Interpretability and explainability are closely related in supporting humans in understanding the reasoning behind a model's predictions. Although they are often used interchangeably in the literature, they are not monolithic concepts, and their precise and formal definitions remain subjective in the specialized literature. No consensual specifications for what an interpretable algorithm would be or a proper way to generate and evaluate explanations have been reached [13].

The need to explain the behavior of non-interpretable learning algorithms that can affect people's lives is not only a desirable property, but also a legal demand in some places. As an example, the European Union introduced the right to explanation in its General Data Protection Regulation (GDPR), including algorithmic decision-making guidelines, to mitigate the social impact of computational systems [24]. Among other requirements, GDPR defines the right to information, i.e., the need for ''*meaningful explanations of the logic involved*'' in automated decisions requiring ''*the controller must ensure the right for individuals to obtain further information about the decision of any automated system*'' [25], [26].

GDPR started institutional discussions about more requirements for compliance with Artificial Intelligence use. The U.S. Food and Drug Administration (FDA) proposed a regulatory framework for medical devices supported by Artificial Intelligence/Machine Learning [27]. The framework defines the need for submission to the FDA evaluation when continuous learning algorithms introduce changes that significantly affect a medical device's performance. However, implementing those requirements for product development is still an open problem [28].

Similarly, the World Health Organization (WHO) released a long guidance report on the Ethics & Governance of Artificial

Intelligence for Health [29]. The WHO document identifies ethical, trust, and transparency challenges in designing or deploying intelligent-based models applied in healthcare. Specifically, the WHO guidance requires health "*technologies should be intelligible or understandable to developers, medical professionals, patients, users, and regulators*," with explainability as the approach to improve transparency and provide an understanding of why an intelligent system made a particular decision.

In addition, the recently introduced California Consumer Privacy Act (CCPA) [30] defines rights regarding use and protection of personal information, which has influenced privacy legislation in the United States. Therefore, explaining black-box decisions is now a legally mandatory desirable subject, motivating the recent explosion in XAI research interest and development techniques [4], [31].

Explanations of the reasons that lead an intelligent model to its discovered patterns, i.e., the reasoning behind predictions, can be even more important than the predictive performance itself [14]. In this sense, Explainable Artificial Intelligence can add a new layer to the undeniable success of Machine Learning, going beyond the usual performance metrics and aiming to provide a direct understanding of the behavior of learning models.

B. CONTRIBUTIONS AND ORGANIZATION

In recent years, XAI has become one of the most popular subjects in Artificial Intelligence and Data Science communities, and explaining machine learning is essential, since complex learning-based models are now part of our lives, making decisions that may influence people's interactions. However, XAI is not yet a mature research domain, often lacking formality in definitions and objectives [32]. In this paper, we investigate the multiple aspects of XAI, providing beginners and experts with the way the concepts of explainability translate into practical applications for understanding machine-learning decisions. With a comprehensive study of the XAI literature, we identified gaps and organized a detailed review of the theoretical foundations and objectives related to explainability research.

In contrast to previous studies that presented a large number of techniques, the present one discusses the latest and main applications devoted to opening black-box problems from different perspectives (e.g., locality or model dependence) and using different mechanisms (e.g., feature importance, inspection, or counterfactuals), highlighting their operational aspects, advantages, and limitations. We also carefully reviewed feature importance explainability methods due to their leading position among XAI approaches [33], with a detailed analysis of LIME and SHAP.

Our focus is on demonstrating the importance of XAI tools for providing an additional layer of trust to automated decision systems by detecting hidden biases and noises that can lead to unfair decisions. We also address the limitations of current XAI approaches and future research directions toward helping researchers design comprehensive explanations. As a result of our investigations, we report an overview of practical applications where XAI has been successfully applied to turn opaque decisions into more transparent information.

In summary, the main contributions of this research are:

- A comprehensive discussion on XAI theory, including motivations, terminology clarification, and objectives of explainability in Machine Learning.
- A concise review and taxonomic categorization of recent and widely used XAI methodologies.
- A presentation of the challenges, limitations, and promising paths toward explainability evolution.
- An in-depth review of feature attribution/importance methods, including an analysis of the problems related to relying on Shapley-based explanations.
- A high-level discussion of cases from various domains where explainability has been successfully applied.

The remainder of the paper is organized as follows: Section II defines the research methodology and basic terminology used; Section III presents some previous research; Section IV briefly overviews the evolution of Machine Learning; Section V is devoted to an algorithmic complexity discussion; Section VI addresses the objectives of explainability; Section VII clarifies the theoretical foundations of XAI and presents the needs, challenges, and a taxonomy; Section VIII reviews recent approaches in the XAI domain and Section IX provides examples of successful implementations of explainability; open problems and future research directions are discussed in Section X; finally, Section XI presents our final remarks.

II. BACKGROUND STATEMENTS

We conducted a content investigation of published literature to understand the evolution of XAI over the last few years. Such research systematically evaluated the available scientific communication, clarified terminologies, described objectives, identified fundamental contributions and applications, and indicated future research opportunities. XAI is a research area that emerged not long ago and still lacks some definitions and further discussions, addressed in this review.

A. METHOD OF THE SYSTEMATIC REVIEW

We combined four databases, namely, Association for Computing Machinery (ACM) Digital Library, IEEE Xplore Digital Library, Citeseer Library, and Elsevier's Scopus, for a comprehensive search of XAI theory and applications and search engines such as Google Scholar, Elsevier's ScienceDirect, and Thomson Reuters' Web of Science were used in association with them.

Queries on terms "explainable," "interpretability," "explainability," "black box," "understandable," and "transparency" merged with "artificial intelligence" or "machine learning" mainly restricted to (but not only) the 2010–2024 period and based on publications' title, abstract, and keywords were performed. The following two criteria

were employed in the search results for selecting publications for further revision:

- Papers published in relevant peer-reviewed scientific journals as articles available online in English. We extended the scope to conference proceedings, arXiv e-prints, theses, and books.
- Previous studies explicitly employing explainable artificial intelligence or explainable machine learning. We excluded papers that only listed XAI in keywords, alluded to XAI, or applied some XAI method with no discussions or a reference to the XAI methodology employed.

The queries returned 439 papers. We removed duplicates after fine-granulated filtering from abstract and introduction readings, carefully reviewed the remaining ones in full, and reported 296 references here.

B. BASIC TERMINOLOGY

A term often used in this research is ''model,'' which can denote diverse meanings in different areas. Some misunderstandings are still possible to occur, even when the scope is limited to Machine Learning. Although machine learning algorithms such as Artificial Neural Networks, SVMs, or Random Forests are not models, they generate models after training procedures. Therefore, the meaning of ''model'' must be defined in this research. Whenever used here, it is employed as a simplified reference to some machine learning algorithm or its generated (trained) model, which follows the usual terminology in XAI literature. In addition, when discussing any aspect of XAI approaches applied to explain learning models, we refer to a trained model.

The models addressed here are usually trained on multidimensional datasets (e.g., tabular data, time series, 2D images, 3D point clouds, videos, or semantic segmentation), which contain m individual instances composed of a collection of characteristics formally expressed as

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\} \tag{1}$$

where each vector $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbf{X}$ is a data instance in $\mathbb{R}^n$. By convention, x_i elements that characterize the instances in such a dataset are called attributes, features, or variables. On the other hand, multiple elements that compose machine learning models updated during learning procedures, i.e., those that vary when the model is trained, are called model parameters or simply parameters.

III. PREVIOUS WORK

Research on XAI has introduced a wide variety of approaches and methods so that several researchers have committed to discussing the XAI environment for defining multiple theoretical and practical particularities of techniques [34], [35] and metrics related to XAI [36], [37], [38], [39].

Lipton [13] was one of the pioneers in organizing the main definitions of interpretability in Machine Learning. Although the final publication dated 2018, its first version was available in 2016, compiling a discussion on the needs and motivations of interpretability according to the literature at the moment. Doshi-Velez and Kim [21], Chakraborty et al. [10], and Došilović et al. [7] introduced the concepts and taxonomy. Despite their valuable overviews introducing early advances in XAI, the studies were seminal in terms of concepts, lacking clear definitions of interpretability and explainability. Zhang and Zhu [9] reviewed XAI, restricting the research on visualization strategies applied for Convolutional Neural Networks (CNNs).

Miller [22] conducted a broad survey on XAI, approaching theories from human sciences such as cognitive and social psychology, which was foundational in relating Artificial Intelligence and how humans explain decisions. Guidotti et al. [26] analyzed several interpretable and explainability methods; they categorized them according to the problem type for which each XAI method was indicated and described a detailed taxonomy from Data Mining and Machine Learning viewpoints. Despite formalizing important concepts of XAI and focusing on interpretability processes, the authors only highlighted the need for evaluation metrics and did not discuss the elements of evaluation comprehensively. Similarly, Murdoch et al. [5] presented an updated conceptual overview and Molnar [40] offered an extensive review of both conceptual elements and characteristics of the main XAI approaches.

Adadi and Berrada [4] and Arrieta et al. [2] conducted comprehensive research in the literature and presented detailed views of the XAI scenario from the fundamentals, contributions, and toward solutions for dealing with the different needs for explainability. Both publications address ethical concepts such as fairness (in the sense of impartiality) and compliance in Machine Learning, differing in a critical aspect, with Adadi and Berrada [4] introducing questions on (the lack of) explainability evaluation and Arrieta et al. [2] distinguishing transparent and post-hoc methods and suggesting guidelines for the development of socially responsible intelligent systems. The two articles provided enriched discussions on XAI, but described the leading strategies in general terms. Linardatos et al. [32] defined a taxonomy of interpretability methods, concluding most XAI methods were proposed for tasks on Neural Network models. The authors also included links to code repositories with XAI implementations. Speith [41] critically reviewed several commonly adopted taxonomies of explainability methods, highlighting their similarities, differences, and inconsistencies.

Tjoa and Guan [42] and Amann et al. [16] discussed the concepts and applications of XAI; however, they concentrated their research on the explainability of black-box systems used in medicine. The two later studies addressed important matters on the risks of omitting clear explanations within medical applications, with Amann et al. [16] highlighting the need to fix the multiple XAI terminology and properly validate explanations.

Regarding specialized publications on explainability, several recent studies have reviewed XAI for applications in different domains where machine learning has been

used, highlighting medicine, specifically cardiology [43], breast cancer diagnosis and surgery [44], medical image analysis [45], radiology [46], and healthcare [28], [47], [48]. XAI reviews dedicated to other areas such as genetics [49], anomaly detection [50], automotive industry [51], [52], automation and smart industry [53], materials science [54], and language processing [55] can also be found, proving the interdisciplinary relevance of XAI research. Although those specialized surveys provide valuable insights into trends, advances, promises, and limitations of XAI under the view of domain experts, they focus on applications for specific contexts, which can limit the approach of the theory behind XAI methods.

In this section, we have presented valuable research reviewing various elements of XAI. Despite the rich literature, there is room for improvements in the latest studies. The literature on explainability is no longer in its early days; however, XAI is not yet a mature research field [32]. Machine Learning applications have evolved quickly in the last few years, pushing the need for more transparency. The early publications reviewing XAI have aged, since XAI has also grown fast.

Therefore, we carefully reviewed the current literature to identify gaps in definitions of leading approaches to filling them. In light of almost a decade of research on XAI, proposals of more concrete terminology are required. This study provides a theoretical foundation that differentiates the main concepts of XAI and supports the reader with a clear set of XAI definitions, challenges, goals, categorization, evaluation, and limitations. We do not perform a quantitative review, since previous studies have conducted them. In addition to conceptual discussions, we propose a deep and highly detailed analysis of the most recent and relevant approaches, especially feature importance/attribution methods such as LIME and SHAP, providing a comprehensive review of theory and practical applications of explainability to beginners and experts researching XAI.

IV. A BRIEF OVERVIEW OF MACHINE LEARNING

Artificial Intelligence and Machine Learning are two closely related research fields often associated with the development of intelligent computing systems. Despite a significant symbiosis between technologies and methods to the point the terms are sometimes used as synonyms, there are significant conceptual differences between them.

Artificial Intelligence is a multidisciplinary area of science with applications in many theoretical and practical domains. It focuses mainly on the development of systems that process data or information from the environment to which they are applied and, based on such environmental perception, execute a set of automated actions that best fit the previous knowledge toward achieving desired results [56]. In this sense, the so-called ''intelligent systems'' are those that can make decisions based on their judgment, similarly to the rationalization process for decision-making in human thought.

Artificial Intelligence techniques are traditionally divided into symbolic and connectionist. The symbolic (or classical) paradigm, prevalent until the 1980s, incorporates predicate logic based on symbols and rules representing human knowledge about a given problem. Symbols enable the algorithm to establish a series of logical reasoning processes similar to language. Symbolic representations have a propositional nature and define the existence of relationships between objects, whereas ''reasoning'' develops new logical relationships supported by a set of inference rules [57]. Note the similarity with the human reasoning process, which relates objects and abstract concepts and, from the knowledge acquired, creates association rules for generalizing when exposed to new settings.

An advantage of symbolic Artificial Intelligence is self-explainability, i.e., it is interpretable, enabling the extraction of explanation elements about the rational process, leading to the model's decisions [4]. A serious limitation of this paradigm is the need to define all necessary knowledge explicitly. Furthermore, representational elements must be formalized manually instead of being acquired from data [58]. Such a limitation makes the development of symbolic models a costly process, generally resulting in domain-specific systems, i.e., with low generalizability, which makes symbolic Artificial Intelligence currently considered obsolete.

In contrast, the connectionist paradigm emerged in 1959 with the concept of Machine Learning and Arthur Samuel defining it as ''a field of study that gives computers the ability to learn without being explicitly programmed'' [59]. Machine Learning focuses on computational methods that can acquire new knowledge, new skills, and ways of organizing existing knowledge [60]. It is formally defined as a collection of techniques that enables computers to automate the construction and programming of modeling by discovering and generalizing statistically significant patterns in available data [61]. In other words, machines learn tasks based on training models generated through data or previous experience and adapt themselves to new inputs to make predictions in human-like tasks. This is one of the main reasons why machine learning is widely employed across different domains.

According to the aforementioned definitions, every Machine Learning model is Artificial Intelligence – however, the latter covers a broader scope of techniques, i.e., not every Artificial Intelligence application belongs to the set of Machine Learning models. Although research on Machine Learning algorithms started several decades ago, much of its impacting contributions have been relatively recent owing to the intense development of new algorithms, especially after 2012. One of the reasons for the recent boom in Machine Learning development is the advent of high-performance computing technologies such as GPUs, which enable modern models to learn on large datasets at reasonable times.

This study focuses on the application of XAI to supervised learning, a training paradigm in which the dataset comprises a set of inputs and a known mapping of each input to a desired output. More specifically, in supervised learning, the

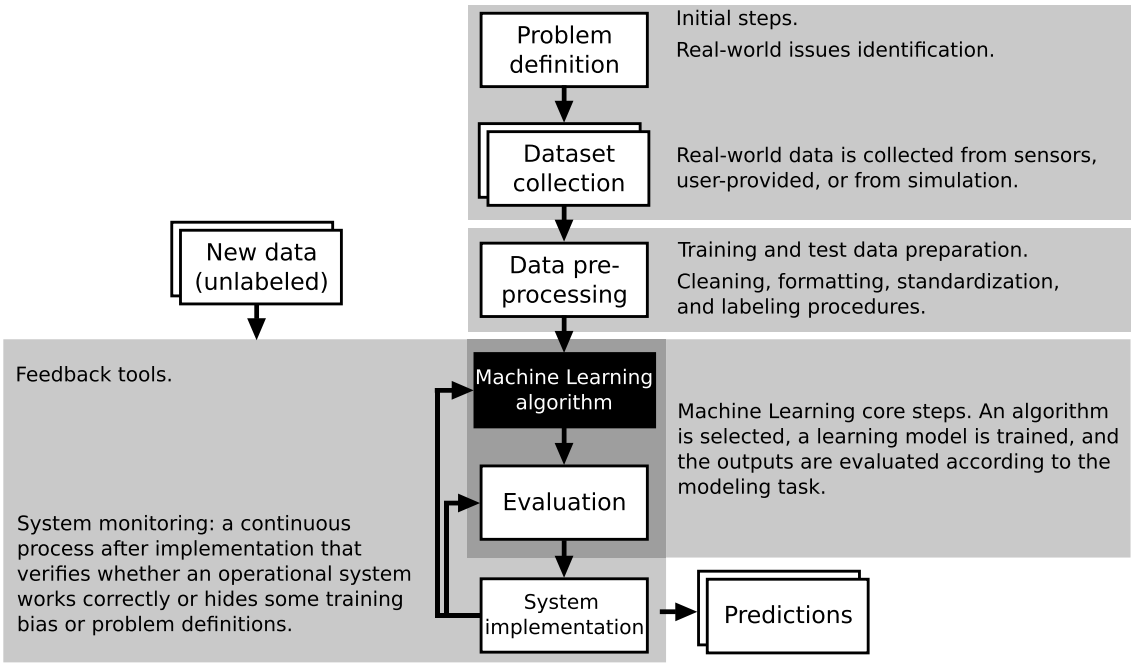


FIGURE 1. Flowchart of multiple tasks tied to the development and implementation of a Machine Learning modeling.

model parameters are adjusted to produce outputs based on a training process that uses input patterns coupled with their desired outputs [11]. The dataset is a pair $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ follows the same definition provided in Equation 1 and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m\}$ is a set defining the respective mappings of each input $\mathbf{x}_k \in \mathbf{X}$, where $\mathbf{y}_k \in \mathbb{R}^c$ for each k.

Supervised learning currently concentrates most of the advances on machine learning [7], hence, on XAI approaches [32] and can be understood as the training of a generalized mapping based on previous data. The supervised learning process aims to determine a mathematical model that minimizes a loss function applied to the difference (or divergence) between all model's predicted values and real values [61]. Therefore, the loss function quantifies the extent to which a prediction is based on the real value for a given data instance, i.e., the more accurate the model predictions, the smaller the loss function results.

Figure 1 illustrates the steps of learning-based modeling from the definitions of the problem to be addressed to the validation and implementation of the model. Supervised learning depends on the historical data available and known (previously processed and labeled) in terms of quantity and quality. Indeed, any Machine Learning application depends on how real-world problems are defined, with data collection and pre-processing steps representing factors that significantly affect both accuracy and efficiency of models in capturing data patterns. However, the generated model does not represent the real world, but rather, only the reality of the data [61].

In general, the data on which the model was trained are not the same used during the evaluation step since the set of rules learned in training might not be the same when the model processes new data. Moreover, the model can memorize the entire training set to increase accuracy, thus leading to an unfair performance rate. Since a good predictive performance is expected for new data, i.e., the ability to generalize, a portion of the data is often excluded and reserved for assessments of the model's performance (the test set).

Regarding the nature of the labels for each input in the dataset, a machine-learning task can be categorized as Regression or Classification. The former trains a function in which the output is a float, i.e., a real-valued vector. In contrast, for classification problems, the model task is to learn a function from the inputs to a finite set, i.e., each input is mapped to one of a set of finite possibilities. Currently, there are several learning algorithms with different specificities and performance abilities, ranging from simpler and more interpretable ones to those more complex and not directly interpretable (black boxes).

Despite our brief introduction to the supervised learning development pipeline, we assume the reader is experienced in Machine Learning. Therefore, no specific model will be deeply reviewed here, and interested readers can consult LeCun et al. [62], Asimov [8], Chen and Guestrin [63], [64], and Ghojogh and Crowley [65] for a detailed overview of some learning models. Due to the remarkable results of current machine learning models and their growing use over the past few years, what elements make such sophisticated models non-interpretable black boxes? The answer is tied to complexity, whose meaning and the reason why it is a critical element in reducing the transparency of machine learning applications demanding explainability are clarified in the next section.

V. WHAT DOES COMPLEXITY MEAN?

Several metrics define and evaluate complexity in computer science. Computational Complexity Theory is a research field of theoretical computing with comprehensive literature on the characterization and classification of computational problems. In the short term, simple and tractable problems can be optimized and solved more efficiently than the existing solutions. However, complex problems require an in-depth investigation of the application domain toward the creation of new tools for solving (or approximating) tasks that require significant resources. Moreover, some problems lack solutions and others are so complex that, although they can be theoretically solved, the solution is unfeasible with current resources, which are intractable.

Algorithmic complexity is a well-known research area related to the definition of the computational time at which a particular algorithm solves a certain problem. However, this study does not focus on time as a measure of complexity. Therefore, for a more detailed discussion of the fundamentals of algorithmic complexity and analysis, we refer to [66] and [67].

The notion of complexity has been used several times so far and will be addressed in several other subjects through this reading, thus raising the need for some clarifications. Specifically for the context between machine learning algorithms and XAI methods, which criteria specify what is considered simple and what is complex? The answer depends on the modeling context under analysis.

A. BIAS AND VARIANCE

Bias and variance are two related topics of extensive debate in Machine Learning; therefore, complexity in learning models cannot be addressed without a discussion on both subjects.

Let us consider a typical supervised learning task for predicting (estimating) variable $\mathbf{y} \in \mathbf{Y}$ from an n-dimensional input $\mathbf{x} \in \mathbf{X}$. There is a function f that captures the true relationships between both variables, i.e., $\mathbf{y} = f(\mathbf{x}) + \epsilon$, with ϵ as a part of $\mathbf{y}$ that cannot be estimated from $\mathbf{x}$. In this context, the learning task objective is to determine an estimator model $\hat{f}$ that approximates the behavior of f, with the estimator describing the relationship between input (explanatory or predictive attributes) and output (dependent or objective variables). A good estimator yields a result as close as possible to the true process that generated the data, which are initially unknown. In an ideal scenario, the estimator model is trained on unlimited data until the predictive patterns can be learned so well that the estimation error tends toward zero.

However, in the real world, we work with training sets of limited size and the generative processes of every data source involve a combination of regular (repeatable) and stochastic components [68]. The objective of a learning model is to train an estimator on the available data in order to acquire the ability to adapt and generalize, generalize, hence, maximize the accuracy for future predictions when the model handles new and unknown data. Technically, ''new data'' are those not used in training [69]. However, accuracy is not maximized by simply learning the characteristics of the training data as precisely as possible [68]. The reason for the loss in accuracy is overfitting, which is a severe problem in Machine Learning. It occurs when the learning function of a model learns (fits) the training data features overly well, which may lead to the generation of less effective (or very incorrect) predictors when applied to unknown data.

A model overfitting the training data tends to capture noise and random aspects of the sampling data (which will not be repeated) as regular elements, therefore, missing broader patterns. On the other hand, when the model is trained in an overly generalistic way (underfitting), the adjustment to new data tends to consider fewer random effects, but at the cost of ignoring regular components [64], [68]. The training fit must be balanced so that the model can learn the true patterns, ignore noise, and minimize the estimation error.

According to Neal et al. [70], a possible way to measure the quality of a predictor model is to quantify its total expected error by the following expression:

$$Err(\mathbf{x}) = \mathbb{E}\left[(\mathbf{Y} - \hat{f}(\mathbf{x}))^2\right] \tag{2}$$

where the difference is squared (for symmetry) for calculating the mean squared error. The total expected error can then be decomposed into the following three components:

$$Err(\mathbf{x}) = \mathcal{E}_{\text{bias}} + \mathcal{E}_{\text{variance}} + \mathcal{E}_{\text{noise}} \tag{3}$$

where $\mathcal{E}_{\text{noise}}$ term represents the intrinsic error independent of the predictor model, $\mathcal{E}_{\text{bias}}$ denotes the bias term estimated with $\mathbb{E}_{\mathbf{x}}[(\mathbb{E}[\hat{f}(\mathbf{x})] - \mathbb{E}[\mathbf{y}|\mathbf{x}])^2]$, and $\mathcal{E}_{\text{variance}}$ is the expected variance of the output predictions estimated with $\mathbb{E}_{\mathbf{x}}\text{Var}(\hat{f}(\mathbf{x}))$. The complete proof of that decomposition is well-known and detailed in Hastie et al. [71], Goodfellow et al. [72], and Ghojogh and Crowley [64]. Note the total approximation error of a predictor is in the function of bias and variance, in addition to an intangible component tied with the noise of the true relationship among the predictive variables (ϵ) that cannot be fundamentally reduced by any model [70], [73]. Therefore, the estimator should simultaneously have low bias and variance (usually hard to achieve) so that the total error is as small as possible.

In other words, a new model $\hat{f}$ is generated at each training process iteration and, owing to data randomness, a variety of predictions is obtained. Bias is the error inherent to the model and reflects the extent to which predictions are far from the objective class. An error due to bias arises from the difference between the expected prediction of the estimator model (or mean) and the correct value of the predicted variable. Variance captures how predictions deviate from each other. The error due to variance can be considered the estimator's sensitivity to small fluctuations as a function of an independent data sample [64], [73].

Geman et al. [74] verified the inconsistency in convergence between bias and variance, claiming the cost of reducing one of them increases the other. Therefore, a predictor must

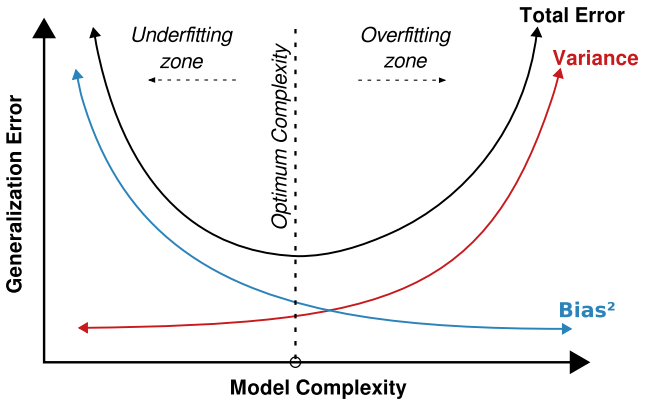


FIGURE 2. Contribution of bias and variance to the generalization error of Machine Learning models as a complexity function.

assume a point on the continuum between bias and variance through learning. According to Briscoe and Feldman [68], the critical parameter that modulates bias and variance of a model is the complexity of its hypothesis. In this context, a standard measure of complexity for estimators is the number of parameters, since it generally establishes the model's degrees of freedom for training data. In other words, more complex hypotheses (models with more parameters) may better fit the training data (high variance), whereas less complex ones (fewer parameters) impose a strong expectation (high bias) on the data, sacrificing fit [64], [68].

A successful learning procedure will produce a model that goes beyond simply memorizing the training data, optimizing the balance between bias and complexity for reducing the generalization error, and providing correct outputs to new input patterns not encountered during training [11]. Figure 2 illustrates the relationship between bias and variance with the generalization error as a function of the complexity of the learning models.

According to Briscoe and Feldman [68], when a model becomes more complex, generalization improves, the error decreases to a minimum, and then starts to increase. The model overfits the training data at high complexities and the predictive performance for new data tends to suffer. It is commonly believed that generalization to unseen test data decreases at high complexities owing to the memorization phenomenon, which occurs when the learning model "memorizes" the training data. Such a phenomenon is a significant matter for theoretical and practical applications of Machine Learning, since it has implications for understanding model generalization and also negative impacts on privacy, for memorization can be explored in attacks to reveal sensitive information from training data [75].

The optimal point of complexity depends on the nature of the patterns to be learned, since the profiles of regular and stochastic processes differ in function of the data source. Therefore, the balance between bias and variance is considered a trade-off between data complexity and fit, i.e., it measures the model's ability to generalize.

Geman et al. [74] claimed the trade-off between bias and variance is universal, which is one of the most significant dilemmas in Machine Learning. However, recent studies have shown it is possible to simultaneously increase the complexity of learning models and reduce bias, with no increase in the total error, owing to advances in regularization and optimization techniques. As an example, Neal et al. [70] provided evidence that bias and variance decrease simultaneously as complexity increases in modern Neural Networks, in contrast to the strict equilibrium intuition of Geman et al. [74]. The results of Neal et al. [70] demonstrated variance decreases in large Neural Networks due to optimizations, whereas sampling variance increases slowly when the network is adequately parameterized.

Zhang et al. [76] indicated modern learning systems tend to fit the training data perfectly while still performing well on the test data. Furthermore, recent studies have investigated the interesting phenomenon of "benign overfitting," which is not restricted to more complex models [77], [78]. Toward deepening the theoretical understanding of the training mechanisms applied to deal with the bias and variance tradeoff and adequately address overfitting, Ghojogh and Crowley [64] published extensive research on regularization and optimization procedures for improving training procedures and reducing the total error of learning models.

Bias and variance are not the only two theoretical complexity measures of machine learning models. However, even those learning models with optimized levels of complexity can be intricate black boxes, providing no single transparent information about their decision-making processes. The balance between bias and variance is a theoretical measure of complexity and influence in overfitting, but the internal structure of a model is another aspect influencing its complexity, as discussed in the following section.

B. NON-LINEARITY

Specifically in the XAI context, complexity is considered a way to translate the number of parameters and the interaction levels among the parameters of a learning model, i.e., its structural configuration. In this sense, complexity describes the transparency and interpretability level of learning models. A simple model is expected to be transparent, hence, more interpretable, because it usually has a reduced number of parameters with few (or no) non-linear relationships with each other. Those simplified parameters can then be inspected directly for evaluations of their effect on each input variable. The model is transparent when such information can be easily obtained by interpreting it, and no other method needs to be applied to generate further explanations.

On the other hand, when the model has a large number of parameters and their relationships are sophisticated (non-linear), obtaining a direct view of their effects for understanding their influence on the decision process is challenging, leading to models with low transparency or opacity. Therefore, a direct interpretation of those models

whose amount and level of non-linearity are too high is virtually impossible [17], [63].

Algorithms derived from the symbolic paradigm are, by definition, transparent (in theory) and require no explanation – even if fully transparent, symbolic algorithms have limited scope and generalizability. Some learning models, such as linear ones, decision trees, rule sets, and Fuzzy Systems, are traditionally considered transparent and interpretable (white-box models) [13], [79], [80]. Linear models are simple, efficient (in specific applications), and interpretable approaches, since their parameters do not have non-linear relationships. However, this is a simplistic and questionable view, for the interpretation of even a linear model may be challenging. Observe the following example:

$$\begin{aligned}\mathbf{y} &= 0.33x_1 + 2.5x_2 + 18.2x_3 - 4.81x_4 + 7.6x_5 + 1.83x_6 \\ &\quad + 43x_7 - 9.1x_8 + 0.15x_9 + 0.01x_{10} - 6.4x_{11} + 3.6x_{12} \\ &\quad + 2.4x_{13} + 2.6x_{14} - 6.3x_{15} - 1.9x_{16} + 4.8x_{17} + 0.25x_{18} \\ &\quad + 6.7x_{19} - 4.2x_{20} + 2.1x_{21} + 5.3x_{22} + 9.01x_{23} - 1.8x_{24} \\ &\quad + 8.5x_{25} + 4.4x_{26} + 7.6x_{27} - 1.1x_{28} - 0.99x_{29} + 7.8x_{30}.\end{aligned} \tag{4}$$

Although such a linear model with 30 parameters can be considered simple from a mathematical perspective, is it easily interpretable? Inspection complexity increases as the number of parameters increases. However, a coefficient is assigned to each feature x_i that linearly describes how each feature affects the model's output, i.e., the effect of each variable can be extracted in a linear model by statistical and graphical support methods such as Friedman H-statistic [81], Partial Dependence plots [82], and Individual Conditional Expectation plots [83], [84]. Even among numerous parameters, only a few may significantly affect the prediction output, making the model inspection task more manageable.

If we can generate predictions and extract information from simplified models (e.g., linear ones), why should we need to employ complex (non-linear) models, such as Neural Networks or Ensembles? The answer does not come from the developers' desires to propose sophisticated models, but from the data – more precisely, from the need to discover hidden patterns in complex data.

The current sources of information are large multidimensional datasets with arbitrary attribute amounts. Many of such attributes can have correlational relationships, following more different non-linear ratios (quadratic, exponential, among other complexities). However, a linear model cannot accurately map patterns hidden under non-linear interactions. When working on complex data with non-linear dependence relationships previously unknown by analysts, simpler models have deficiencies in their generalization abilities to ''unfold'' the intricate correlations among variables.

Non-linear algorithms, such as Neural Networks and Random Forests, can efficiently map non-linear relationships. Data with complex interactions are naturally expected to require more sophisticated solutions to uncover their patterns. Therefore, the models named here ''complex'' are developed to fit such a sophisticated relationship toward solutions in contexts that would hardly be solved by less complex tools. The additional sophistication of modern learning algorithms comes at the expense of their opacity. Whereas a linear transformation can be interpreted by checking the weights associated with the input variables, multiple layers with non-linear interactions inside and among each layer produce complex structures of difficult comprehension, requiring proper tools to obtain explanations for their results [4].

Even linear models considered fully transparent solutions suffer from low accuracy compared with more sophisticated and accurate non-linear approaches. Arrieta et al. [2] observed some exceptional cases in which the data under modeling are ''well structured.'' In those circumstances, simple and accurate models can be trained. Those who develop real machine-learning applications are not expected to continuously operate using controlled and high-quality data; therefore, complex models are more advantageous due to their high approximation flexibility

Simple learning models do not compete with complex ones in terms of predictive performance and generalizability capacity in multiple domains [32]. Once the interpretability challenge placed by complexity has been understood, tools can be designed to explain the opaque outputs of black-box models. The design process starts with the understanding of the goals and requirements of explainability.

VI. XAI GOALS

According to behavioral economist and Nobel laureate in economics Daniel Kahneman, wherever human judgment exists, there will be noise [85]. In other words, humans are susceptible to noise and diverse biases when making choices – e.g., two professional financial analysts can elaborate contrary market forecasts, judges can impose different sentences for a same crime, and doctors can make distinct diagnoses for patients with a same problem. Which elements have influenced those decisions – the weather, the weekday, or the moment they were taken? According to Kahneman et al. [85], those elements are examples of noise that can lead to variability in judgments that should be similar.

However, human actions can be confronted in order to discover the reasoning process that guides a person to make a particular decision and then identify the set of variables that are essential and what is only noise. Let us imagine a decision was based on a complex machine-learning model. In many scenarios, noise can have harmful effects that should not be ignored. However, how are the processes or logical reasons of a ''black-box'' model interpreted? How can a model's main influence be explained?

Despite the significant advances in both definition and construction of learning algorithms, explainability remains a relatively new research topic. The Explainable Artificial Intelligence community has been active, promoting many

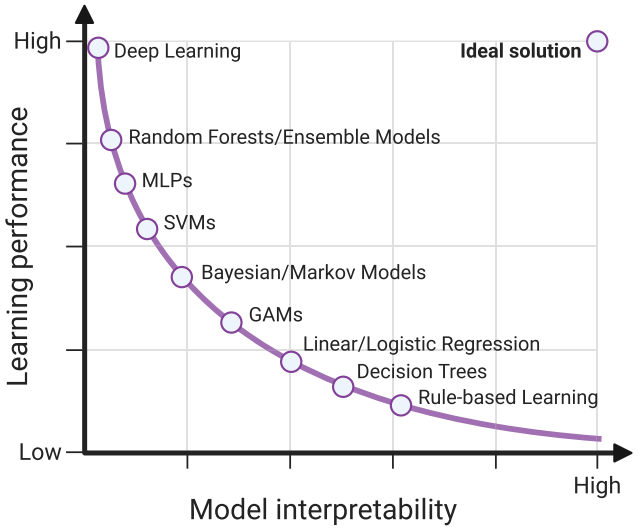


FIGURE 3. Trade-off between interpretability and predictive performance of main learning models. High-performance models are complex and offer low interpretability of their decisions, whereas more transparent ones show low predictive performances.

scientific events dedicated to the subject [4], with high-quality research and papers published in renowned scientific vehicles. This research interest demonstrates the importance of explaining intelligent-based system characteristics not only to satisfy transparency requirements, but also to promote interaction between humans and Artificial Intelligence, thus helping the development, maintenance, and monitoring of learning-based approaches [14].

Došilović et al. [7], Gunning and Aha [23], and Arrieta et al. [2] reported the existence of a tradeoff between interpretability and performance in learning models. Figure 3 illustrates the conflict of goals between predictive performance and transparency of the more commonly used algorithms, from rule-based models and decision trees to Deep Learning-based models. Although it shows a simple comparison, it provides a good overview of the contrast between interpretability and accuracy. In terms of non-linear structures, the more complex the model, the lower its interpretability – the model is more opaque, hence, more difficult to be directly interpreted. High-performance models are often less interpretable and most explainable ones have low accuracy [23].

Note Figure 3 cannot be overlapped with the previous chart in Figure 2, although both have a similar complexity/interpretability axis. More specifically, when we move into the ''Model interpretability'' axis of Figure 3 toward high to low levels, we are approaching models with high learning performances and, concomitantly, moving toward more complex ones, which does not imply high-complex models have high generalization errors (see Figure 2). Although Deep Learning models are known as far more complex than rule-based models, it does not necessarily mean one is more complexity-optimized than the other. As previously discussed, different machine learning models have different performance abilities in function of the application context. Each model has its own underfitting and overfitting chart, with complexity balancing according to the learning algorithm and the needs of the application domain.

The lack of interpretability of more complex models restricts the decision-making processes to the choice of whether or not to execute an automatic decision, with no additional elements supporting the understanding and justifying the decision made [16]. XAI explanations address the opacity of complex machine learning models, helping users better understand the impacts of learning models [4], [86]. XAI does not impose limitations, invalidate, or render Machine Learning unfeasible. Users affected by learning-based applications have the right to appropriately know and understand the essential factors that lead to those decisions [24], [30]; however, those decisions must be explained while maintaining high levels of prediction performance [23].

In summary, XAI research addresses the following considerations regarding design and development of new explainability approaches:

- **Problem**: Machine learning algorithms do not analyze data in the same way as humans do. Learning models use complex mathematical mechanisms to find patterns that a human analyst may not entirely know or understand [87]. Modern learning systems have high discriminating power, at the cost of increased complexity and consequent low interpretability. Such high precision does not guarantee the decisions produced are, in fact, fair and not permeated by some hidden spurious bias. The lack of explanatory power increases trustworthiness problems and transforms learning algorithms into unreliable decision-support systems, making the implementation of learning-based systems in critical real-world domains challenging [88]. Therefore, the right to explanation arises, i.e., learning models applied to decisions that can significantly affect their users' lives must be more transparent and interpretable and provide reasonable explanations of the logical processes behind the results/predictions. Even when explanations are provided, they can be inaccurate, insignificant, or useless. Many XAI methods that have proven valuable tools have drawbacks, including instability, which reduces trust and confidence in their application. Stability is achieved (*i*) when the XAI method generates consistent explanations for multiple runs of a same instance and (*ii*) when it explains similar instances. According to Amparore et al. [25], an explainability method must be at least stable to be reliable. Therefore, the explainability for machine learning must satisfy consistency, accuracy, and trustworthiness requirements to be useful. If the explanations are inconsistent or inaccurate, they are not trustworthy and are useless.
- **Hypothesis**: The empirical success of Machine Learning is due to its computationally efficient algorithms and high-parametric space, with hundreds (millions or even billions) of parameters [2]. If the reasoning involved in

the decision processes of an intelligent system could be "explained by humans," knowledge about the working methods of those algorithms could be extracted, making them more transparent, verifiable, and applicable.

- **XAI goals and basic requirements**: Decisions derived from applications supported by machine learning must be explained for breaking the limits of models that exclusively find patterns in data. XAI aims to elucidate machine learning, helping us discover where the patterns learned came from, why they occurred, and what they imply, leading to our understanding of why decisions were made, thus promoting fairness by detecting hidden biases. An XAI approach should be designed to verify the different aspects of complex learning functions and decompose opaque elements for generating human-interpretable information. They also must consider the target public, since users and developers have different needs and demand different information. XAI explanations make black-box problems more transparent, demystifying the logic behind the outputs or models' internal mechanisms from local or global views. Reliable explanations should accurately and consistently reflect the behavior of the underlying model – both fundamental requirements for promoting explainability, since it would be difficult to trust an explainer that generates inconsistent explanations. Therefore, XAI methods must provide reliable explanation elements for supporting machine-learning decisions and converting black boxes into verifiable tools.

VII. XAI GROUND THEORY

One of the first challenges observed in the XAI literature is the lack of terminology standards – many different terms are often used interchangeably as synonyms with no clear definition. Amann et al. [16] discussed the need to harmonize the XAI vocabulary and argued a direct consequence of that lack of definition is every new publication on XAI must detail the meaning of terms that will support the research, thus causing more confusion and inflation of definitions in use. Note XAI is a young research line influenced by other domains of knowledge, including the humanities, and some of the more frequent terms in its scope are broad and slightly different. Before moving forward and aiming to avoid future mistakes, some recurring terms must be clarified. Following a literature review, we propose concise and clear definitions here.

Although sometimes used interchangeably, interpretability and explainability are two concepts that maintain some differences. *Interpretability* is a relatively elusive general concept that can be characterized differently [13]. The word "interpret" denotes what can be explained in an understandable way [21]. According to Miller [22], interpretability is the extent to which a human can understand the reason for a decision. In machine learning solutions, an interpretable application allows users to observe the outputs, study the model architecture, and then understand how the input data were mathematically and logically mapped into the outputs [4]. In this sense,

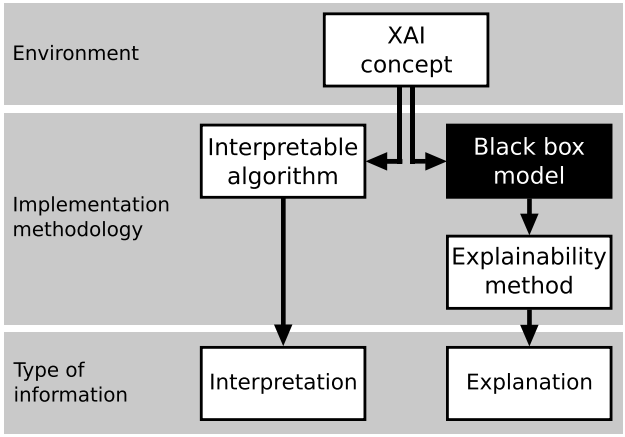


FIGURE 4. Conceptual difference between interpretability and explainability within the XAI context.

interpretability is seen as a passive element, indicating the extent to which meaning can be extracted from a domain with abstract information [2], [7].

In psychological terms, the explanation is "*the currency in which beliefs are exchanged*" [21], [89], i.e., the communication of what has been understood by providing reasoning for something. Although *explainability* is also a broad concept, it can be used in XAI by referring to the additional information generated for verifying how a learning model yields a particular result [16]. According to Bhatt et al. [90], it can be any technique that enables users or developers of machine learning to understand why models behave in the way they do. Explainability is established as an active element, indicating the collection of actions or procedures that clarify or detail a model decision [2], [7]. What makes an explanation better than another depends on the context addressed and the questions to be answered in the explanation task.

Figure 4 illustrates the slight, but significant difference between interpretability and explainability in the XAI environment. Learning algorithms share the same objective of providing accurate predictions, but interpretable and explainable implementations differ regarding the technologies applied.

- **Interpretable algorithms** are those through which a human can inherently and intuitively understand the working logic and extract valuable information.
- **Explainability algorithms** are those applied to open black boxes *a posteriori*, generating useful information about the behavior of opaque models [4], [16].

Neither interpretability, nor explainability is specific or restricted enough to make a definitive formalization. However, when a learning model is not directly interpretable, we can use tools designed to extract information elements and generate explanations, thus providing interpretability.

Other terms frequently used in XAI for Machine Learning are comprehensibility, model transparency, and trust. The former is described in the literature as an interpretability synonym because comprehensibility is the ability of a model

itself to express information [2], [91]. Similarly, a transparent model can be interpreted, i.e., a model with some degree of interpretability [2], [13]. In general, final users are not equipped to understand how data and code interact to make decisions affecting them individually. In this sense, the transparency concept includes various efforts to provide practitioners, especially end-users, with relevant information on how a decision model works [90]. Finally, trust is also a term with a subjective meaning commonly associated with a psychological state of security that, in the Machine Learning context, has often been expressed through the models' good predictive performance (evaluated by performance metrics). However, this study shows this is a simplistic perspective, since more trust criteria must be considered.

When used in isolation, traditional performance metrics can lead to misleading evaluations. We will not open a discussion analyzing the multiple performance metrics for machine learning applications available in the literature, since it is out of this study's scope, although an in-depth understanding of metrics is recommended for any machine learning practitioner. The reader can find detailed descriptions of the metrics used for evaluating learning algorithms in Gareth et al. [92] and the drawbacks of some performance metrics in Batista et al. [93].

An accurate measurement of a model's prediction error is essential for assessing its quality. The primary goal of machine learning modeling is to build models that make accurate predictions of the target value of new data (data not used in training). A performance metric should reflect the modeling objectives; however, instead of reporting the model error on new data, traditional metrics are often applied to a test dataset, which is a portion of the data split from the training dataset – although assessment and recurrent mechanisms consider new or residual data to check the quality of predictions and adjust the model ''on the fly,'' if necessary. Any model is naturally optimized to describe the data on which it was trained. In this sense, the information generated by the methodology typically used for error measurement in learning models can be misleading, resulting in the selection of inaccurate and inferior models [69].

According to Amann et al. [16], transparency is one of the main requirements for the establishment of trust in intelligent systems. Therefore, regarding applications based on complex black-box models, efforts should be made to include transparency, with explainability proposing tools for achieving transparency. Other frequent terms encountered in the XAI literature are derivations in context or semantic meaning of those addressed and clarified in this section, which are the core terms in XAI vocabulary. The reader can find helpful definitions for other such terms in Arrieta et al. [4], Adadi and Berrada [2], and Bhatt et al. [90]. In the following subsections, we describe additional important concepts of XAI theory.

A. XAI NEEDS AND CHALLENGES

The Machine Learning literature has been ''algorithm-centric,'' assuming the approaches and models developed are intrinsically interpretable [94], [95], but with no further verification of the interpretability of the algorithms [42]. Munroe [96] addressed this subject using sarcasm. Machine-learning applications are constructed by mathematical modeling tools derived from linear algebra and calculus. From the user's viewpoint, such complex models are black boxes in which input data enter on one side and answers are collected on the other. However, what is the solution to ''incorrect'' results, i.e., those results that do not meet performance criteria or statistical metrics? The mathematical tools may be adjusted by hyperparameter optimizations until the answers begin to appear correct.

Real research on Machine Learning goes beyond simple adjustments in models. Despite such a satirical view of the learning modeling algorithm-centric process, a matter of significant importance for the establishment of trust in machine learning applications that have been in the background must be considered, i.e., results that ''look correct.''

Assuming a learning algorithm is intrinsically interpretable is not always an incorrect or problematic view. In some cases, interpretability may not be necessary (e.g., when the algorithmic decisions are not leading to significant consequences that affect user safety, when there is no possibility of generating injustices, or when the task solution is generally well-known and was already sufficiently tested) [21], [42]. However, the range of decisions made by intelligent systems based on machine learning increases daily and is no longer restricted to academic and research environments. Handling incorrect results requires understanding the source of the errors and not only their relation to model adjustments. Errors can sometimes be related to biases in the training data learned by the model. In this case, adjusting the model will improve the metrics' results, but may hide biases. Understanding the source of errors in black-box applications goes beyond assuming interpretability is unnecessary because a model has high accuracy rates. Trained models can hide biases, requiring the exploration of their reasoning.

Deep Learning and ensemble learning models have intricate and complex internal mechanisms that are virtually impossible to interpret. Moreover, the reasons leading to a decision cannot be understood, thus obscuring verification tasks that try to assess the logic behind predictions [97]. Opaque models are black boxes in a setup where input data enter one side and predictions are output on the other side, with the processing details remaining obscure or unknown. Black box components do not clarify their reasoning, hampering the understanding of the way they achieve a given result [4]. The top part in Figure 5 illustrates a typical supervised learning application. Each learning model has its own capacity and each data context may demand different capabilities; therefore, different models can have distinct accuracies in a same dataset. In this context, performance metrics guide data scientists in selecting the most accurate model for each application.

After training, testing, and achieving predetermined accuracy requirements, the learning model can be deployed to

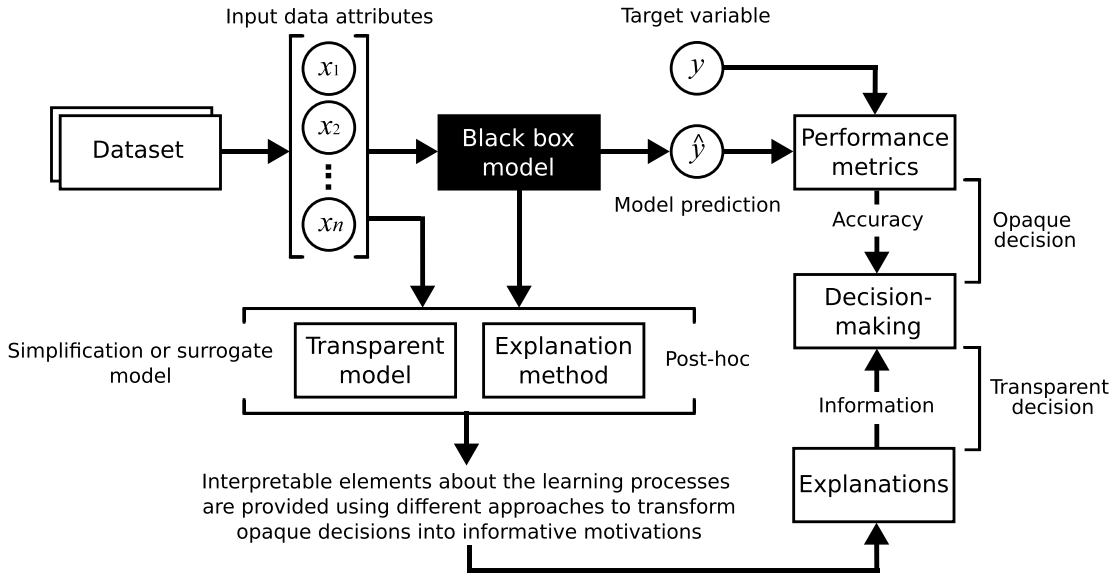


FIGURE 5. Explainability is positioned as a complement to Machine Learning. Complex models work as black boxes and XAI tools explain black-box decisions in interpretable terms, enabling practitioners to make decisions based on more transparent information.

classify unlabeled data. At this moment, it will be on unknown ground and apply all the knowledge acquired (learned patterns) during the training-testing procedure to the new input data. Model monitoring is challenging due to the significant diversity in the relationship modes between data and mapped spaces [14]. It is essential to verify whether an operative model classifies new data correctly or if its performance is considered satisfactory owing to some training bias or problem definition. However, in practice, when running on unlabeled data, the model can generate incorrect classifications or classifications based on incorrect reasons and, consequently, problems in machine-learning applications, such as spurious noise or bias, can remain hidden in the decision-making process.

The internal elements of machine learning models are commonly illustrated graphically, although the models are essentially mathematical functions. Such a representation is often used because the reading of a complete function of a model (e.g., Neural Network) can be problematic, and visual representation as a network of neurons enhances readability [65], [98], [99]. Despite graphical representations supporting explanations of models' architectural elements, understanding how learning models work is difficult and discovering how complex non-linear functions transform data is an overwhelming task.

However, the assurance of trust in intelligent-based applications requires more than the supply of results that appear correct – they must be both correct and fair. That is when XAI can be applied, generating explanations that can check whether model predictions are correct for correct reasons, thus providing compliance guarantees justifying black boxes' decision-making processes. The bottom of Figure 5 illustrates such a scenario.

Although novel learning architectures are constantly developed toward better performances in most different domains [7], [8], [100], their understanding has been primarily ignored [101]. Some studies have demonstrated the weaknesses of high-end models, proving not everything is perfect, even when Machine Learning can reach high accuracy rates. As examples, ethnic biases were detected in a model that predicted criminal recidivism, software used by Amazon excluded ethnic minorities while determining areas in the United States that would receive discount offers [26], and a model trained to predict the probability of death from pneumonia assigned lower risk to patients with asthma [102].

Szegedy et al. [103] demonstrated how a powerful deep Convolutional Neural Network (CNN) could be manipulated to misclassify pictures of school buses as being ostriches by simply introducing a visually imperceptible noise to the test images. Goodfellow et al. [104] discussed the susceptibility of Artificial Neural Networks to adversarial attacks, a type of attack used to discover minimal changes to be made to input data for "foolling" the network and causing wrong classifications. Su et al. [105] analyzed the extreme case of a one-pixel attack and Haim et al. [75] showed sensitive training data encoded in the parameters of a trained classifier could be recovered in a training-data reconstruction attack.

A critical problem that may remain unnoticed by commonly used evaluation metrics is the generalization ability of learning methods. Lapuschkin et al. [106] developed interesting research demonstrating cases whose model's predictions were based on spurious correlations unrelated to the learning objectives, known as Clever Hans phenomenon. If complex models provide difficult-to-interpret decisions in sensitive contexts that can affect people's lives (e.g., credit scores, public administration, and medicine), the interactions between attributes that provide predictive accuracy must be understood [17]. Once the reasons behind those decisions have been comprehended, it is possible to verify whether

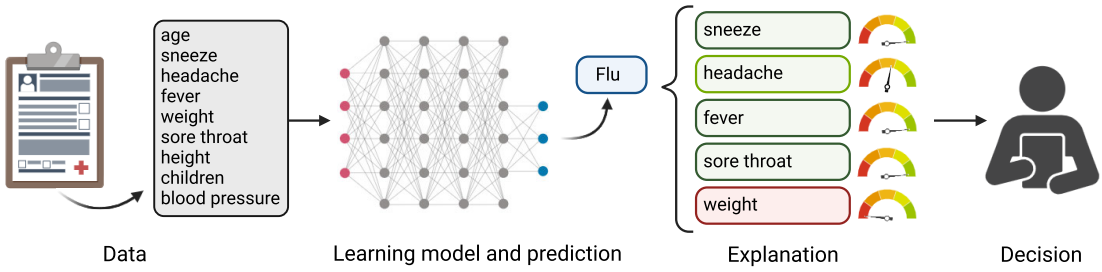


FIGURE 6. Explanation of the influence features in a prediction diagnosis of flu. Symptoms that contribute to the result are highlighted in green and those that do not contribute are highlighted in red.

the model results are reliable or based on spurious biases or noise.

Spurious biases can be incorporated into the knowledge space of learning models in more diverse ways, leading to unfair decisions. Systematic biases in training datasets (socially constructed biases, inaccuracies, and errors in data collection [2]), problems or errors with modeling definitions and algorithms, limitations in training, or lack of evaluation are typical sources of biases. Once incorporated, hidden biases are difficult to detect and treat in complex models. However, explainability can be applied to support bias detection by providing insights into the models' decision processes. Zhang and Zhu [9] and Ras et al. [107] investigated the adoption of XAI approaches for bias detection.

XAI can support information-based decisions. Let us consider a medical diagnosis support system based on a classification model that predicts a patient's flu according to his/her history of symptoms or is verified by the hospital team. The application explaining which symptoms have influenced the decision apart from prediction would be helpful and very informative to doctors so that they could have a better basis for diagnosis instead of simply making a decision based on automatic results [97]. Figure 6 illustrates the situation and the importance of adequate explanations.

In addition to the need for decision-making processes supported by information, regulations such as the European Union's implementation of GDPR (discussed in Section I-A) also require intelligent-based applications observe ethical matters. However, the regulation is not sufficiently detailed and does not define the tools or requirements to be designed or made available for ensuring compliance with the law.

In such a context, explainability can verify whether the model is sufficiently fair in its decisions, mainly when training data include biased or incomplete cuttings [108]. Correlation does not imply causality; however, causality involves correlation. In this sense, explainability can help validate predicted outcomes by revealing possible correlation relationships related to specific outcomes [2]. Explaining learning-based applications can promote the discovery of potential failures, helping data scientists identify causes of errors more efficiently and indicate what the model has really learned from the data [90].

"Opening" the intricate black boxes that modern machine learning systems have become is not a trivial task. State-of-the-art models are black boxes difficult to understand [42]. Even linear models, which are simpler and more transparent, are fully transparent only in limited contexts, i.e., explanations of high-dimensional linear models are challenging [26]. In some cases, and contrarily to expectations, models considered transparent and of effortless understanding may decrease users' chances of error detection owing to high amounts of information [108].

Omitting explainability can lead to challenges for model trust. However, toward properly including explainability, the generated explanations must be helpful, understandable, valid, accurate, and consistent to be useful [25], [109]. Only because an explanation "makes sense" does not imply it is automatically valid [110]. One of the most significant challenges within the XAI environment is to provide reliable explanations supported by robust validations [16].

B. XAI CATEGORIZATION OF APPROACHES

The first initiative for the generation of explanations for artificial intelligence models dates back to the 1980s, when researchers introduced questions about the negligence of symbolic systems regarding explanations [111]. However, the concept of XAI has been consolidated recently owing to the growing need to explain the results of complex learning models [2].

The XAI research community has been devoted to the creation of multiple new methodologies and approaches that explain specific models or even those independent of any model. XAI techniques can be classified according to their functional characteristics and objectives and black-box problems can be "open" in two ways, namely, (*i*) by constructing transparent systems accurately enough to replace black-box ones or working with them in a support or redundancy mode [20], [112] or (*ii*) through post-hoc explainability. Figure 7 shows a diagram of the XAI approach.

It is an intuitive approach to replace a black-box model with a transparent one; however, many Machine Learning tasks are so complex that they cannot be solved by constructing an accurate, transparent model. In those cases, there is the need for explainability through post-hoc methods designed to

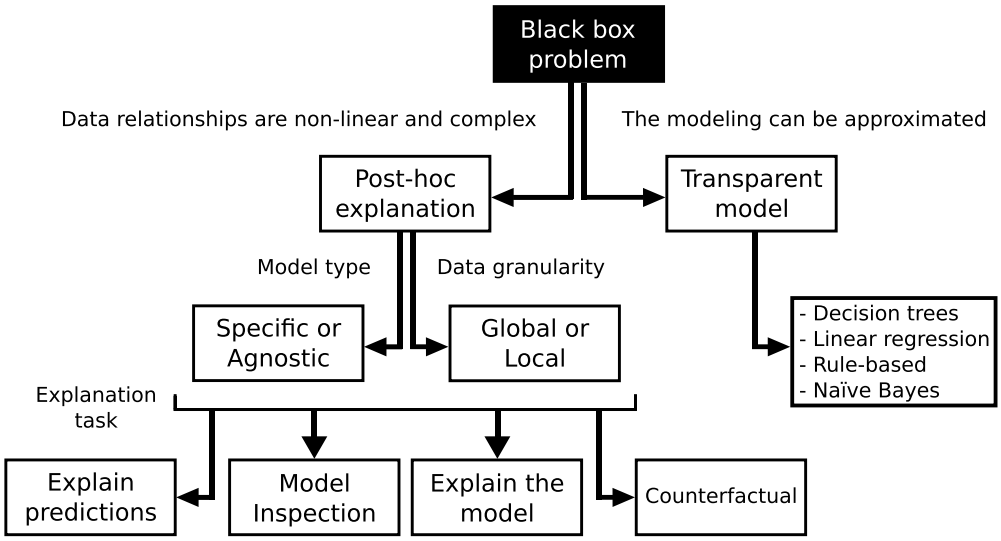


FIGURE 7. Taxonomy of explainability approaches regarding the different understanding objectives over a black-box problem. Explanations can be derived by approximating black boxes through transparent models or post-hoc methods.

explain models that are not directly interpretable and cannot be efficiently approximated by a transparent model [2]. Post-hoc approaches are applied after a black-box model has been trained and used to explain its predictions based on tools that improve the understanding of black-box applications. In addition, post-hoc explanations are typically not intended to unravel the way a learning model works internally, but rather, to provide helpful information to users and data scientists, such as importance of specific modeling parameters or data attributes [26].

Based on Figure 7, the following subsections categorize the XAI approaches according to the underlying learning model, data granularity, and explanation tasks.

1) ACCORDING TO THE MODEL

Post-hoc techniques can be classified according to the type of model they are designed for. XAI methods created for a specific class of models are indicated when the explainability objective is to unravel the logic of specific classifiers or when an advantage can be taken from the model's architecture. Model-independent methods, or model-agnostic, can (theoretically) be applied to any learning model, regardless of the underlying architecture or algorithm, because they are not designed to consider any characteristics of a specific model.

Model-specific methods may perform better, for they explore the functional specificities of the model's class under explanation. However, they have limited application capabilities – since they are planned to work with a singular class of learning models, they may not be flexible enough to work with any other type of model out of the class they were designed for [4]. In contrast, model-agnostic methods aim to understand the reasoning behind predictions using simplifications, relevance estimates, or visualizations [2]. They separate predictions from explanations and search for explanatory elements without entering the classifier's internal logic [4].

Ribeiro et al. [97] argued model-agnostic techniques are more valuable because they provide explanations for any algorithm, enabling different models to be compared by a same explainability technique. Conversely, Chen et al. [113] claimed such techniques rely excessively on modeling *a posteriori* of arbitrary learning functions and, consequently, may suffer from sampling variability when applied to models with many input attributes, which can hamper convergence among results. In this sense, whether model-specific or model-agnostic, a more suitable choice will depend on the explainability task. A model-specific method can be employed if learning models from different classes do not require comparisons and a superior model-specific method is available for the model under explanation. Otherwise, a model-agnostic method will be more flexible.

However, series of models is a type of architecture that has imposed significant challenges on transparency. When the outputs of a predictive model are used as inputs for another predictive model, we have a series of models architecture [114], which are complex pipelines composed of different types of black boxes, such as linear, tree, and deep models [115], [116]. Such a design hampers explainability in comparison to a single model. As an example, different proprietary models are distributed across different institutions in consumer scoring tasks. Each pipeline branch has thousands of data segments about consumers for simulating distinct elements related to consumer scores (fraud scores, credit scores, and health risk scores, among others).

Multiple high-stake applications use series of models, which demand approaches designed to explain such an architecture as crucial in XAI research due to the lack of transparency

of this complex structure and the need for debugging and building trust in applications based on series of models [114]. A natural solution might be to apply model-agnostic methods for explaining the entire pipeline of a series of models at once. Although standard model-agnostic techniques can explain a series of models, they do not work appropriately because model-agnostic methods suffer from some shortcomings in this context, i.e., they require access to every model in the series (but institutions sometimes cannot share their proprietary models) and have a high computational cost, which may not be tractable for large pipelines [114].

Standard model-specific XAI methods are often faster than model-agnostic alternatives. However, they cannot explain series of models, since model-specific methods are designed to operate by considering one type of black box, and a series of models may comprise many types of predictive models in its pipeline. Series of models have been little explored in XAI literature. The explainability solutions for such models demand hybrid model-specific XAI tools that can handle distributed pipelines but, simultaneously, sufficiently generalist to manage the diversity of the model's compositions [114].

2) ACCORDING TO GRANULARITY

Another categorization in XAI is related to the granularity of explanations. According to Ribeiro et al. [97], there are two main classes of methods with respect to granularity, namely,

- **Global**: Methods designed for global explainability are applied when the task is to obtain an overview of the model's behavior regarding the more influential elements [117]. Global methods provide a summarized description of a model's behavior when the scope is the entire dataset (or a significant cut) [118]. The strategy is commonly used to compare the global relevance of a variable and understand which other variables are more relevant in population-level decisions (population-wise). As an example, estimating global behavior in settings such as climate change or drug consumption trends is more valuable than providing explanations for all possible modeling elements [4], [119].
- **Local**: Methods devoted to local explainability are indicated when the task is to retain a more accurate description of the details connected to a single-instance prediction (instance-wise) [120]. At a high level of knowledge, the goal is to understand the motivation for a specific prediction. Local explanations are valuable for complex models with different behaviors when exposed to different combinations of input variables [117].

The user must trust the model will exhibit an appropriate behavior when deployed. In the modeling stage, evaluation metrics were applied to the model generated from the training data (validation process) for emulating real-world behaviors. However, data content and the real world significantly differ. Global explainability can investigate whether a model reflects its modeling expected behavior.

The capture of an overview of all learned mapping can be difficult or less informative, especially in models with high numbers of attributes, since it demands the explanation method to discover an optimum in detecting any functional dependence between all input data and targets, which can be an NP-hard problem in general [121]. Users must trust the prediction to make a decision based on it. Apart from evaluation metrics, predictions must be tested individually by local explainability for justifying them, particularly when the consequences of an action can be catastrophic (e.g., incorrect medical diagnosis or counter-terrorism).

Individual explanations can justify which input features of a data instance lead to a specific decision when a global view of the model is not sufficiently descriptive [4]. However, for graph-based learning models, sometimes the goal of local explanations is not limited to explaining input features of a specific node; rather, it can be more valuable to explain which nodes in a neighborhood were most important for a decision.

According to Ribeiro et al. [97], to be meaningful, an explanation must maintain local accuracy, i.e., the correspondence between the explanation and the behavior of the model in the neighborhood of the predicted instance. On the other hand, the authors also claimed local accuracy does not imply global fidelity simultaneously, since globally important characteristics may not be locally important, and vice versa. An utterly faithful global explanation cannot be obtained without a complete description of the entire model. As reported by Wojtas and Chen [121], a simple collection of instance explanations may not work at the population level characterization because local explanations are specific to the instance level and often inconsistent with global explanations. Therefore, identifying globally faithful and interpretable explanations remains a challenge [97].

3) ACCORDING TO THE EXPLANATION TASK

Some XAI methods are designed to understand learning models' structures and internal mechanisms, i.e., model explanation methods. Such a category of methods is typically found in Neural Network applications, in which information visualization is applied to generate visual representations of the internal patterns of neural units. However, contrarily to intuition, Poursabzi-Sangdeh et al. [108] indicated exposing the internal mechanisms of a learning model reduces the users' ability to detect faulty behaviors for unusual instances. Amann et al. [16] claimed such an interpretability reduction might be related to the overhead induced by the large amount of information users are exposed to during the understanding process, even in transparent models. Note the findings of Poursabzi-Sangdeh et al. [108] do not invalidate model explanation methods, but rather alert developers to design tools that synthesize large amounts of information carefully.

Model inspection is used when the explanation task is to verify the model's sensitivity, i.e., the behavior of the learning algorithm or its predictions when the input data are varied through perturbations. On the other hand, prediction explanation methods display visual or textual elements

that provide a qualitative/quantitative understanding of the relationship between the input variables and a prediction for clarifying the factors that influence the model's final decision.

According to Ribeiro et al. [97], prediction explanation methods promote trust between users and learning applications faithfully and intelligibly. The explanation of predictions does not require all the classifier's internal logic to be unraveled. Moreover, such methods should explain individual predictions of a complex model, regardless of whether it is correct or not. Prediction explanation is one of the leading research areas in XAI, with multiple techniques devoted to identifying and quantifying the contribution of input elements to predictions of complex models [4], [33].

Explanations can be provided by a global method or a local attribution one that assigns some measure of importance to each input datum in both granularity, i.e., for a collection of instances or a set of input attributes of a specific data instance. Finally, the output of a learning model can be interpreted through evidence-based (or factual) explanations. In this context, contrastive and counterfactual methods seek justifications for why a decision was not different from that one predicted and how it can be modified, respectively [122].

C. XAI AS A FAIR PLAY ELEMENT TO ARTIFICIAL INTELLIGENCE

Establishing trust is one of the main foundations of XAI. Therefore, it is unreasonable to use a black-box explainer as a black box itself. According to GDPR, regarding the right to explanation, "*if a decision-support system provides inconsistent explanations for similar instances (or the same instance), those explanations provided to the user cannot be trusted*." A suitable explainability method cannot provide (completely) different explanations when executed multiple times to explain a same instance or similar ones. In other words, an explainer must be consistent to be reliable [25]. Stability is a critical requirement XAI methods must verify to ensure consistency of explanations, which must be constant [123]. The lack of consistency may lead to problems with the explanation's general trustworthiness, thus questioning the entire purpose of explainability.

It is necessary to highlight that explanations are context-dependent, i.e., it is impossible to define the questions an XAI method will answer without considering the needs of the target audience to whom explanations are addressed [90], [97]. Domain experts and developers may be interested in auditing the behavior of models for discovering errors or vulnerabilities hidden within complexity. Improving learning models or prediction understanding would lead to the correction of flaws in the application of the modeling [2]. End users and regulatory agencies may demand logical and verifiable outputs from decision-making systems that can affect them, clarify doubts, and ensure the observation of compliance criteria, which are important for indicating a learning application is trustworthy. Data scientists and corporate managers demand tools to verify whether their data are being transformed into useful information for the right reasons. Liao et al. [124] defined an "XAI question bank" as a set of how-, why-, and what-based questions users might ask about Machine Learning for guiding XAI developers' good design practices. Bhatt et al. [90] identified explainability needs according to the audience and developed a framework with a set of goals for explainability to facilitate end-user interaction.

Other significant challenges in XAI beyond the existing technical challenges in explaining complex models should be highlighted. Important principles that must always guide any Artificial Intelligence system development and implementation (e.g., security, privacy, and data protection guarantees) must also be included in explainability approaches. Regarding GDPR again (see Section I-A), algorithmic decisions "*which produce legal effects concerning (a citizen) or of similar importance shall not be based on the data revealing sensitive information, for example about ethnic origins, political opinions, sexual orientation*" [25]. In other words, explaining a prediction does not mean disclosing and exposing sensitive data that should not be published [125]. However, defining what is sensitive data must comply with social principles like ethics and fairness. Fairness refers to the ability of learning models to make fair decisions with no influence of hidden biases that might mistakenly affect (negatively or positively) them [126]. Intelligent computer applications must be impartial concerning social aspects such as religion, socioeconomic background, political opinions, or ethnic origins [127].

In addition to respecting those elements, XAI methods should work toward improving learning-based applications and ensuring they accomplish their tasks accurately and responsibly. Interested readers can find extensive discussions on the needs and challenges in promoting responsible Artificial Intelligence in Arrieta et al. [2] and Tjoa and Guan [42].

VIII. XAI APPROACHES AND METHODS

In this section, we provide an overview of the most recent and relevant XAI methods by analyzing their characteristics and functionalities for generating explanations for black-box problems. According to Molnar [40], the decision process of a learning-based model can be interpreted analyzing the influence of each variable (attribute) on instance prediction. Such individual influences (or importance) can be easily verified in a linear model f, formalized as follows for an n-dimensional data instance, $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbf{X}$:

$$f(\mathbf{x}) = \omega_0 + \omega_1 x_1 + \cdots + \omega_n x_n \tag{5}$$

where x_i is each attribute value i of instance $\mathbf{x}$, with ω_i being its associated weight and $\omega_i x_i$ describing the effect of each variable (weight multiplied by attribute value). The influence (ϕ_i) i-th variable implies on prediction $f(\mathbf{x})$ is calculated as

$$\phi_i(f) = \omega_i x_i - \mathbb{E}[\omega_i \mathbf{X}_i] = \omega_i x_i - \omega_i \mathbb{E}[\mathbf{X}_i] \tag{6}$$

where $\mathbb{E}[\mathbf{X}_i]$ is the expected value of variable i. The contribution of each attribute to a prediction can be inferred according to the difference between its effect and the expected

value. Adding all variable influences of the instance leads to

$$\begin{aligned}\sum_{i=1}^{n}\phi_i(f) &= \sum_{i=1}^{n}(\omega_i x_i - \mathbb{E}[\omega_i \mathbf{X}_i]) \\ &= (\omega_0 + \sum_{i=1}^{n}\omega_i x_i) - (\omega_0 + \sum_{i=1}^{n}\mathbb{E}[\omega_i \mathbf{X}_i]) \\ &= f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})] \end{aligned} \tag{7}$$

as the difference between instance $\mathbf{x}$ prediction and prediction expected value.

Explaining other classes of machine-learning models in such a simplified way would be interesting; however, the concept of ''effect'' worked directly only in that case because of the model's linearity. Even a moderate amount of non-linearity in relationships among attributes may increase the complexity of the linear modeling formulation, thus reducing its interpretability. In other words, even linear models can be sufficiently complex to be interpretable [113]. Feature effect/impact, variable contribution, and feature-level interpretation are terms often used in XAI literature to describe how or to what extent each input feature contributes to the model's prediction, i.e., feature importance [33], [84].

Breiman [17] proposed one of the first solutions for identifying important features. His approach permutes each feature (randomly shuffling the feature values) to assess its individual contributions. More specifically, permuting feature values breaks the connection between the feature and the target variable, resulting in a significant loss of prediction performance if the feature is important. Therefore, the amount of performance deterioration indicates the extent to which the model depends on that feature [84]. On the other hand, Breiman's method is model-specific for trained Random Forests.

Despite the significant assortment of XAI methodologies, explaining predictions through feature-level interpretations is a common goal of XAI approaches [33] and several authors (including the ones of the present research) classify XAI methods according to other elements or mechanisms applied to accomplish the task of feature importance. Feature attribution is at the core of feature importance. However, as demonstrated here, that is not the only way to verify importance. The main terminology related to feature importance problems is defined as follows:

- **Feature Attribution**: Measures the contributions of individual input features to the performance of a supervised learning model, fairly distributing the predicted values among the input variables for quantifying each variable's relevance [84], [121].
- **Additive Importance**: An explanation modeling according to which the summation of all feature importances should approximate the original predicted value [109].
- **Sensitivity**: Measures how the predictive performance of a learning model varies (increases or decreases) by perturbing each input feature [128]. From the perspective of sensitivity analysis, the more important the variables, the more significantly their contribution to the predictive performance.
- **Gradient-based**: A particular case of the sensitivity approach that assesses the behavior of the machine learning model through infinitesimal size perturbations [90].
- **Feature Selection**: Identifies a combination or a subset of p important or most contributing features (from an original dataset holding n features) that train a model with the minimum loss of accuracy. In practice, $p \ll n$ for most feature selection tasks [129].

We distinguished feature selection from feature extraction. Both methodologies aim to improve the performance of the data-driven models by reducing the original feature space. However, feature extraction methods are more closely tied to dimensionality reduction tasks, creating a set with new features from the original data through linear or non-linear transformations that map a significant low-dimensional representation from high-dimensional data while preserving previously defined information [130], [131]. In contrast, despite also aiming to reduce dimensionality, feature selection is performed by dropping data axes based on canonical projections instead of learning mappings.

A. XAI BASED ON APPROXIMATIONS

When data science tasks demand the application of sophisticated and accurate models, their strong non-linearity results in a lack of transparency, requiring explainability approaches. Explainability can be achieved by intrinsically interpretable algorithms for approximating the predictive performance of the original black-box model. More specifically, a black-box model can be used as a ''trainer'' to transfer knowledge to a more transparent and interpretable model that approximates and explains the original predictor's outputs, which is also known as model distillation [132].

A well-known interpretability standard derives from decision trees because the logical sequence of a decision tree can be intuitively interpreted by a human analyst [16]. Other interpretable strategies include logistic regression and rule-based learning [26], which have significant limitations. As an example, decision trees tend to have low generalizability in addition to being prone to overfitting, and logistic regression assumes input data are linearly separable, which rarely occurs in real-world situations. According to Tan et al. [33], explanations based on decision trees have low accuracy, and, in some cases, can be less accurate than those explanations based on linear models.

Guidotti et al. [133] developed a solution through rule-based classifiers using a genetic algorithm to sample the neighborhood of a given instance, train a decision tree, and then generate an explanation. Although considered transparent, rule-based methods have scalability limitations, similarly to linear models. In some cases, generating a massive set of rules is necessary for the obtaining of good classification levels, rendering the analysis unfeasible. Rule-based models are best suitable for approximations in reduced domains [134] and

simpler and transparent models are not sufficiently efficient in handling high-dimensional data with complex relationships.

Lou et al. [135] presented an approach based on GAMs (Generalized Additive Models) as an interpretable alternative to complex regression models. GAMs are linear smoothing models that decompose a predictive function into an aggregation of one-dimensional components defined for each predictive variable [136]. They can then capture the individual non-linear relationships of the variables under modeling. Lou et al. [112] improved the previous version of Lou et al. [135] using a new and optimized mathematical formulation and Caruana et al. [20] chose a GAM-based approach because it is a ''gold standard'' in interpretability. However, GAMs are limited by their low performance in generating explanations for more complex modeling [2].

Interpretable models can perform similarly to their non-interpretable counterparts. Tan et al. [33] presented a comparative study indicating explanations generated by distilled transparent models achieve accurate results in contexts of additive explanations. However, those models build different representations of the latent space, which can affect the predictive performance [137]. Unfortunately, training an interpretable model as a best-fit approximation that mimics complex black boxes is sometimes unfeasible. According to Linardatos et al. [32], the construction of a competitive and transparent model in some domains, such as language processing and computer vision, is very difficult because the gap in performance compared to deep-based models is unbridgeable. Furthermore, knowledge transfer between different domains is another limitation imposed on transparent models.

Post-hoc explainability must be considered when approximation through transparent methods is not feasible [40]. The universe of post-hoc XAI is extensive in the literature. The following subsections present the main classes of approaches created to convert black-box problems into explainable information and the proposed classification extends the taxonomic diagram shown in Figure 7.

B. XAI BASED ON INFORMATION VISUALIZATION

Information visualization maps data into graphical formats, simplifying their representation, hence, helping analysts visually discover trends, patterns, and characteristics [138]. Visualization techniques have been long adopted in multiple application domains, including XAI, where the visualization community has made considerable efforts to using graphics to provide interpretation from Neural Networks [139] to the recently introduced Deep Learning architectures based on Transformers [140]. The use of visualization to explain the training process enables analysts to inspect model learning, thus, monitoring its performance [141].

Marcílio-Jr et al. [142] designed a model-agnostic tool based on coordinated views to visualize similarity between classes. It measures the importance of features using optimization to induce perturbations in individual features that simultaneously minimize the model's performance and perturbation. Chan et al. [118] developed a graphical interface for the inspection of predictions from different density levels. The system provides a summarized overview so that users can ''browse'' from global to local explanations. The interface shows the importance of an instance from different contexts generated through groupings of similar instances into topic vectors of different granularities.

Partial Dependence plots [71], [82], [143] are graphical methods used to understand supervised learning models by visualizing the average marginal effect (partial dependence values) between input variables and predictions [4]. Partial dependence can capture monotonic relationships, but can also obscure heterogeneous effects and complex relationships resulting from feature interactions [40]. Individual Conditional Expectation [83] curves handle this problem by disaggregating the partial dependence output and visualizing the extent to which the prediction of an individual instance changes when the value of a selected feature changes [84]. Heatmaps have been applied to highlight and explore the most relevant elements of neuronal units in image classification problems [9], [106], [134]; however, they are difficult to aggregate, making the visual detection of false positives at scale challenging [90].

Xenopoulos et al. [144] developed GALE (Globally Assessing Local Explanations), a TDA-based [145] methodology for extracting simplified representations from sets of local explanations. The method generates a topological signature of the relationships between the explanation space and the model predictions. Based on a visual inspection of the representations, the parameters of the underlying local XAI method can be assessed and tuned or the similarities among different explainability methods can be quantitatively compared. GALE acts more as a visual assessment tool for well-known local XAI methods than as an explainability method.

Cabrera et al. [146] demonstrated how to detect learning biases and assess fairness with an interactive visualization interface that enables investigations of similar subgroups and impacting features. However, it supports only binary classification and tabular data and suffers from scalability.

Multiple XAI visualization-based methods actively use dimensionality reduction techniques or multidimensional projections [131], [147] to generate interpretable representations of the feature spaces of learning models, such as relationships between neurons and their influence on data [141]. Cantareira et al. [148] developed a method that describes the activation data flow in hidden layers of Neural Networks. The information enables verifications of the network evolution during the training process and representations are created when one layer transmits the information to the next layer. Rauber et al. [98] introduced a similar method that visually explores the way artificial neurons transform input data while they pass through the hidden layers of deep networks.

SUBPLEX [120], [149] is an interactive visual analytics tool that connects multidimensional projections with local explanations and aggregates large sets of local explanations

at the subpopulation level for reducing visual complexity. From aggregated explanations, users can interactively explore explanation groups in detail using feature or instance selections for identifying patterns and comparing local patterns in multiple subpopulations. SUBPLEX did not introduce a new XAI approach or technique itself; instead, it applied well-known XAI techniques and projection methods to generate aggregated explanations in a graphical user interface. However, the tool requires computation for real-time processing of large amounts of data.

UMAP [150] and t-SNE [130] are two robust multidimensional projection techniques frequently used in visualization-based XAI methods [42]. However, dimensionality reduction techniques have usability limits in terms of number of points visualized simultaneously [98], [141]. Explainability through information visualization faces scalability challenges related to dealing with large numbers of elements in addition to adequately describing their relationships [147].

C. XAI BASED ON DECISION BOUNDARIES

Explaining the behavior of machine-learning models by investigating their decision boundaries is a little-explored approach in the literature. Karimi et al. [101] developed DeepDIG, a method based on the generation of adversarial samples sufficiently close to the decision boundary, i.e., synthetic instances between two different classes. More specifically, DeepDIG works on a previously trained Deep Neural Network (DNN) and generates border instance samples with classification probabilities for two distinct classes as closely as possible, resulting in classification uncertainty. Those synthetic border instances are then used for measuring the complexity or non-convexity of the decision boundary learned by the trained DNN.

One of the main properties of DNNs is their remarkable generalization power, achieved through sophisticated combinations of non-linear transformations. As a result, DNNs can map data with complicated and high-dimensional relationships, which introduces the question of whether the complexity of data in the input space is reflected in the transformed (learned) space of the network.

Toward addressing that issue, Karimi et al. [101] designed two metrics to characterize the complexity of decision boundaries – one for the original space (input data) and another for the transformed space. The authors then verified the hypothesis presented by Li et al. [151] concerning the decision boundary resulting from the last layer of a DNN trained with backpropagation converging to the solution of a linear SVM trained on the transformed data. In a similar context, Guan and Loew [152] developed a metric to assess the complexity of decision boundaries, arguing models with simpler borders have optimal generalization ability.

Englhardt et al. [153] proposed a technique to discover the minimum sampling with (almost) uniform density containing border points using original data. The technique is based on optimization for retaining the necessary samples and ensuring the correct delimitation of the decision boundaries. Sohns et al. [154] designed an interactive interface to visualize the topology of decision boundaries and other graphic tools to explore partial dependence and feature importance. However, as a visualization tool in complex domains, the approach has scalability limitations.

Yousefzadeh and O'Leary [99] calculated flip points, which are points close enough to the decision boundaries of trained models such that they can be classified into both classes (considering Neural Networks with two outputs). The study introduced the following issues: flip points can be used to determine the minor change in input data enough to modify a prediction; incorrectly classified data instances tend to have smaller distances from a flip point than correctly classified ones; points close to their flip points are more influential than distant ones in determining decision boundaries between classes; and using flip points as synthetic data during model training can improve accuracy when the model is biased. Flip points exist for any model, and not only for Neural Networks and, if appropriately confirmed, the aforementioned issues can turn flip points into key elements in providing interpretation for checking trust in predictions [99].

D. XAI BASED ON CONTRASTIVE AND COUNTERFACTUAL EXAMPLES

The right-to-information regulations demand meaningful information about the logic behind automatic decisions [155]. Although they do not define "meaningful," it is reasonable to expect explanations will go further than technical aspects and translate machine-learning decisions into human-descriptive language. The goal of introducing a human-centric nature to XAI has motivated researchers to pay more attention to social-based considerations on the requirements of explanations, with notions of contrastive and counterfactual argumentation arising as natural reasoning paths of humans for explaining why or how some decision was made or could be made.

The concept of contrastivity is acquired from social sciences, which establish human explanations derive essentially from contrastive processes [156]. Contrastive explanations clarify why one event occurred in contrast to the other. Therefore, the contrastivity property specifies an explanation should answer questions related to why an event occurred in terms of its possible causes (hypothetical alternatives). As an example, a "reasonable" explanation to a question such as "why did event A happen instead of event B?" would provide the causal reasons that directed the model to event A [122]. In XAI, contrastive methods offer insights into why a model made a specific prediction, highlighting the features that led to that prediction and contrasting them with features that would lead to alternative outcomes.

Similarly, different scenarios can be described for a particular prediction in case of slight modifications in the input data, thereby explaining the possible

"contrary-to-fact" consequences of those modifications. Counterfactual explanations have a long history in philosophy and psychology because, for human explanations follow counterfactual dependence patterns [157]. In XAI, counterfactual methods generate instances or scenarios close to the input for which the output of the classifier changes [90], describing characteristics that will change in the prediction in case of any perturbation, deletion, or inclusion of values in the predictive features [124]. Counterfactual explanations do not explicitly answer "why" a model predicts a decision; instead, the broad goal is to describe a link between what could have happened if a certain input had been changed in a particular way, providing directions toward the desired prediction [128], [157].

Poyiadzi et al. [158] developed a method that generates actionable counterfactual explanations by constructing a weighted graph. It then applies Dijkstra's shortest path algorithm to find the instances that generate explanations according to density-weighted metrics and users' requirements, providing suggestions on how much a change in the input would lead to the users' desired outcomes. Raimundo et al. [159] introduced MAPOCAM (Model-Agnostic Pareto-Optimal Counterfactual Antecedent Mining), a multi-objective optimization algorithm that determines counterfactuals. Its input attributes are handled as a set of cost functions applied in a tree-based search mechanism for identifying the changes that give rise to counterfactual antecedents. Multi-objective optimization has still not been well explored in the counterfactual XAI literature; however, MAPOCAM is computationally expensive and does not support categorical data.

Contrastive and counterfactual methodologies are generally similar, but not equivalent and cannot be used interchangeably. Contrastive explanations are more restrictive than counterfactuals [155], with counterfactual explanations usually being seen as contrastive by nature [160]. Both strategies provide influencing factors through alternative scenarios or conditions for helping users better understand the opearation of machine learning models. That is the reason why XAI literature often uses term "contfactual" to agglutinating explanation solutions covering both concepts independently and their fusion [122].

We observe a synergy between the contfactual methodology and the approach based on the characterization of decision boundaries, previously discussed in Section VIII-C. Both investigate minimal modifications in the input data that modify a prediction context, offering an opportunity for researchers to design contractual methods that benefit from advances in the decision boundaries literature and vice versa. Aïvodji et al. [161] and Dissanayake and Dutta [162] leveraged the fact that counterfactual explanations lie close to decision boundaries and investigated model extraction by strategically training surrogate models using counterfactuals. Such works highlight concerns about privacy in XAI since the information provided by counterfactual explanations can enable adversary attacks aiming to build faithful copies of original (target) models, which might lead to sensitive information leaks.

The interest in contfactual explanations research has grown because of their appealing alignment with human reasoning, which could enhance machine learning transparency by transforming XAI applications into "human-like" explainers [22], [163]. However, the contfactual methodology has limitations. Bhatt et al. [90] indicated many current contfactual methods make crude approximations, since finding plausible counterfactual explanations (feasible in both input data and real world) is non-trivial and computationally expensive. Contfactual explanations are highly domain-specific, which leads to a lack of standardization in evaluation procedures [122]. Furthermore, current implementations of contfactual explanations are model-specific or work for models that are not black boxes in nature.

Verma et al. [157] reviewed and categorized research on counterfactual XAI, describing the desirable properties and evaluating the advantages, disadvantages, and open questions of the current methods. Interested readers can also consult Stepin et al. [122], who conducted an extensive literature review from the theoretical characteristics and differences to the current state-of-the-art of contrastive and counterfactual XAI in addressing explanations of causal and non-causal dependencies. Van Looveren and Klaise [164] addressed some limitations of the counterfactual approach.

E. XAI BASED ON EXPLANATION OF GRAPH MACHINE LEARNING

Graph Neural Networks (GNNs) represent a powerful class of models in Graph Machine Learning (GML) applied to generating predictions on data associated with a graph as its underlying structure. Several real-world applications naturally arise as graph models (e.g., social networks, fraud detection, knowledge graphs, bioinformatics, molecules modeling in chemistry, street maps, document citation, and infrastructure optimization) [165], [166].

The set of different specific approaches involved in GML is large, since feature vectors can be associated with graph nodes (e.g., document content), graph edges (e.g., messages between users in a social network), and/or the whole graph (e.g., toxicity of a molecule). The Machine Learning task may also be a node-level prediction (e.g., predicting the class to which the documents belong), edge-level prediction (e.g., forecasting traffic flow in the streets of a city), graph-level prediction (e.g., forecasting the solubility of a molecule), link prediction (e.g., recommending users who might follow each other), and graph-to-graph interaction (e.g., predicting the side effects of taking two drugs simultaneously).

Accounting for not only the features associated with graph elements, but also the complex interactions defined by the graph structure can make the explainability of GNNs challenging, leading to less extensive literature in comparison to a non-graph XAI scenario.

However, some recent and notable progress has been made in XAI research to providing explanations for GML complex models. Ying et al. [167] proposed GNNExplainer, a perturbation-based method devoted to explaining individual predictions made by trained GNNs that highlights the more influential nodes and edges in the input graph by computing the gradients of the prediction concerning node embeddings. Despite providing valuable insights, the method has some limitations, such as sensitivity to initial node embeddings, need for approximations for large graphs, and difficulty in explaining complex interactions. PGM-Explainer [168], another explainability technique for GNNs, builds a probabilistic model for the desired node to be explained. It generates a local dataset by randomly perturbing the node features of a subgraph that contains the target node multiple times. An interpretable Bayesian network then fits the dataset and explains the GNN predictions by focusing on node explanations.

In contrast to previous methods devoted to the explainability of graph elements or features, SubgraphX [169] aims to identify important subgraphs; It employs the Monte Carlo Tree Search (MCTS) algorithm to explore subgraphs by pruning the nodes. The importance of each subgraph is measured by an efficient approximation of Shapley values [170] that considers interactions within the message-passing mechanism. Although SubgraphX yields human-interpretable subgraphs, its computational cost is higher because of the need to explore different subgraphs through MCTS.

XGNN [171] is a model-level graph explainability technique that employs graph generation, i.e, instead of making computations directly in the input graph, it trains another model to produce graphs that optimize the GNN prediction. Such graphs are expected to contain discriminatory patterns and thus provide the desired explanations. The framework can incorporate constraints to ensure interpretable explanations, such as restraining the node degrees or the number of nodes in the generated graph. However, one of its limitations is XGNN is suitable only for GML graph-level classification problems.

For further details on XAI for the Graph Machine Learning context, we refer to [166], [169], [172], and [173] and to [165], [167], [174], and [175] for metrics and benchmark datasets to assess GML explanations.

F. XAI BASED ON EXPLANATION OF ATTENTION MODELS

In traditional sequence models, such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory Networks (LSTMs), information tokens are sequentially passed from one step to the next. Such a sequential nature limits the extent to which the context can be captured, even with RNNs and LSTMs bearing structures designed to hold information for a longer duration [176], [177].

In contrast, attention-based approaches do not process the inputs sequentially. The attention mechanism enables the model to assign relative importance to different parts of the input sequence and ''pay attention'' to certain parts when making predictions. Each information token is attended to every other simultaneously in parallel, enabling more efficient and scalable computations [65]. In addition, the attention mechanism promotes selective focus on different parts of the input sequence, contributing to context preservation, which is particularly effective for applications involving sequential data in, for instance, the Natural Language Processing (NLP) domain [178].

Vaswani et al. [65] introduced Transformers, a type of Neural Network architecture that relies on attention mechanisms (or scaled dot-product attention) to capture relationships between different words or tokens in a sequence. At a high level, the self-attention mechanism enables a token in a sequence to focus on other tokens in the same sequence, thereby assigning different levels of importance to each token. The transformer approach is not restricted to fixed-size contexts, enabling tokens to directly influence each other, regardless of their distance in the sequence [65].

Transformers have demonstrated state-of-the-art performance across many NLP tasks, but are widely adopted in various domains. Although Vaswani et al. [65] claimed attention mechanisms employed on Transformers could yield more interpretable models, such transparency is questionable [95], i.e., the interpretation of internal workings of a transformer model can be challenging and must be better understood [179], [180].

Vig [179] briefly discussed studies that developed tools to visualize attention in NLP models, from heatmaps to graph representations. The author presented an improved version of BertViz [181], a visualization tool composed of three views following the small multiples design pattern for exploring transformer models at attention-head, model, and neuron levels and also demonstrated an interesting case in which an attention-based model encoded gender biases. However, BertViz has certain limitations. It can show a slow performance when handling extensive inputs or large models and only a few transformer-based models were included in it. The presentation of heat maps of attention weights can be misleading, thus causing unclear interpretations. Furthermore, counterfactual experiments can generate alternative heat maps that yield equivalent predictions [95], although Wiegreffe and Pinter [100] claimed the existence of another explanation does not mean the one provided is meaningless or false.

Pythia [178], a benchmark framework for evaluating Large Language Models (LLMs), includes several open-access and pre-trained transformer-based models, covering a wide range of scales up to 12 billion parameters. The study also highlights the critical role of model size in language modeling performance and provides analyses of gender biases and memorization. Although memorization in LLMs has become a significant concern, few tools enable data scientists to detect and prevent it. Biderman et al. [180] introduced an overview of memorization and proposed measures to understand and predict it.

Garde et al. [182] introduced DeepDecipher, an interactive interface for the visualization and interpretation of neurons in

the MLP layers of transformer models. It provides information on the behavior of neurons toward the understanding of when and why an MLP neuron is activated based on a pre-defined database of sequences and a method that creates a graph of tokens [183]. However, a neuron view may not represent its general behavior, and DeepDecipher does not introduce a novel explanation method.

Since large and complex attention-based models have become increasingly influential in intelligent applications, interpretability must be urgently provided for them. Hundreds of new studies have been recently published and, despite their general success, transformer models must be better understood [180]. Many XAI solutions to attention models apply visualization tools. Visualizing attention weights illuminate one part of the predictive process, but not necessarily provide a reasonable explanation [95]. Chefer et al. [184] proposed a gradient-based method to compute relevancy scores for transformer models. We address the gradient approach in the following section.

Research on how a given transformer model learns and represents data can potentially impact the next generation of software. One of the reasons for the gap in explainability research is the lack of available large models that are also openly accessible for tests and development [178]. For further valuable discussion on attention transparency and explainability, we refer to [94], [95], and [100].

G. XAI BASED ON GRADIENTS AND SIGNAL DECOMPOSITION

Gradient-based methods utilize the partial derivatives of learning models to explain their predictions. They attribute importance to the input features by analyzing the amount to which small perturbations in the input features impact the model's output. Furthermore, computing the gradients of the output concerning the input is analog to verifying the coefficients of a Neural Network model [185], generalizing the deconvolutional network reconstruction procedure [186]. The early gradient-based proposals focused on determining inputs maximizing neuron activity of unsupervised network architectures [187] and generating visualizations for convolutional layers of deep networks [188].

Simonyan et al. [186] introduced the use of gradients to generate saliency maps for supervised models – such an approach is referred to as Vanilla Gradient by the XAI community [189], [190]. It directly computes the model's output gradients through a first-order Taylor expansion [191] around a perturbed instance and a bias term. The product of the gradient and input feature values (with no modifications) is interpreted as a feature importance attribution. Despite its simplicity, the approach lacks fine-grained sensitivity and is prone to noise within the gradients and neither the perturbation procedure, nor bias term were adequately specified [192].

Similarly, T-Explainer [193] relies on Taylor expansions to approximate the local behavior of black-box models and perform feature attributions. However, the method computes gradients through input perturbations in a finite differences-based optimization procedure not dependent on the model's architecture. T-Explainer works with tabular data, although it has limitations with categorical features.

Bach et al. [192] proposed LRP (Layer-wise Relevance Propagation), which explains the predictions of complex non-linear models by decomposing the outputs in terms of input variables. The method is mathematically based on DTD (Deep Taylor Decomposition) [191] for identifying pivotal properties related to the maximum uncertainty state of the predictions. It redistributes the predictive function in the opposite direction through the projection of signals from the output to the input layer by a backpropagation mechanism uniformly applied to all model's parameters [4], [106], [194]. LRP is deemed a model-agnostic method because it avoids *a priori* restrictions on specific algorithms or mappings. However, it was designed as a general concept for black-box architectures based on non-linear kernels, such as Multilayered Neural Networks and Bag of Words, which include several well-known models strongly tied to image classification tasks [192].

Let f be the learning model and $\mathbf{x}$ be the input instance, e.g., the pixels of an image. LRP algorithm assumes the black-box model can be decomposed into l layers of computation to attribute a vector of relevance scores $\mathcal{R}_d^{(l)}$ over each intermediate layer. The attribution process is iterative and starts on the real-valued output in the last layer, from where the calculated scores are backward propagated until they approximate the first (input) layer of the model as follows:

$$\sum_d \mathcal{R}_d^{(1)} = \cdots = \sum_{d\in l} \mathcal{R}_d^{(l)} = \sum_{d\in l+1} \mathcal{R}_d^{(l+1)} = \cdots = f(\mathbf{x}) \tag{8}$$

with d representing the indices of the neurons in each layer.

LRP has been successfully used to generate measurable values describing the processing of variables in Neural Networks because its redistribution strategy follows relevance conservation and proportional decomposition principles, which preserve a strong connection with the model output [195].

Lapuschkin et al. [106] applied spectral clustering to LRP score vectors to identify atypical patterns and behaviors in patterns learned from a pre-trained Neural Network. The study demonstrated unnoticed biases in the training dataset, where many images from a specific class had a URL source tag. As a result, new images not associated with that class, but artificially manipulated to presenting the source tag, were incorrectly classified. The resulting LRP relevance scores were rendered through heatmaps of same dimensionality of the input data (relevance maps) as an interpretable visualization tool. The study of Lapuschkin et al. [106] illustrates a clear example of how explainability tools can assist data scientists in discovering hidden biases in learning models. Montavon et al. [191] and Kohlbrenner et al. [195] conducted reviews evaluating LRP approaches applied to Neural Networks.

The convolutional layers of Convolutional Neural Network (CNN) architectures apply specialized filters across input images to learn complex visual patterns, such as spatial information and high-level semantics. DeConvNet [188] visualizes the activity of intermediate layers of a CNN by using the same layer components in reverse order. Grad-CAM (Gradient-weighted Class Activation Mapping) [196] generates explanations for any CNN-based model by attributing importance scores to each neuron of the final CNN layer. The attribution process uses class-specific gradient information [197] from the backward pass of backpropagation for producing a localization map, highlighting the most influential regions in the input image for the model's decision.

Specifically, Grad-CAM computes an importance score matrix w_c^k generating a localization map $\mathcal{L}^c_{\text{Grad-CAM}} \in \mathbb{R}^{m \times n}$, where m and n represent, respectively, width and height of an input image belonging to any target class y^c, based on the gradients from neuron weights on each feature map $\mathcal{M}^k$ of the last convolutional layer, which is computed before the application of SoftMax function, $\frac{\partial y^c}{\partial \mathcal{M}^k}$, by passing back over m and n dimensions as follows:

$$w_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial \mathcal{M}_{ij}^k} \tag{9}$$

where Z represents the number of pixels in the feature map, which is used for outputting normalization.

Importance scores w_c^k represent a partial linearization of a CNN and describe the importance of feature map k for class c. Finally, the importance scores are linearly combined by globally averaging them with their corresponding feature maps, passing them in a *ReLU* layer and plotting the final scores map in a heatmap:

$$\mathcal{L}^c_{\text{Grad-CAM}} = ReLU\left(\sum_k w_c^k \mathcal{M}^k\right). \tag{10}$$

Grad-CAM and its recent variants [198], [199] generate interpretable visualizations by overlaying the scores heatmap on the original input image, providing visual information that enables identifying the regions of the image that are more influential in the decision process. Grad-CAM does not require architectural modifications or retraining; although it is agnostic regarding different CNN models, it is restricted to working on CNNs. Furthermore, it depends on activating a ReLU layer for proper gradient sensitivity, may have limitations for accurately determining the coverage of class regions, and is prone to instability when locating multiple instances of an object within an image [32].

Input × Gradient [200] attributes feature importance by computing the element-wise product of the model's output gradients and the corresponding inputs, a process known as sensitivity mapping. However, sensitivity maps are prone to instability due to input noises leading to fluctuations in the partial differentiation. Smilkov et al. [201] designed the SmoothGrad approach to run on top of existing gradient methods, enhancing them by averaging multiple slightly perturbed input samples generated by the addition of small levels of Gaussian noise [202]. The process makes the final explanations less sensitive to noise in the input data and, formally, averages $\mathcal{M}_k$ sensitivity maps resulting from those perturbed samples:

$$\mathcal{M}_{SG}(\mathbf{x}) = \frac{1}{k} \sum_1^k \mathcal{M}_k(\mathbf{x} + \mathcal{N}(0, \sigma^2)) \tag{11}$$

where $\mathbf{x}$ represents the input instance, k is the number of perturbed samples, and $\mathcal{N}(0, \sigma^2))$ is the Gaussian noise with σ as the standard deviation.

The Integrated Gradients method [185] differs from other gradient-based attribution methodologies by determining a set of interpolated samples between the input under explanation and a baseline (usually an instance with "neutral" values, e.g., mean values or zeros). It computes the gradients of interpolated samples and integrates them along the path from the baseline to the target input. Let $\frac{\partial f(\mathbf{x})}{\partial x_i}$ be the gradient of a model f along the i-th dimension of input data $\mathbf{x}$ and $\mathbf{x}'$ representing the baseline. Integrated Gradients is then defined as

$$\mathcal{IG}_i(\mathbf{x}) = (\mathbf{x}_i - \mathbf{x}'_i) \times \int_{\alpha=0}^{1} \frac{\partial f(\mathbf{x}' + \alpha \times (\mathbf{x} - \mathbf{x}'))}{\partial \mathbf{x}_i} d\alpha \tag{12}$$

Integrated Gradients is suitable for generating global or local feature attributions for (theoretically) any Neural Network model. However, it can generate inconsistent explanations, since its performance is closely tied to the baseline choice, which depends on the domain context.

DeepLIFT (Deep Learning Important FeaTures) [200], [203] is based on the concept of importance scores derived from LRP. The method aims to explain Deep Neural Networks (DNNs) propagating attributions at each layer of the deep model to compare the difference between a neuron activation and a "reference activation" used as a baseline. DeepLIFT applies non-linear transformations based on a chain rule to network multipliers. More specifically, it computes multipliers for any neuron to its immediate successors (a target neuron) using backpropagation, which is similar to the application of the chain rule for partial derivatives. According to Lundberg and Lee [109], this composition can be equivalent to linearizing the Neural Network's non-linear components.

Shrikumar et al. [203] also defined the rescale rule as an improvement of the LRP's chain rule upon computing the gradients regarding the output of backpropagation. Chain-rule methods generally do not hold for discrete gradients [204], making DeepLIFT and LRP violate the implementation invariance property [205]. Saliency maps and input gradient-based methods suffer from the so-called neuron saturation problem [203]. Reference-based methods such as Integrated Gradients and DeepLIFT address that limitation by comparing input features with reference values, avoiding the saturation issue [206]. However, the reference

activation (baseline) choice is made heuristically, leaving significant open problems, such as empirical computation of a good baseline and propagation of importance beyond simply applying gradients [203].

H. XAI BASED ON SIMPLIFICATIONS

Explainability through simplification comprises techniques in which a new explainer model is built based on a trained model to be explained [2]. The goal of the simplified model is to perform a behavior similar to that of the original model with less complexity, i.e., it must retain a predictive performance similar to that of the original model, but based on more transparent structures. Thiagarajan et al. [207] developed TreeView, a tool that visually interprets complex models. It identifies discriminatory factors across data classes using sequential elimination through a hierarchical partitioning of the feature space, clustering the instances into groups for each factor, where undesirable associations are discarded.

LIME (Local Interpretable Model-Agnostic Explanations) [97] is among the well-known and widely applied explainability techniques. It determines an interpretable linear model that locally approximates an original model. LIME generates a neighborhood of synthetic samples around the instance under explanation through perturbations on the instances of the original dataset. The synthetic samples are then classified by the original learning model, which weights them by applying a weighting kernel according to their proximity to the point under explanation. LIME then determines a linear model on the neighborhood, minimizing a non-fidelity function, and the predictions are explained through the linear model interpretation.

More specifically, let f be the trained model, $g \in G$ be an interpretable model, and G be a class of potentially interpretable models, such as linear regression or decision trees. Toward explaining an n-dimensional instance $\mathbf{x} = (x_1, \ldots, x_n)$, an interpretable model g is determined to minimizing loss function $\mathcal{L}$ according to

$$\mathcal{E}(\mathbf{x}) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_{\mathbf{x}}) + \Omega(g) \tag{13}$$

where $\pi_{\mathbf{x}}$ defines the weighting kernel centered on $\mathbf{x}$ responsible for maintaining the explanation's local fidelity, and Ω is a complexity term (which should be kept low) applied to g.

Some authors have also classified LIME as a Local Surrogate Model, defined as the class of methods that explain individual predictions through a locally trained substitute model [40], called local fidelity. LIME has a simple and informative graphical interface. Figure 8 displays an example of a LIME-generated explanation for an instance of the well-known Iris[1] dataset that, for simplicity, was adapted for containing only two classes. Regarding a binary classification task, LIME provides explanations using a pattern of two colors (orange and blue, in this case).

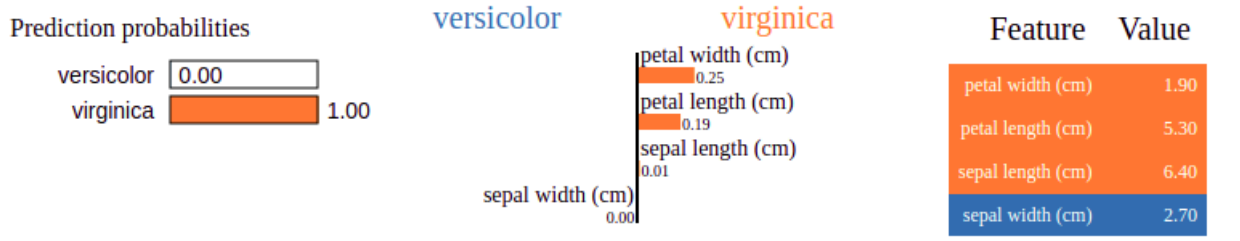
[1]https://archive.ics.uci.edu/ml/datasets/iris, visited on June 2023

Figure 8a shows the classification probabilities of the instance under investigation, predicted by the black-box model as belonging to Virginica class. Figure 8b displays the most relevant attributes in order of importance for the prediction, with the float point values on the horizontal bars informing the LIME's importance values attributed to those features. Figure 8c provides an overview of the instance under investigation with the original values of each feature. The colors are distributed according to the contributions, i.e., attributes in orange contributed to Virginica class and those in blue contributed to versicolor class. Color coding is consistent across charts. However, the way each attribute contributes positively or negatively to the LIME results is unclear. The authors also developed two LIME extensions [208], [209], including an improved version with provides clearer textual explanations based on rules [210].

Among the main advantages of LIME is its flexibility. Any interpretable model can be used as a surrogate and even changing the original learning model, explanations can be generated for the dataset by the local interpretable model generated by LIME. The method is one of the few that works on tabular, textual, or image data [40], although it is not suitable for applying to time series, since LIME independently constructs simplified models for each instance [211]. As an example, LIME builds neighborhoods in image classification by segmenting the input image into superpixels and perturbing the segmented image by randomly switching the superpixels with a background color. The Boolean states of the superpixels are then used as attributes of the simplified model [194]. That type of strategy can be extended for enabling other XAI methods to operate with images.

Although LIME is considered an outstanding solution for XAI, it has some significant limitations. It is stable in explaining linear classifiers, but unstable in other cases, i.e., it sometimes provides explanations that do not align with human intuition and can change its explanations entirely only by running the code a few times, owing to the sampling variance [25], [90]. Aas et al. [117] argued LIME does not guarantee perfectly distributed effects among variables. Furthermore, different models can fit the sampled data, with LIME randomly selecting one of them without guaranteeing it is, in fact, the best local approximation. No solid theoretical guarantee indicates a simplified local surrogate model adequately represents more complex models [4], i.e., original and surrogate models always produce similar predictive behaviors.

Defining a meaningful neighborhood around an instance of interest is a complex task. LIME bridges such a difficulty by constructing a neighborhood around the point under explanation using the center of mass of the training data. The strategy can contribute to the instability of the technique by generating samples considerably different from the instance of interest, despite it increasing the probability of LIME learning at least one explanation [40]. LIME also relies on simplistic assumptions regarding the decision boundaries of learning models, assuming they are locally linear. However, decision



(a) Classification probabilities. **(b) The most important features.** **(c) Original values of the attributes.**

FIGURE 8. LIME explanation for an instance from Iris dataset, classified as belonging to Virginica class. LIME provides visualization tools with different information on model classification, locally attributed importance values, and the input instance itself.

boundaries of models such as Neural Networks can be highly non-linear, even locally, and a linear approximation in this context might lead to unstable explanations [212].

LIME output values lack comparative meaning; it is not straightforward to understand what the values attributed to each input feature mean (Figure 8b) and the relationship between those values and the model prediction. Moreover, LIME linear weighting increases the influence of unperturbed samples [194]. Since no reasonable way estimates the weighting kernel or even an appropriate choice of its width ratio [40], LIME then chooses critical parameters, such as weighting kernel, neighborhood size, and complexity term heuristically, leading to inconsistent behaviors, which might affect the local fidelity [25], [109].

Deterministic versions of LIME [213], optimization [214], [215], and learning-based [216] strategies have been proposed to reduce instability; however, those alternatives have the cost of increasing the number of parameters to be tuned.

I. XAI BASED ON SHAPLEY VALUES

Derived from classical game theory modeling, Shapley values [170] describe a way to distribute the total gains/costs of a cooperative game among players, satisfying fairness criteria [119], i.e., determining Shapley values is a cost-sharing problem [205]. According to Moulin [217], cost-sharing problems are central subjects in several areas that require splitting joint costs and proportionally allocating their shares among each individual contributor.

As an example, electricity is a public utility with a long production chain that, in a simplified way, starts at the power-generating units and moves through transmitters and distributors until it reaches the final consumer. Determining how much the consumer will pay and fairly distributing this value to each link in the production chain is a typical cost-sharing problem.

A Shapley value represents the average marginal contribution of a player evaluated over all possible combinations of players, i.e., it is a weighted average of individual contributions related to all possible compositions of individuals [40]. An aspect of Shapley values lies in their solid theoretical foundation, which axiomatically ensures a fair distribution of gains/costs among the participants of a collaborative game. According to Kumar et al. [218], a collaborative game comprises a set of n players and a characteristic function v, which maps subsets $S \subseteq \{1, \ldots, n\}$ into real values $v(S)$, satisfying $v(\varnothing) = 0$. The characteristic function describes the extent to which the final gain can be attributed to individual players cooperating as a team in the game. Therefore, Shapley values represent a method of distributing the total value of cooperation, $v(\{1, \ldots, n\})$, among n individuals.

Let us consider $v(i)$ the characteristic function applied to attribute i (a player) from a subset of S attributes, i.e., $i \in S$. The Shapley value can be computed as a weighted average between the attribute's i marginal contributions regarding every possible subset of attributes $S \subseteq \{1, \ldots, n\}$ and the number of permutations of S:

$$\phi_v(i) = \sum_{S \subseteq n \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \quad (14)$$

where $\phi_v(i)$ is the Shapley value of the i-th attribute, $v(S)$ is the expected value of the characteristic function conditional on subsets $S \subseteq \{1, \ldots, n\}$, i.e., $\mathbb{E}[v(S)]$, n represents the total number of attributes, and $|\ .\ |$ denotes cardinality [119]. Note $v(S \cup \{i\}) - v(S)$ describes the marginal contribution of a player to a combination of players S, i.e., variation $\Delta_v(i, S)$ generated when i is included in S [218].

Figure 9 illustrates the concept behind the calculation of Shapley values in a set holding three attributes $\{a, b, c\}$. Each possible combination of attributes must be considered so that the attributes' individual contributions can be verified, i.e., all possible subsets S, with $|S|$ ranging from 0 to n ($n = 3$ in this case), must be computed and verified.

Each vertex in Figure 9 depicts a combination of attributes and each arrow corresponds to an attribute inclusion not present in the previous combination. Note the original approach for the calculation of Shapley values requires verifying 2^n combinations, which is the number of possible subsets of $\{1, \ldots, n\}$, making the Shapley value computation an NP-hard problem (since n grows, Equation 14 tends to be unfeasible).

However, calculating the Shapley value in datasets containing few attributes is relatively simple. Let us consider the example shown in Figure 9. The Shapley value of attribute a is computed assessing the marginal costs among

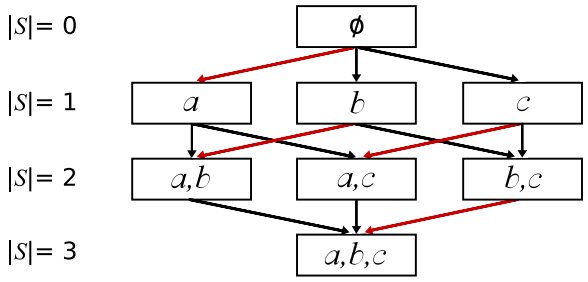


FIGURE 9. Diagram with all possible combinations of attributes in a dataset with three attributes. The calculation of the Shapley value of a specific attribute requires computing the marginal gain/cost of its inclusion in combinations that do not contain it.

all combinations of attributes $S \subseteq n \backslash \{a\}$, leading to a subset containing attribute a, in this case $\{\varnothing\}$, $\{b\}$, $\{c\}$, and $\{b, c\}$. Red arrows in the Figure 9 indicate the path between a combination without attribute a and another with it, i.e., the marginal contribution of including a in a combination previously without it. Applying Equation 14 leads to the following setup for the marginal contributions and weighting factors:

$$\begin{aligned}\phi_v(a) = &\frac{0!\,(3-0-1)}{3!} \times \Delta_v(a, \{\varnothing\}) \\ &+ \frac{1!\,(3-1-1)}{3!} \times \Delta_v(a, \{b\}) \\ &+ \frac{1!\,(3-1-1)}{3!} \times \Delta_v(a, \{c\}) \\ &+ \frac{2!\,(3-2-1)}{3!} \times \Delta_v(a, \{b, c\}) \\ = &\frac{1}{3}\Delta_v(a, \{\varnothing\}) + \frac{1}{6}\Delta_v(a, \{b\}) + \frac{1}{6}\Delta_v(a, \{c\}) \\ &+ \frac{1}{3}\Delta_v(a, \{b, c\}) \end{aligned} \tag{15}$$

where $\Delta_v(a, S)$ is the marginal contribution of a conditional in the subset of attributes $\{S\}$ (see red arrows in Figure 9). Note a Shapley value is not only the difference in prediction output when an attribute is removed from the model, but also a weighted sum of marginal costs [40].

Several authors have studied applications of Shapley values in Machine Learning [84], [109], [205], [217], [218] and cited the following theoretical properties as desired for a cost-sharing problem solution, satisfied by Shapley values. These axioms can be considered definitions of fairness in cost sharing.

- **Accuracy**: The sum of Shapley values for all attributes equals the total cooperation value, $\sum_i^n \phi_v(i) = v(\{1, \ldots, n\})$. Accuracy means the full value of a game is divided among its players.
- **Missingness**: For an attribute i, if $\Delta_v(i, S) = 0$ for all subsets S, then the attribute will not impact the modeling result, i.e., $\phi_v(i) = 0$.
- **Consistency**: For an attribute i and a non-decreasing characteristic function v, contribution $\phi_v(i)$ should be increased only if the value of i increases and the values of all other attributes are fixed. This axiom implies if v is monotone on i, $\phi_v(i)$ increases if i increases.
- **Additivity**: For an attribute i and two characteristic functions v and t, $\phi_v(i) + \phi_t(i) = \phi_{v+t}(i)$, where $(v + t)(S) = v(S) + t(S)$. This axiom defines arithmetic sum.
- **Symmetry**: For two attributes i and j, if $\Delta_v(i, S) = \Delta_v(j, S)$ for any subset of attributes S, then the contributions of i and j must be equal, $\phi_v(i) = \phi_v(j)$.

In the machine-learning context, Shapley values quantify, for each attribute, the changed value in the expected prediction when learning models are conditioned to combinations of that attribute [109]. The applied models were equivalent in terms of hyperparameters and training data (the complete dataset). The difference lies in the combination of the attributes included in each model. Shapley value modeling has a long application background – Lipovetsky and Conklin [219] used Shapley values to analyze the global importance of attributes in linear regression models and Štrumbelj and Kononenko [220] measured feature effects in classification tasks.

More specifically, a Shapley value attributes an importance value to an input attribute, representing its contribution to the final prediction by including it in the model. The machine learning model under explanation is taken as the characteristic function for the calculation of importance attributions by Equation 14. It is applied to subsets with and without an attribute of interest, i.e., the attribute's marginal contributions, thus extracting a weighted average between contributions [119]. Since all combinations of attributes must be computed for the extraction of the Shapley values of all dataset attributes, the computational cost increases exponentially as the number of attributes increases, hampering operation modelings on high-dimensional data [40].

Toward overcoming that limitation, Štrumbelj and Kononenko [221] developed an approximated version based on Monte Carlo sampling. The estimation of Equation 14 using sampling assumes predictions are generated from instances (randomly sampled) containing randomly selected attribute permutations (except for the attribute under investigation) [17], [84] rather than all n original input features. Therefore, the Shapley value of an attribute is iteratively estimated from the randomly selected samples in each iteration. Variations in the predictions are weighted for each sample according to the probability distribution of the data and the result is computed as an average. The procedure is repeated for each attribute for the estimation of all Shapley values [40].

Lundberg and Lee [109] formulated SHAP (SHapley Additive exPlanations), one of the most successful approaches based on Shapley values. It estimates them approximating the original learning model through a conditional expectation function over vectors with a permutation of simplified attributes. It then measures the gain/loss in a prediction, simulating presence and absence of attributes by sampling the values of the marginal distribution of each attribute. Note conditional expectation is the usual estimator that summarizes the probability distribution in prediction

applications. However, SHAP assumes feature independence and uses a marginal distribution to replace the conditional distribution [117], enabling the conditional expectation approximation to estimate the Shapley values directly through an Additive Feature Attribution modeling.

Additive XAI methods assign an effect to each attribute (see Equation 7) and the sum of all feature attribution effects should lead to a value that makes sense for the original model prediction [109]. Note the correspondence between cost-sharing and attribution problems – Shapley values distribute collaborative game costs among its players, the learning model can be taken as equivalent to the characteristic function, with the game (total cost) as the prediction value, the players as input features, and cost shares being the importance attributions [205].

Formally, let f be the learning model under explanation, g be the explainer model, and $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbf{X}$ be the n-dimensional instance to be explained. Using $\mathbf{x}'$ as a simplification of $\mathbf{x}$, as defined by [208], i.e., $\mathbf{x}' \approx \mathbf{x}$, SHAP defines a mapping function $h_{\mathbf{x}}$ for original instance $\mathbf{x} = h_{\mathbf{x}}(\mathbf{x}')$, such that $g(\mathbf{x}') \approx f(h_{\mathbf{x}}(\mathbf{x}'))$. Even if the simplified instance retains less information than the original instance, $h_{\mathbf{x}}$ ensures no significant loss of information occurs. SHAP then generates an explanation model g that locally approximates original model f determining the Shapley values for each attribute of $\mathbf{x}$ additively according to

$$f(\mathbf{x}) = g(\mathbf{x}') = \phi_0 + \sum_{i \in n} \phi_i x_i' \tag{16}$$

where ϕ_0 represents the prediction expected value, $\mathbb{E}[f(\mathbf{X})]$, and ϕ_i is the Shapley value related to attribute x_i, calculated by Equation 14, where the marginal gain is estimated by $f(h_x(\mathbf{x}')) = f(\mathbf{x})$ as a characteristic function. Determining $\mathbb{E}[f(\mathbf{X})]$ for an arbitrary dataset is not a trivial task; in practice, SHAP estimates the prediction expected value through the average model output across training dataset $\mathbf{X}$ when feature values $\mathbf{X}_i$ are not known.

Equation 16 indicates SHAP approximates learning model f through a linear additive model g, enabling locally estimating the predicted value of f based on the Shapley values of an instance's attributes as parameters of a linear model g. Since the application of Equation 14 can be costly because of the large number of possible combinations, SHAP employs an approximation strategy based on Monte Carlo integration of a permutation version of Shapley's classical equation, with samplings taken separately for each attribution [109].

According to Lundberg and Lee [109], SHAP locally estimates the contribution of each feature, respecting the *local accuracy*, *missingness*, and *consistency* axioms (SHAP can also estimate feature importance at the global level [119]). In contrast to LIME, SHAP allows contrastive explanations, i.e., comparing a prediction in the context of a specific subset of instances or even a single instance instead of only comparing predictions with the entire dataset's average prediction [40], [218]. LIME is not necessarily locally efficient, for its explanation values do not add up to the original prediction.

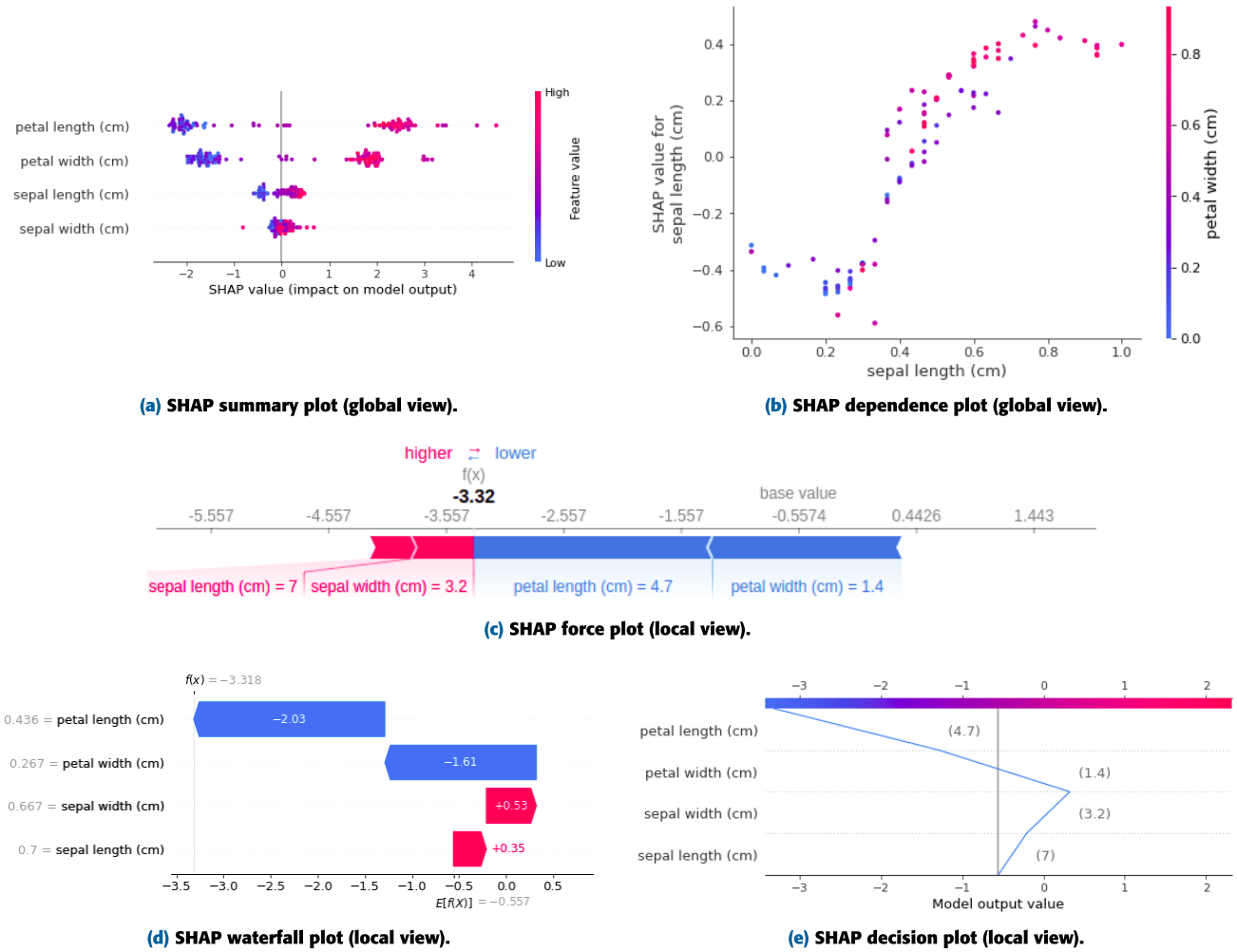
By transforming explanations into an additive linear modeling, SHAP promotes a connection between Shapley values and LIME [40]. Lundberg and Lee [109] developed KernelSHAP, a SHAP version based on the concept of LIME. Although the LIME formulation differs from the Shapley value one, both LIME and SHAP are additive attribution methods. However, LIME heuristically chooses weighting kernel, loss function, and complexity term, violating the consistency axiom and affecting local accuracy, resulting in unstable explanations [40], [109]. Equation 14 is a difference of means. Because the mean is the best least-squares point estimated for a dataset, a weighting kernel can be found by the least-squares method. Lundberg and Lee [109] demonstrated how to determine weighting kernel, local loss function, and complexity term in the context of Shapley values.

Given the linear formulation of LIME, KernelSHAP can estimate Shapley values through regression-based solutions, which is computationally more efficient than directly computing the classical equation of Shapley values. Note SHAP and LIME (the original version) explain predictions differently. LIME indicates the feature that is most important for a prediction, whereas SHAP indicates the contribution of each feature to the prediction. Although both methods compare predictions under an explanation with an average probability, SHAP verifies the difference between predicted and expected values of the global average prediction. Simultaneously, LIME explains the difference between prediction and local average prediction (from neighborhood sampling) [117].

SHAP is currently one of the leading state-of-the-art methods in feature attribution/importance XAI due to its tolerance to missing values and Shapley's theoretical guarantees regarding local precision and consistency [25]. SHAP Python library[2] offers a range of methods and a valuable set of graphical tools for visually analyzing explainability (see Figure 10). Iris dataset was used for the generation of the charts in Figure 10, which, for simplicity, was adapted to containing only two classes, namely, Versicolor and Virginica. The setting is identical to that of the LIME example in Figure 8.

Figure 10a displays explanations for all dataset instances, summarizing the importance of the attributes, indicated on the left side of the chart and ordered vertically according to their global mean importance. Each point represents the Shapley value attributed to each variable from a data instance. The points were horizontally dispersed according to the Shapley value; therefore, the more distant from zero in the positive direction, the more influential the attribute on the predicted class of the instance. Values farther from zero toward the negative side indicate the attribute has more importance to classes other than to the predicted one. Finally, points are accumulated vertically for indicating Shapley value density distribution per attribute and colored according to the attribute's values, from smallest to largest.

[2]https://shap.readthedocs.io/, visited on February 2024



(a) SHAP summary plot (global view).

(b) SHAP dependence plot (global view).

(c) SHAP force plot (local view).

(d) SHAP waterfall plot (local view).

(e) SHAP decision plot (local view).

FIGURE 10. SHAP library provides visualization tools to transform attributions into graphical information, describing important features and their relationships, whether locally or globally. In this example, SHAP visualizations were used to explore the most important features of Iris dataset (global view) and important ones of a sample classified in Virginica class (local view).

Figure 10b shows the average partial relationship (marginal interaction effects) between one or more input variables, where each point represents the prediction of a data instance. The chart considers all instances and describes the global relationships between the variables with respect to the model prediction. Instances were selected from Virginica class in the example. The abscissa axis depicts the sepal length variable and the ordinate axis represents the Shapley value attributed to the respective sepal length variable, indicating the extent to which the sepal value modifies the prediction for each instance. Values farther from zero indicate the variable is important for Virginica class. The points are colored according to the petal width, which is the most significant attribute of the interaction effect related to sepal length.

The chart in Figure 10c provides a local analysis of a specific data sample so that the contribution of each variable to a single prediction can be understood. The force plot shows the predicted value of the instance under explanation, the base value (expected output for the model's average prediction in the training data), and the value of each variable. Under the horizontal line, the colors of the arrows indicate influence of the variables – red color denotes attributes that contributed positively and blue indicates those that contributed negatively. The larger the arrow, the more significant the variable impact. Although the force plot is a succinct metaphor, its horizontal arrangement is inefficient for handling many variables simultaneously. Such a limitation is avoided in the chart in Figure 10d, which shows similar information, but organized vertically in a "waterfall" format. The lines in Figure 10e represent each variable effect, enabling the visualization of the profile of feature importances from one or more instances simultaneously.

We used a Random Forest binary classifier model to generate the chart sequence shown in Figure 10. The expected output of a binary learning model is a probability value between 0 and 1, indicating which of the two classes the samples is most likely to be classified. However, the final SHAP values are different, since, by default, SHAP explains

a classifier's prediction in terms of its marginal result before applying the output activation function. Therefore, SHAP units are log-odds contributions rather than probabilities. According to Aas et al. [117], such a design choice is not appropriate for promoting a direct interpretability of SHAP results due to the difficult interpretation of the output in the log-odds space. Explaining a model's probability output rather than the log-odds output can yield more naturally interpretable explanations [114].

The SHAP authors published other studies with optimizations and new tools. Lundberg et al. [222] proposed TreeSHAP, a specific and faster version of SHAP for complex models based on gradient-boosted tree ensembles that can analyze interaction effects, since it considers dependency relationships among features, which is a well-known limitation of permutation methods. TreeSHAP breaks the feature's contributions into its main and interaction effects by taking advantage of the hierarchy in the decision tree's features for estimating some degree of dependency (but not all dependencies) among the inputs using valid tree paths (weighted average of the final nodes reachable by the permutation subsets) [117]. That study also presents other contributions, such as supervised clustering, which addresses one of the most challenging problems within the unsupervised clustering context. Supervised clustering uses the feature attribution concept to convert input variables into values with the same units of the model output and then determines the weights (metric distance) toward direct comparisons of the relative importance among variables with different metric units.

Lundberg et al. [14] developed an improved version of SHAP, extending the concept of local explanations to separate and capture the individual interaction effects of variables not only on single instances, but also on pairs of instances, providing explanations in a matrix of feature attributions. Chen et al. [113] extended SHAP from the perspective of DeepLIFT, creating DeepSHAP, devoted to providing explanations for Deep Learning-based models and Gradient-boosted Trees [64].

Lundberg and Lee [109] revealed a connection between DeepLIFT and Shapely values such that DeepLIFT can be considered a fast approximation of Shapely values [203]. DeepSHAP explains deep models by performing DeepLIFT using a baseline reference – in this case, the baseline is a subset of samples with (not necessarily) randomly adjusted values, called ''background distribution.'' The instance under analysis is explained from a set of variables to be ''missed,'' which refers to corresponding values in the background distribution. The SHAP value of the instance was obtained for each sample from the background distribution and the final value is calculated by averaging the importance attributions from the background distribution. DeepSHAP is compatible with PyTorch and TensorFlow; however, it can be less accurate than SHAP due to more complex approximations.

Chen et al. [114] presented Generalized DeepSHAP (G-DeepSHAP), a local feature attribution method developed as an improvement over DeepSHAP and DeepLIFT that explains the complex distributed series of models. In contrast to defining the absence of a feature by masking features according to a single baseline, G-DeepSHAP selects a baseline distribution using k-means clustering. The sample under explanation is compared with that distribution of baselines, which decreases the bias of relying on a single baseline and enables feature attributions to answer contrastive questions. G-DeepSHAP generalizes the rescale rule introduced in DeepLIFT for explaining series of models architectures by propagating attributions to a series of mixed-model types rather than only layers in a deep model. The group rescale rule of G-DeepSHAP also reduces the dimensionality of highly correlated features; however, a limitation is G-DeepSHAP does not guarantee it satisfies Shapley's desirable axioms.

Among the vast range of recent publications based on Shapley theory, Casalicchio et al. [84] proposed a local Shapley-based approach to measure the feature importance of individual observations, providing visualizations through visual tools related to Partial Dependence and Individual Conditional Expectation plots. Hamilton et al. [194] extended SHAP (and LIME) for explaining deep CNN models applied to similarity image searches and retrievals.

J. PROBLEMS IN XAI RELYING ON SHAPLEY VALUES

Explaining predictions based on Shapley values is a popular topic in XAI research. This subsection extends the discussion on that class of methods due to the multiple proposals derived from the Shapley concepts in the literature, providing the reader with an in-depth review of Shapley-based XAI. The previous section presented Shapley's theory, evolution, and advantages of application for explaining Machine Learning. However, the approach and SHAP, its main derivation, also have significant limitations.

SHAP can be applied to non-structured data, such as text and images, relying on additional assumptions and heuristics for generating a feature set. Slack et al. [223] proposed an adversarial mechanism to generate a biased classifier that cannot be detected by SHAP, thus highlighting its vulnerabilities and the need for validation in XAI. A Shapley value allocates an importance value to a variable rather than an interpretable prediction model such as LIME. Therefore, most methods derived from Shapley values cannot verify the way a prediction changes through modifications in input data. KernelSHAP addresses that limitation enabling LIME to estimate Shapley values [40].

Amparore et al. [25] described the advantages of using SHAP in different scenarios, suggesting it is the most suitable choice when the analytical objective is local concordance, achieving exceptional levels of concordance. However, the study also reported SHAP is not much more stable than LIME and its alleged advantage can be exploited in practice only for datasets with few variables. The exact computation of Shapley values is resource-intensive – Aas et al. [117]

indicated its intractability for datasets with more than only ten variables. In most cases, only approximated solutions are feasible [40]. Hooker et al. [224] reported SHAP exhibits a deterministic behavior on low-dimensional data; however, when applied to high-dimensional data, it uses statistical sampling techniques based on Monte Carlo integration, tending to unstable explanations.

The correlation among features can be a severe problem, since importance tends to be split among correlated features, masking the true importance of each feature. Hooker and Mentch [225] showed explainability from feature importance methods that rely on permutation-based strategies, such as SHAP, can be highly misleading. They addressed the creation of subsets of features that hold impossible or unlikely combinations within the original data context. Specifically, permuting the set of input features in SHAP is performed naively and might result in subsets combining features that do not necessarily make sense in the real world. The strategy works well for data with total independence among the variables; however, statistically independent features are rarely found in observational studies and machine learning problems [117] and assuming complete independence between features is similar to ignoring the ''whole is more significant than the sum of its parts'' concept in predictive modelings.

When submitted to combinations of correlated attributes following unrealistic configurations, the learning model is forced to extrapolate to unknown regions of the learned space, even assuming independence in cases with at least a few of the features with a high degree of correlation. Since XAI permutation-based methods are sensitive to the way the model extrapolates, the extrapolation behavior becomes a significant source of error, leading to the generation of explanations based on the capture of unwanted or distorted information [226]. Furthermore, ignoring dependency structures by assuming independent distributions, as in SHAP, is a property for which the consequences must be carefully studied [205].

A solution to the limitations related to assuming feature independence is conditional sampling, in which variables are conditionally sampled according to those already in the explanation setup. However, it violates the symmetry axiom [227]. Wojtas and Chen [121] proposed an interesting dual-net mechanism for learning and selecting optimal feature subsets in feature-importance tasks, although the dual-net architecture imposes a computational burden on the training. Aas et al. [117] claimed, in theory, TreeSHAP considers dependence among input features, but, in practice, it is potentially inaccurate when the variables are dependent. The authors also extended KernelSHAP using TreeSHAP elements to handle the dependent attributes; however, the solution suffers from computational complexity.

Kumar et al. [218] and Kaur et al. [228] argued Shapley values are not a natural solution to human-centric explanations due to the lack of clarity in the analysis of the method, which may lead analysts to confirm biases or even some overconfidence. Kumar et al. [218] introduced mathematical problems related to XAI that rely on Shapley values, demonstrating the solutions for avoiding those mathematical problems introduce more complexity with no significant gain in explainability capacity. As an example, when a set of variables is arbitrarily large, a meaningful set must be selected for providing a concise explanation. However, explanations can change considerably in function of the selected variables. Furthermore, it is not clear whether two statistically related attributes can be considered individually, such as in SHAP's additive attribution modeling, even when the feature independence assumption has no direct impact on the final result.

Kumar et al. [218] claimed it is unclear whether the solution of computing an average of the sums describing ''all possible explanations'' is an acceptable way to provide explanations in the context of Machine Learning. Kumar et al. [86] verified information losses of Shapley values concerning interactions among correlated attributes and developed Shapley residuals, a proposal that is not an explanation or evaluation method per se. Instead, it quantifies the information lost by Shapley values. Shapley residuals highlight the limitations of Shapley values, indicating when importance attributions can rely on missed relationships. In such scenarios, Shapley-based explanations should be taken with some skepticism.

What elements influence Shapley values? How can features distribution influence a Shapley value? How can a Shapley value explanation change according to different predicted outputs of a same model? Kumar and Chandran [119] highlighted the difficulty in addressing those questions, which is related to Shapley values (Equation 14) lacking a closed solution within feasible computational times and numerical estimates being costly. The authors demonstrated in addition to depending on the attribute's values and the learning model, Shapley values depend on data distribution. As an example, when the overall variance is low, most observations fall into a small region of the space, implying the probability curve can be approximated through a line in that region. On the other hand, high variance implies a volatility situation in which few observations are concentrated around zero, which is the point over the probability curve with the highest derivative, with Shapley values increasing in magnitude as the variable value deviates from the mean.

Shapley values are the most common type of importance-based explanation [90] and one of the most prominent approaches in the recent XAI literature. However, Kumar et al. [218] highlighted a Game Theory framework does not automatically solve the importance attribution problem, although it is an adequate general solution for quantifying the importance of variables.

K. EVALUATION METHODS AND METRICS

Despite the existence of a wide range of XAI methods, a proper evaluation of results from explanation methods still faces some difficulties. A problem in XAI evaluation is, in general, we do not have ''ground truth,'' i.e., the literature reports

no reliable references of what an adequate explanation for each black-box problem could be. As an example, machine-learning engineers still consider domain experts' judgments an implicit ground truth for explanations [90]. Therefore, the solutions proposed often rely on axioms for determining the desirable properties of explanations or use simulated data for checking and comparing what can be computed in terms of explanations [117]. Even synthetic datasets designed to hold ground truth explanations [229] suffer from a significant drawback due to the lack of guarantees learning-based models trained on those synthetic datasets will adhere to the underlying ground truth [175], [189]. Furthermore, defining ground truth is generally unavailable in most real-world applications that use explainability [230].

Relying on axioms to explain high-stakes Machine Learning may be insufficient. In this context, evaluation methodologies follow four main lines, namely, (*i*) measurement of the sensitivity of explanations to perturbations in the model and input data, (*ii*) inference about the behavior of explanations from feature removal, (*iii*) evaluation of explanations from controlled setups in which the importance of attributes is previously known, and (*iv*) visual assessment based on insights of human analysts (human-in-the-loop process) [110].

Qualitative and quantitative evaluations in XAI commonly correspond to the plausibility of explanations and fidelity to model behavior, respectively [114]. Bodria et al. [231] discussed quantitative tests and Yang and Kim [110] organized good research references regarding each of the four evaluation methodologies and proposed an evaluation framework based on perturbations. They also held a raw discussion on false explanations and highlighted evaluating explanations is as essential as developing XAI methods. The study also proposed two metrics for evaluating the extent to which and how the explanations should change according to modifications in the input attributes (for image data) in a controlled setting. Yang and Kim [110] did not guarantee an XAI technique that performs well in their framework would also perform well on real data and argued their metrics are simple tests for XAI techniques and those that fail in simple tests are likely to fail in more complicated scenarios.

Ablation is a line of XAI sensitivity assessment that attempts to evaluate global or local explanations by removing information from input features according to their importance ordering. In other words, an ablation study assesses the relative performance of a learning model perturbing its input features in a rank order of importance measured by the explanations [232], [233]. Theoretically, if an XAI method is adequately applied, the important features decrease the model performance when it is perturbed. Hooker et al. [224] presented ROAR (RemOve And Retrain) framework based on ablation to benchmark explanations from image processing models. It is costly, since it retrains the model for every perturbed input and the retraining processes diverge from the post-hoc paradigm. Furthermore, ablation assessment is dependent on baseline approximations. According to Haug et al. [234], the closer the approximation of a baseline to the original data distribution, the more discriminative the approximation, i.e., baselines that deviate from out-of-distribution (OOD) can produce invalid explanations.

Closely related to ablation, feature selection algorithms have been used for improving the assessment of the role of features in prediction and explanation tasks [129]. Some feature selection methodologies select a subset of important features from an original set, one-by-one or group-wise, according to their relative importance to each other and through feature elimination (top-down) [235] or inclusion (bottom-up) [129] approaches. Although feature selection reduces data dimensionality and computational complexity, it also depends on retraining processes.

Weerts et al. [236] applied qualitative analyses and Adadi and Berrada [4] extensively discussed the lack of evaluation in XAI, claiming the existence of few studies on XAI evaluation was due to the subjective aspect of explainability and highlighting rare studies were dedicated to the challenges of generating explanations truly understandable by humans. Researchers have also argued contrastive explanations can be applied in assessment contexts, for human-social interaction is based on contrastivity [22]. However, Hooker et al. [226] indicated contrastive and counterfactual explanations tend to be extrapolation problems (similarly to SHAP), which makes them potentially misleading. Furthermore, identifying optimal counterfactuals is an NP-hard task [237].

According to Yang and Kim [110], one reason for the lack of concrete works with human analysts is the human-in-the-loop evaluation process is complex and costly, since it involves sociological and psychological considerations. However, expert knowledge may enrich the context of explanations, making them more understandable [238].

In contrast to qualitative evaluations, quantitative ones are relatively independent of the model under explanation and almost exclusively devoted to the feature attribution context [114]. Amparore et al. [25] indicated a need for a consensus on fundamental metrics and the lack of definitions for quantifying the effectiveness of explanations. They demonstrated an unexpected behavior in LIME and SHAP and proposed four metrics to verify the different aspects of XAI techniques. Liu et al. [229] also showed some flaws in commonly used explainability methods and defined a set of metrics and a methodology for generating synthetic data for application in evaluations. Hooker et al. [224] presented a method for evaluating important variables and discussed the difficulty in directly "buying" results from XAI methods such as LIME and SHAP. Alvarez-Melis and Jaakkola [239] evaluated the drastic effects minimal perturbations can exert on XAI methods, claiming most XAI approaches are not sufficiently robust even when applied to explain robust learning models.

DeYoung et al. [240] presented a benchmark framework with metrics to measure the faithfulness and plausibility of explainability. *Faithfulness* refers to the extent to which an

explanation accurately represents the reasoning process of the model and *plausibility* evaluates the agreement of explanations with human-provided rationales [241].

Petsiuk et al. [242] introduced the *Prediction Gap on Important feature* metric (PGI), which measures the predictive faithfulness based on the change in a model's prediction probability from the selection of k features deemed as the most important ones and determined by a post-hoc XAI method. PGI is defined as

$$\text{PGI}(\mathbf{x}, f) = \frac{1}{m}\sum_{i=1}^{m}[|f(\mathbf{x}) - f(\tilde{\mathbf{x}})|] \tag{17}$$

where f represents a learning model, $\mathbf{x}$ is the original input data, and $\tilde{\mathbf{x}}$ is the same input, but holding the k most important features. The higher the PGI value, the more faithful the explanation. Barr et al. [233] compared metrics applied to measure the faithfulness of local explainers, showing most of the current metrics do not agree, which is a gap that should be investigated by the XAI community.

Stability metrics measure the sensitivity of explanations to changes under specific modifications of the model's hyperparameters or input data [123]. However, the XAI literature reports no agreement on stability (also known as sensitivity or robustness). Mishra et al. [128] discussed different definitions of stability metrics for feature importance and counterfactual explanation methods under the umbrella of robustness as a unified term.

Alvarez-Melis and Jaakkola [123] introduced *local Lipschitz continuity*, one of the first metrics for evaluations of the stability of local explanation methods. It generates a neighborhood $\mathcal{N}_{\mathbf{x}}$ of sampled instances $\mathbf{x}'$ by adding small perturbations around original input $\mathbf{x}$. $e_{\mathbf{x}}$ is expected to be similar to explanations $e_{\mathbf{x}'}$ because instances $\mathbf{x}'$ are sampled to be similar to $\mathbf{x}$. Agarwal et al. [243] improved local Lipschitz continuity by introducing the relative stability concept that evaluates the stability of an explanation considering perturbations in the inputs and output predicted probabilities of the underlying model. Therefore, *Relative Input Stability* (RIS) and *Relative Output Stability* (ROS) metrics are formulated according to

$$\text{RIS}(\mathbf{x}, \mathbf{x}', e_{\mathbf{x}}, e_{\mathbf{x}'}) = \max_{\mathbf{x}'} \frac{\|\frac{e_{\mathbf{x}} - e_{\mathbf{x}'}}{e_{\mathbf{x}}}\|_p}{\max(\|\frac{\mathbf{x}-\mathbf{x}'}{\mathbf{x}}\|_p, \epsilon_c)} \quad \text{and} \tag{18}$$

$$\text{ROS}(\mathbf{x}, \mathbf{x}', e_{\mathbf{x}}, e_{\mathbf{x}'}) = \max_{\mathbf{x}'} \frac{\|\frac{e_{\mathbf{x}} - e_{\mathbf{x}'}}{e_{\mathbf{x}}}\|_p}{\max(\|\frac{f(\mathbf{x}) - f(\mathbf{x}')}{f(\mathbf{x})}\|_p, \epsilon_c)}, \tag{19}$$

$\forall \mathbf{x}' \in \mathcal{N}_{\mathbf{x}}$ restricted to $\hat{\mathbf{y}}_{\mathbf{x}} = \hat{\mathbf{y}}_{\mathbf{x}'}$ (only perturbed instances predicted as the same predicted class of $\mathbf{x}$), where p is the l_p norm for measuring input/output changes and $\epsilon_c > 0$ is a clipping threshold for avoiding zero division. The larger the RIS/ROS values of the underlying explanation method, the more unstable the method to input/output perturbations.

Agarwal et al. [189] introduced OpenXAI, an open-source framework for evaluating and benchmarking post-hoc XAI methods. The tool includes a synthetic data generator, collection of real-world datasets, pre-trained models, feature attribution methods, and implementations of diverse quantitative metrics for evaluations of faithfulness, stability, and fairness of the included XAI methods. OpenXAI can benchmark new explanation proposals, despite a certain difficulty in customizing some of its parameters.

Quantus [244] is a comprehensive and well-documented Python library that provides several metrics from various evaluation categories, enabling comparative analyses of XAI methods and attributions.

L. XAI LIMITATIONS

What is a good explanation? Miller [22] defined a good explanation as one that is true in reality. Reality in machine learning is limited to the "truth" learned from training scope, which can hide unknown biases.

A way to promote trust is to increase the transparency of intelligent applications. An essential part of increasing transparency is the application of explainability [16], which can potentially unravel the black boxes of complex learning algorithms, enabling the introduction of more trust elements for supporting systems that actively use intelligent models. However, considerable discussions on the limitations of XAI and the aforementioned concerns on the lack of evaluation have been held. Although this study discussed several concepts, needs, challenges, and XAI methods, Kaur et al. [228] showed not all data scientists know how to apply XAI correctly in machine-learning pipelines.

Several investigations (including ours) have raised concerns over the need for more precise definitions of XAI basic terminology [4], [16]. Significant efforts were made in this review toward identifying the main concepts and specifying definitions as clearly as possible for each of the main elements of XAI theory in light of the literature. Although such definitions will contribute to clarifying many conceptual doubts in future research, the XAI area lacks agreement. As a result, every new paper addressing some XAI aspect or method has to introduce the same related terms in its own way, which is unnecessary and, consequently, leads to a profusion of similar (and confusing) terminology.

Krishna et al. [190] discussed the frequency at which explanations produced by state-of-the-art methods disagree with each other. The disagreement phenomenon may be tied to a lack of common objectives among explainability methods [245]. The authors also conducted a user study with data scientists on solutions to such disagreements in explanations. Han et al. [245] unified popular local feature importance methods into a same framework and demonstrated no one could generate optimal explanations across all data neighborhoods. Their results showed disagreement can occur because different explainers approximate the model using different neighborhoods and loss functions. The authors also established guidelines for choosing XAI methods according to faithfulness to the model.

Permutation techniques are among the leading XAI approaches and are easy to describe, develop, and use. They

show appealing results by producing ''null features,'' breaking the connection between features and target variables [17], [84]. Although permutation methods may be effective under a global null, they may fail to yield accurate explanations in cases that differ from the global null [246]. Hooker et al. [226] demonstrated how simple it is to generate examples in which permutation-based explanations can be misleading or distorted.

Explanation through causal relationships has rarely been explored in the XAI literature. Uncovering causality in Machine Learning is far from trivial, although regarded as a meaningful goal in explainability [90].

Non-numeric variables are another limitation that can be considered a challenge for XAI methods and properly handling categorical variables is also a challenge in learning algorithms. The solution traditionally used in this context is one-hot encoding, a simple method derived from digital circuits that transforms non-numeric features into binary matrices. However, in large datasets with many categorical attributes holding high numbers of distinct categories for each attribute, one-hot encoding significantly increases the degree of data sparsity which, in turn, increases data dimensionality, with most of the encoded values added as extra columns of little individual importance. An alternative for one-hot encoding is target encoding [247], which converts each value of a categorical attribute into its corresponding expected value. The resulting transformation does not add extra columns, avoiding turning the dataset into a sparser high-dimensional dataset. Aas et al. [117] suggested alternative approaches from clustering literature that describe distribution functions for manipulating non-numerical data [248] and generalizations of Mahalanobis distance for mixtures of nominal, ordinal, and continuous attributes [249].

Some studies have contested the real need for explanations [250], [251]. As discussed, explanations on learning models are not always required. However, after our argumentation, a clear understanding of the problems associated with relying exclusively on black-box models based on good metric performance is expected, since auditability is required for our understanding of the reasons behind predictions of complex models. In this study, we consider the questions about the ''human-centric'' lack of significance in explainability and the impacts on comprehension's mental models of users very pertinent – both as a consequence of the lack of explanations completeness and the information overload generated by XAI methods [218].

However, black-box model unboxing should be integrated into machine-learning development pipelines. Although significant, the present limitations will mature with research evolution in the XAI domain and more careful development of future tools. Such criticisms must be raised to addressing them and enabling the evolution of XAI.

M. SUMMARY AND CHARACTERIZATION OF XAI METHODS

According to Bhatt et al. [90], feature importance is the most widely used class of explainability technique. While traditionally characterized in different categories, many ''explain prediction'' methods such as LIME, a feature importance/attribution method often classified as a surrogate or simplification method, provide explanations through feature importance tasks. Similarly, LRP is an attribution method based on gradients and signal propagation framework.

Data scientists use feature importance for a variety of tasks (e.g., verifying the existence of biases in datasets or deficiencies in models, comprehending the underlying learning process, and providing insights for further feature engineering). Note feature importance methods can generate information on the importance of a particular feature to a prediction made by a specific trained model. However, they do not disclose that feature's generalizability or possible importance to any other learning models, although such a lack of generalizability between explanations in different models might be extended to other XAI approaches.

Table 1 shows an overview of the literature reviewed in this research, according to XAI approach and explanation objectives (note some methods can be classified into more than one category). We highlighted the main strengths and limitations commonly related to each approach.

Table 2 summarizes the objectives of the explanation tasks shown in Table 1. For a more detailed discussion on XAI categorizations, see Section VII-B.

IX. PRACTICAL EXAMPLES OF SUCCESSFUL XAI APPLICATIONS

This section provides a high-level discussion of examples from various domains where XAI has been successfully applied. According to Google Scholar,[3] we used as a search engine, SHAP [109] and Grad-CAM were cited in more than 20 thousand publications each, LIME [97] was cited in more than 16 thousand ones, Vanilla Gradient [186] counted nearly eight thousand citations, Integrated Gradients [185] showed around six thousand citations, and LRP [192] and DeepLIFT [203] roughly received four thousand citations each. The numbers show the impact of those techniques, since they are recent publications. Proceeding with a comprehensive analysis of every XAI application would be an overwhelming endeavor, despite the removal of possible duplicates and studies with no in-depth discussions on XAI; therefore, we selected impacting publications from high-level venues covering different areas toward an overview of how the elements of XAI translate into practical applications.

XAI has been widely addressed in **medicine and healthcare** research. Caruana et al. [20] used a high-performance GAM model to predict the risk of 30-day medical readmissions related to pneumonia cases. They presented detailed case studies with impressive results, leading to the discovery of patterns that previously prevented complex models from being deployed for medical applications. Lundberg et al. [12] applied SHAP to ensemble learning-based models in a medical context and predicted complications during surgical

[3]https://scholar.google.com/, visited on March 2024

TABLE 1. Characterization of Explainable Artificial Intelligence research and approaches.

Approach	Explanation Task	References	Strengths	Limitations
Approximation or distillation	Explain the model or predictions	[20], [33], [112] [133], [135], [136], [210]	Most proposals are agnostic. Can provide local or global explanations. Reduce the computational overhead.	Suitable only for limited contexts. Using a transparent model does not ensure interpretability in practice.
Visualization	Explain the model Explain predictions Model Inspection	[9], [98], [141], [148] [106], [118], [120], [144], [149], [194], [222] [71], [82], [83], [134], [142] [143], [146]	Can provide graphical representations that facilitate pattern discovery. Enable comparisons of different models or data.	Most proposals are model-specific. Information overload is present when high-dimensional feature spaces are visualized. Scalability limitations may result in visual clutter or loss of detail, limiting the approach's effectiveness.
Decision boundaries	Explain the model	[99], [101], [151], [152], [153], [154]	Enable understanding how models classify different regions of the feature space.	Most proposals are model-specific. Highly non-linear decision boundaries are difficult to characterize.
Contrastive and counterfactual	Based on examples	[155], [156], [158], [159], [160], [161], [162], [163], [164]	Valuable insights into factual relationships through human-like causal reasoning.	Can be computationally expensive. Counterfactual perturbations may be subjective and highly context-dependent and may not be realistic because of model extrapolations.
Graph Machine Learning models	Explain the model or predictions	[165], [166], [167], [168], [169], [171], [172], [173], [174]	Can provide local or global explanations. Connections among vertices can explain individual predictions.	Large-scale graphs can be very complex to explain. Scalability limitations.
Attention models	Explain the model	[137], [178], [179], [180], [181], [182], [183], [184]	Can provide local or global explanations. Exploring attention weights can reveal the most relevant elements to the prediction.	Attention mechanisms can be very complex. Limited regarding model types or architectures.
Gradient-based	Explain predictions	[106], [185], [186], [188], [192], [193], [195], [196], [197], [198], [199], [200], [201], [203]	Can provide local or global explanations. Suitable for complex domains, e.g., image classification.	Suitable only for models with differentiable parameters. Computationally expensive for large models and high-dimensional input data.
Simplification	Explain predictions	[97], [133], [194], [207], [208], [209], [210], [213], [214], [215], [216]	Most proposals are agnostic. Flexible to different simplified models. Suitable for different data.	Oversimplification of decision boundaries may introduce approximation errors. Tend to instability because of sampling procedures.
Shapley-based	Explain predictions	[12], [14], [84], [86] [109], [113], [114], [117], [119], [194], [205], [218], [219], [220], [221], [222], [225], [226], [228], [236]	Most proposals are agnostic. Shapley values are endowed with desirable axiomatic properties. They can provide local or global explanations.	Exact solution is NP-hard. Approximated solutions can introduce errors and are prone to instability. Shapley values can lead to misleading explanations because of model extrapolations. Complex interactions can be ignored by assuming feature independence.
Evaluation	Assessment and validation	[25], [110], [123], [129], [175], [224], [229], [232], [233], [236], [239], [240], [242], [243], [244]	Essential process to ensure both faithfulness and robustness of any XAI method.	Lack of comprehensive methodologies. Lack of ground truth. Qualitative approaches can be expensive and prone to objectiveness.

procedures. Medical specialists tested the methodology by using a graphical web interface, returning positive feedback.

Meske and Bunde [252] employed models to detect malaria in cell images and then used LIME to explain which part of the cell caused the model to make its prediction. The impact of research of such nature enhances transparency and trust in automated disease diagnostics. Zhang et al. [253] discussed interesting approaches to modeling the understanding of clinical texts using Transformers. Properly processing the semantic information contained in clinical notes provided

TABLE 2. XAI explanation tasks and their objectives.

Explanation Task	Description
Explain the model	Describes the general behavior of a learning model, clarifying how it processes data and makes decisions. The goal is to support the model's trust globally. In addition, it can provide an understanding of how a learning model works internally, visualizing its structure and internal mechanisms.
Explain predictions	Explains why a model made a particular prediction for a given input. The goal is to support users' trust in the model's predictions. It usually provides an understanding of how learning models behave locally in the neighborhood of an input.
Model Inspection	Verifies the model's behavior and performance, often utilizing sensitivity analysis through data perturbations. The goal is to evaluate whether the model behaves as expected and remains robust and reliable under different conditions.
Based on examples	Provides factual reasoning for the model's predictions. The goal is to explain why a particular prediction was made instead of another or how the prediction could change under different conditions.
Assessment and validation	Evaluates the effectiveness of XAI methods in providing interpretable, consistent, accurate, reliable, and relevant explanations. It includes quantitative metrics, qualitative evaluations by domain experts, and comparisons using ground truth or human-based knowledge.

by doctors could automate several applications related to medicine and healthcare. The authors proposed using the visualization tool provided by Vig [179] for extracting existing relationships learned by the Transformer (e.g., symptoms and body parts).

Lawhern et al. [254] developed a Convolutional Neural Network (CNN) to classify electroencephalogram signals using DeepLIFT to provide feature importance and support the predictions' confidence. According to the authors, DeepLIFT results suggested the network had learned relevant features closely aligned with results from the literature. Qiu et al. [255] proposed a learning framework to classify clinical information from individuals into different cognitive levels for supporting neurologists in detecting Alzheimer's. DeepLIFT was then applied to assess the contribution of imaging and non-imaging features to the diagnoses.

The **COVID-19 pandemic** posed a critical and urgent threat to global health, thus motivating remarkable research efforts in medicine and other areas, including Machine Learning. Zoabi et al. [256] trained a model to predict positive SARS-CoV-2 infections and applied beeswarm and summary plots from SHAP to identify features impacting the model's predictions. [257] also used SHAP to identify important features of long COVID cases from electronic records in the USA.

Hu et al. [204] designed a deep-learning model to classify COVID-19 infections from computed tomography images, relying on Integrated Gradients to provide lung lesion localization. Using chest X-rays, Brunese et al. [258] developed a deep-learning model to detect COVID-19. Grad-CAM was applied for inputs classified as positive for highlighting the symptomatic areas related to the disease, thus reducing the diagnosis time. A similar setup for explaining COVID-19 detection from X-rays was proposed in [259]. Oh et al. [260] developed a patch-based CNN architecture where the COVID-19 class can hold patches with different scores. Then, Grad-CAM was patch-wise adapted by weighting its saliency maps with the classes' probabilities.

Characterizing elements that influence **cancer** is a biological and clinical challenge [261]. Chen et al. [262] presented a framework based on different Neural Networks for cancer diagnosis and prognosis. The system combines morphological and molecular information from histology imaging and genomic features, respectively, and enables interpretability by applying Grad-CAM and Integrated Gradients. Elmarakeby et al. [261] introduced a deep-learning model trained to predict cancer stages in patients diagnosed with prostate cancer based on molecular data from their genomic profiles. DeepLIFT evaluated the importance of specific genes in the model's prediction and attributed high scores to genes known as prostate cancer previously related to metastatic disease drivers, thus inspiring new hypotheses for cancer studies.

Genetics is a research- and technology-intensive area that has demanded more machine-learning solutions for making predictions in genomics research. According to Novakovsky et al. [49], explanatory elements promoting insights into genetic processes can be more significant than the predictions themselves for genomics researchers. CRISPR-Cas9 is a cutting-edge tool in genetic engineering that enables a wide range of genome editing in different organisms. Wang et al. [263] used TreeSHAP and DeepSHAP to evaluate the influence of the position-dependent nucleotide features and improve a Recurrent Neural Network (RNN) model to predict gene activity for CRISPR-Cas9 design.

Bar et al. [264] trained Gradient-boosted Decision Trees with the human serum metabolome to predict metabolite levels and measure biomarker agents of different diseases. They used TreeSHAP to find associations between representative genomes and metabolite levels and discovered diet and

microbiome increased the models' predictive power. The effect of feature values on predictions was modeled following a directional mean absolute importance values setup, which relates the SHAP values from TreeSHAP with the sign of a Spearman correlation between target and features.

Avsec et al. [265] constructed a deep learning architecture based on convolutional layers and self-attention mechanisms to predict gene sequences with high expression levels from large-scale sequencing data in evolutionary studies. They computed gradient scores [203] to assess the contributions of different gene expressions and understand the gene sequences impacting the model predictions. Buergel et al. [266] trained a Deep Neural Network (DNN) as a metabolomic model using a large dataset with 168 metabolic markers as input for learning disease-specific states and predicting multi-disease risk for 24 conditions, including neurological disorders and cancer. DeepSHAP was initially applied globally to detect the metabolites that most affected disease risk, considering all investigated diseases. Global SHAP values were visualized through heat maps and circular charts. Next, they applied DeepSHAP locally to attribute risk profiles for individual predictions, visually analyzing them by projecting the entire set of SHAP values with UMAP [150].

Chklovski et al. [267] used TreeSHAP to highlight the contribution of specific genomic features and pathways to predictions of a model trained on metagenomic data. For further discussions, Novakovsky et al. [49] conducted a recent literature review focusing on sequence-to-activity models and the emerging applications of XAI used in genetics research to investigate spurious correlations and complex interaction relationships between features.

In **language modeling**, hate speech is a cultural threat resulting from the increase in online iterations. Despite the proposals of hate speech detection models, more research on their interpretability must be developd. Mathew et al. [241] used LIME and attention methods to detect significant tokens related to hate speech. The authors verified models with high performance in hate speech classification do not perform well on explainability metrics such as faithfulness and plausibility [240] and introduced a benchmark dataset covering multiple aspects of hate speech detection.

Pratt et al. [268] combined open-vocabulary models with large language models (LLMs) to generate sentences describing the characteristics of images. They computed Shapley values to understand the importance of different image regions highlighted in the descriptions and used heat maps to visualize the results. Sarzynska-Wawer et al. [269] used LIME to understand the reasons behind predictions of a language model applied in psychiatry for diagnosing schizophrenia-related symptoms. The authors focused the explainability task on misclassified patients, with LIME results identifying types of words as indicative of thought disorder and revealing their language model was sensitive to context and word meanings. For a comprehensive overview of explainability within the language processing domain, we refer to [55].

Regarding **industrial applications**, Hong et al. [270] used force and decision plots from SHAP to analyze the results of a model based on deep CNN and LSTM networks applied to sensor data for predicting turbofan-type aircraft engine maintenance. Brito et al. [271] proposed an unsupervised approach for detecting and classifying faults in rotating machinery and applied SHAP for feature rankings in anomaly diagnosis. Black-box models have been used in the automotive industry for enabling vehicles to perceive the environment and make driving decisions with less or no human intervention. In this context, transparency is critical for accepting autonomous vehicles on commercial scales. Omeiza et al. [51] surveyed the XAI applications and challenges of autonomous driving and Zablocki et al. [52] focused their study on XAI methods for vision-based self-driving deep learning models. We refer to [53] for further details on the applications of XAI methods in modern industries.

XAI has also been applied in **time series** analysis. Xu et al. [272] used SHAP in an interactive system for analyzing the relationship between input feature importance and the output of multidimensional time-series forecasting models. Parsa et al. [273] modeled real-time data from Chicago metropolitan expressways to detect the occurrence of traffic accidents and applied SHAP and dependency plots to analyze the impact of input features (e.g., speed) for accident detection.

In the **chemistry** domain, Sanchez-Lengeling et al. [274] employed graph learning explainability to tackle the quantitative structure-odor relationship (QSOR) problem. The challenge was to understand which molecule's substructure was responsible for the specific scent of that material (e.g., fruity, weedy, medicinal). Preuer et al. [275] trained a Deep Graph CNN model to classify drugs into toxic and non-toxic; they derived explanations from Integrated Gradients to detect the molecular substructure causing the prediction, arguing chemists can design methods to modify the responsible elements and thus avoid molecule toxicity.

McCloskey et al. [276] used a Message-Passing Graph Neural Network to study the binding properties between molecules and proteins and then relied on Integrated Gradients to explain which parts of the complex molecule-protein scheme were causing the chemical bond to happen. The explanations enabled them to discover spurious binding correlations in predictions despite that network achieving perfect classification accuracy. Schwaller et al. [277] applied Transformers to learn chemical reaction mechanisms based on the grammar of organic chemical interactions. The model input is the stream of tokens concerning atoms in the chemical chain of the molecules involved in the reaction. They explained the complex atom mapping between reactants and products by visualizing the relationship learned by the Transformer attention heads using the tool provided by Vig [179].

Yang et al. [278] trained learning models on chemical descriptors to predict gas permeability and design

high-performance polymer membranes. SHAP extracted the contributions of the different chemical components linked to permeability and selectivity and, according to the authors, it identified impacting molecular substructures, thus encouraging future chemical and polymer research toward taking advantage of explainability. Jiang et al. [279] applied SHAP to explore the importance of molecular descriptors in drug discovery tasks. Jiménez-Luna et al. [280] addressed the technical challenges that XAI approaches faced when applied in drug discovery supported by machine learning, highlighting the need for a collaborative effort among deep-learning developers, cheminformatics experts, chemists, and other domain specialists for promoting models' reliability with XAI in the chemistry area.

Artificial Intelligence can also assist **mathematicians** in discovering new theorems and proposing solutions for long-standing open deductions. Davies et al. [281] proposed an intriguing learning-based framework to recognize potential patterns and relationships in pure mathematics problems. When the model finds some relationship, it applies Integrated Gradients to explain its nature, thus enabling mathematicians' intuition in proposing new conjectures.

A considerable number of recent papers from diverse research areas have tied XAI methods to their machine-learning studies. Most of them only cited explainability as a future direction to face the transparency challenges of using potent black boxes. However, several studies published in high-impact venues improved their results with the support of XAI approaches, as demonstrated in this section. We highlight the prevalence of gradient-based methods in applications handling image data, since most of those methods are designed for architectures based on Neural Networks, which are almost standard in image modeling applications. Moreover, multiple studies using XAI can be found in medicine and genetics, two areas with intensive research efforts that have been increasingly open to machine-learning solutions, thus proving the practical value XAI can add to cutting-edge research.

Table 3 summarizes the XAI applications presented in this section organized according to the research area in which they were employed.

TABLE 3. Summary of the XAI applications presented in Section IX, according to their research domains.

Application domain	References
Medicine and healthcare	[12], [20], [252], [254], [255], [256], [257], [261], [262]
Genetics	[49], [263], [264], [265], [266], [267]
Images	[204], [258], [259], [260]
Language modeling	[55], [241], [268], [269]
Industrial	[51], [53], [52], [270], [271]
Time series	[272], [273]
Chemistry	[274], [275], [276], [277], [278], [279], [280]
Mathematics	[281]

A. PACKAGES AND LIBRARIES FOR XAI APPLICATION

Despite many papers promising explainability for tackling the opacity of black-box models, XAI developers have made significant efforts to implement explainability methods, making them available through software packages and libraries. LIME [97] and SHAP [109] are openly available libraries providing methods and tools related to their base approaches (see Sections VIII-H and VIII-I). Other initiatives include explainers covering different XAI approaches.

IML [282] was one of the first XAI packages. It is an **R** toolkit focused on classical implementations of some well-known global and local model-agnostic methods. InterpretML [283] is a Python package that provides interpretability through a set of transparent models and post-hoc explainers, as well as visualization tools for feature importance analysis. Captum [284] is a PyTorch library focused on gradient and perturbation-based attribution methods for explaining Neural Networks and also provides a visualization tool and sensitivity-based evaluation metrics. Similarly, the iNNvestigate library [285] provides several gradient and LRP-based methods for explaining Neural Network architectures. AIX360 [286], [287] includes eight local and global explainers, evaluation metrics, and demonstration tutorials. Alibi Explain [288] is a Python library that implements transparent and black-box models and nine explanation methods, including counterfactuals and bias detection, for generating local and global explanations.

OmniXAI [289] is a comprehensive library that provides explanations through a wide range of specific and model-agnostic XAI methods and visualization tools for various types of models (e.g., Scikit-learn, PyTorch, and TensorFlow implementations) and data. In addition to those valuable XAI frameworks, the previously discussed OpenXAI [189] and Quantus [244] packages are devoted to providing multiple evaluation metrics for XAI methods and explanations.

Bhatt et al. [90] outlined how several organizations have applied XAI strategies to their workflows, highlighting which explainability methods work best in practice. According to the authors, local explainability is typically the most relevant form of model transparency for end users. However, they concluded most XAI advances are far from end users due to the limitations of current approaches for generating direct information to those users, with machine learning engineers being the primary users of XAI implementations most for sanity-checking procedures during development processes. Despite the diversity of successful application cases, there are significant opportunities for improvements in future XAI research, as discussed in the next section.

X. DISCUSSIONS AND FUTURE OPPORTUNITIES

In this study, we conducted qualitative comparisons, providing a comprehensive view of the strengths and limitations

of XAI approaches. Interested readers can find extensive studies on quantitative comparisons among XAI methods in Amparore et al. [25], Krishna et al. [190], and Tan et al. [33]. Regardless of the considerable efforts devoted to this study, examining all relevant elements, techniques, and publications involved in the XAI environment would be unfeasible. Therefore, other surveys and reviews have been indicated for clarifying some specific concepts beyond the scope of this text.

As addressed elsewhere, the number of decisions made with the support of intelligent systems has grown and new proposals have continuously emerged and been adopted. Multiple domains in which learning algorithms are inserted can potentially influence the way society interacts, challenging new tools to mitigating possible negative consequences. Although the scientific community devoted to Machine Learning has successfully improved the predictive performance of models, the trade-off between precision and transparency still must be adjusted. High precision indicates high rates of true-positive predictions, hence, low rates of incorrect decisions. However, ignoring the logical processes that generate decisions is unacceptable [12].

According to Lundberg and Lee [109], the best explanation for a simple model is the model itself, for it represents the learned space perfectly and is easy to understand. However, it is impossible to use a trained model based on complex architectures, such as Random Forests and DNNs, to explain itself due to difficulties in interpreting their decision structures. Non-linear learning models create mapping spaces that ''carve'' space segments for different data classes, and such regions can be fully connected, which is challenging to unravel [101]. Interpretability arises from the model design. However, when a model cannot be directly interpreted, explainability methods must be applied to transforming obscure elements into interpretable information. Multiple aspects that can affect the ability of XAI tools to disclose the logic of complex models must be considered.

We remind the reader not all learning-based systems require interpretability. Further explanation is not necessary in case of no significant consequences for an algorithm results or if the problem has already been sufficiently tested in real situations [21]. However, since complex learning models have been increasingly adopted for making critical decisions in critical contexts, such as precision medicine, algorithmic transparency has been clearly demanded due to the reluctance of humans to employ techniques that are not interpretable, treatable, or reliable, thus limiting the applicability of Machine Learning [290]. Unjustifiable decisions, or decisions that do not enable detailed explanations about the underlying logic, can lead to dangerous situations or even profound impacts on social dynamics [2], [291], [292].

The ''right to information,'' defined by the European GDPR or the Californian CCPA (see Section I-A), constrains personal data usage and demonstrates the inherent lack of ethics many people feel regarding the decision-making process involving automated systems and humans when no reasonable explanations are available [218]. Therefore, research and development of comprehensive XAI techniques are essential for understanding the decisions made by Machine Learning applications, offering data scientists and users the ability to summarize views on information hidden by the complex and intricate parametric spaces of modern learning models.

Bhatt et al. [90] detected a gap between explainability and the goals of transparency, since most of the current XAI deployments primarily serve developers, who internally use explainability to debug models, rather than external end users affected by Machine Learning, who are intuitive consumers of explanations. Prediction explanation is currently one of the leading research lines in XAI. Explaining a prediction is particularly interesting because the model discovers patterns that can be even more meaningful than the predictive performance [14]. However, the performance metrics typically used in Machine Learning cannot appropriately verify learned patterns. Such a lack of assurance introduces a sensitive issue regarding the broad, sometimes thoughtless, use of complex non-linear models for decision-making in domains ranging from science to industry [106].

Lundberg et al. [12] argued only providing information about which variables are more important for the learning model does not imply uncovering causal relationships, since importance values do not represent a complete scenario for explaining a learning model. In other words, explaining a prediction simply by providing unique aspects within today's sophisticated learning systems is only part of the promotion of interpretability. Nevertheless, understanding the features that influence a decision is information that analysts can use to formulate explanations of the right reasons or biases that guide a model's predictions due to inferring the logic behind complex black boxes with no support of an XAI method is challenging.

Instability is a critical matter in XAI because it reflects one of the most damaging factors related to the integrity of an explanation provider, namely, difficulty in promoting trust. As an example, if an application supported by decision algorithms provides users with inconsistent explanations for similar (or same) data instances, those explanations cannot be considered reliable [25]. State-of-the-art XAI techniques suffer from instability due to the randomness of the approximation strategies used to circumvent the need for computational resources. Another instability factor is the extrapolation behavior derived from impossible or improbable scenarios that can occur during the inclusion of correlated features in the marginalization strategy. Moreover, some XAI methods are based on heuristic or empirical solutions, with weak justifications for the choice of important parameters.

Most XAI methods are non-robust and their adoption for understanding models devoted to safety-critical applications can be risky [128]. Rudin [293] argued interpretable models are more appropriate for applications in which high-stakes decision-making is more important than attempts to explaining

the outputs of black-box models through limited XAI methods that do not offer all theoretical guarantees. However, due to performance and computational restrictions, transparent models and exact axiomatic XAI solutions cannot be applied in several real-world circumstances. In some machine learning tasks, it is not reasonable to sacrifice performance by using transparent solutions, thus demanding complex high-performance models. In such cases, XAI approaches represent a promising direction for debugging or identifying insights that warrant deeper investigation, enabling users to build explanatory elements [114].

A comprehensive XAI tool capable of producing consistent, stable, and comprehensive information about different aspects of a learning procedure can provide a helpful explainability scenario. If carefully developed and applied, XAI adds a new perspective to the vast horizon of Machine Learning, which can enrich future debates on whether computer devices can genuinely exhibit intelligent behavior [106]. There is a long avenue of research in XAI related to developing stable and consistent explainability solutions based on evaluated and validated explanations for accomplishing such aims. A modern XAI solution would include more than only prediction explanations or model inspection; it should be a comprehensible tool that includes explanations from different XAI approaches enhanced by information visualization and human interaction mechanisms – in this sense, we cite the impressive proposal of Lundberg et al. [12].

However, XAI researchers and applicants face a significant open question that hampers the development of comprehensive XAI tools. The current regulations have established the right to explanations, but only in high-level conceptual terms. No law has defined system requirements for XAI, baselines, guidelines, assessment or validation standards, or presentation formats for explanations, leading to a lack of common objectives and formalization in current XAI approaches. XAI community has established all such needs by questioning machine learning opacity.

Technology evolves rapidly and the needs of each target audience can also change quickly. Moreover, the target audiences of XAI tools can be very heterogeneous. In this sense, we recommend designing processes based on multidisciplinary teams with data scientists, psychologists, domain experts, and law specialists as an important future direction for XAI evolution toward the creation of XAI tools tailored to the target audience requiring explanations and observing privacy, data governance, accountability, and fairness principles. For a valuable discussion connecting the research gaps between XAI and fairness, we refer to [294].

Providing users with realistic and highly informative explanations is a desirable way to satisfy criteria such as robustness and faithfulness. However, more informative explanations might leak non-trivial information about the underlying model, which could be explored in model extraction attacks and raise privacy issues [161]. Protecting data privacy while enhancing transparency is of paramount research importance in XAI. The literature does not fully explore the extent to which model explanations can unintentionally reveal sensitive details about users' data. In this context, Aïvodji et al. [161] addressed the trade-off between privacy and explainability and Nguyen et al. [125] surveyed the recent findings in privacy-preserving mechanisms within model explanations to avoid privacy leaks or deciphering attacks by malicious entities. Additionally, we restate the need for more careful evaluations. Future research on XAI must carefully create and systematically apply qualitative and quantitative assessment methodologies and metrics for robust explanations.

Explainability goes beyond the simple belief in raw statistical measurements within the Machine Learning environment, thus benefiting everyone involved (and affected) through applications relying on artificial intelligence-supported decision-making. Machine Learning is revolutionary and can be applied as a powerful element to shift paradigms, leading to future changes in society. However, not everything related to technological advances derived from Artificial Intelligence is perfect, as discussed by several studies addressing the weaknesses of high-end learning algorithms referenced in this review. Some of those drawbacks go beyond the problems related to challenges in understanding such intricate models, and, as discussed here, the ability of Machine Learning to adapt and generalize may sometimes be overestimated.

The reality appears much more evident in hindsight than in foresight [295]. Therefore, new learning algorithms and methodologies that are more robust and meet ethical and legal requirements can emerge, taking advantage of the support of XAI techniques. Users who choose to merely believe in black-box model outputs without contesting the motivation behind predictions may eventually be fooled by randomness and outcome bias [296].

XI. CONCLUSION

Artificial Intelligence is not the future; it is the present. Modern high-performance learning-based applications have become a near-ubiquitous reality; however, several institutions still use classical models (e.g., linear regression and rule-based ones) due to the need for more transparent solutions [293]. An explainable artificial intelligence application provides detailed elements clarifying the models' decision-making procedures, thus facilitating the understanding of the contributions of features to predictions and their impact on predictive performance, making such models' rational processes more transparent and verifiable [2]. Black boxes that only output decisions with no further explanations of the underlying mechanisms are difficult to trust and do not supply contextual directions to support their outcomes when confronted by users [12].

In this sense, knowing what goes in and out of automated decision-making systems with no comprehension of what occurred behind the scenes no longer satisfies the information access needs, which is against recent ethical and legal directives. XAI can generate human-comprehensible explanations for machine decisions, helping detect hidden biases. Introducing such clarifications might lead to the

TABLE 4. List of abbreviations and acronyms.

AI	Artificial Intelligence
CCPA	California Consumer Privacy Act
CNN	Convolutional Neural Network
Contfactuals	Contrastive and/or Counterfactual explanations
DeConvNet	DeConvolutional Network
DeepDIG	Deep Decision boundary Instance Generation
DeepLIFT	Deep Learning Important FeaTures
DeepSHAP	Deep Learning SHAP
DNN	Deep Neural Network
DTD	Deep Taylor Decomposition
FDA	U.S. Food and Drug Administration
G-DeepSHAP	Generalized DeepSHAP
GALE	Globally Assessing Local Explanations
GAM	Generalized Additive Model
GDPR	General Data Protection Regulation
GML	Graph Machine Learning
GNN	Graph Neural Network
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
KernelSHAP	Kernel-based SHAP
LIME	Local Interpretable Model-Agnostic Explanations
LLM	Large Language Model
LRP	Layer-wise Relevance Propagation
LSTM	Long Short-Term Memory Network
MAPOCAM	Model-Agnostic Pareto-Optimal Counterfactual Antecedent Mining
MCTS	Monte Carlo Tree Search
ML	Machine Learning
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
OOD	Out-of-Distribution
PGI	Prediction Gap on Important feature
QSOR	Quantitative Structure-Odor Relationship
ReLU	Rectified Linear Unit
RIS	Relative Input Stability
RNN	Recurrent Neural Network
ROAR	RemOve And Retrain
ROS	Relative Output Stability
SHAP	SHapley Additive exPlanations
SVM	Support-Vector Machine
t-SNE	t-Distributed Stochastic Neighbor Embedding
TDA	Topological Data Analysis
TreeSHAP	Tree-based SHAP
UMAP	Uniform Manifold Approximation and Projection
URL	Uniform Resource Locator
WHO	World Health Organization
XAI	EXplainable Artificial Intelligence

development of procedures or algorithms to identify and handle biases, which is valuable to organizations potentially responsible for unfair decisions.

This concise literature review showed explainability for machine learning is a fruitful area, with multiple strategies already in use and others under development toward filling the current gaps in bringing transparency to opaque systems. We held discussions on the current XAI literature, contributing from theoretical characterizations to critical investigations of XAI methods and applications. XAI explanations reach domains where transparent models cannot be applied. However, a black-box explainer must also be explainable and robust. Despite the multiple XAI approaches proposed, the literature reports no gold-standard XAI system and, as we highlighted, design considerations and proper assessments are critical concerns to be more methodologically addressed in future XAI research. Properly validated explanations can improve the confidence levels of artificial intelligence-powered applications, promoting sustainable computer decisions according to social responsibility, reliability, and security needs.

APPENDIX
ABBREVIATIONS AND ACRONYMS

The abbreviations and acronyms used in this review are summarized in Table 4.

ACKNOWLEDGMENT

The authors acknowledge Angela C. P. Giampedro for her valuable comments. Figures 3 and 6 were created with BioRender.com. The opinions, hypotheses, and conclusions or recommendations expressed in this material are the authors' responsibility and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2010.

[2] 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.

[3] D. M. West, *The Future of Work: Robots, AI, and Automation*. Washington, DC, USA: Brookings Institution Press, 2018.

[4] 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.

[5] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019.

[6] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7499–7519, Jun. 2024.

[7] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, Opatija, Croatia, May 2018, pp. 0210–0215.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 7553.

[9] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, Jan. 2018.

[10] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models: A survey of results," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, San Francisco, CA, USA, Aug. 2017, pp. 1–6.

[11] S. Ullman, "Using neuroscience to develop artificial intelligence," *Science*, vol. 363, no. 6428, pp. 692–693, Feb. 2019.

[12] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, and S.-I. Lee, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 749–760, Oct. 2018.

[13] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.

[14] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020.

[15] O. Maier and H. Handels, "Predicting stroke lesion and clinical outcome with random forests," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Athens, Greece: Springer, 2016, pp. 219–230.

[16] J. Amann, D. Vetter, S. N. Blomberg, H. C. Christensen, M. Coffee, S. Gerke, T. K. Gilbert, T. Hagendorff, S. Holm, M. Livne, A. Spezzatti, I. Strümke, R. V. Zicari, and V. I. Madai, "To explain or not to explain—Artificial intelligence explainability in clinical decision support systems," *PLOS Digit. Health*, vol. 1, no. 2, Feb. 2022, Art. no. e0000016.

[17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[18] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data Soc.*, vol. 3, no. 1, pp. 1–12, 2016.

[19] W. Samek, T. Wiegand, and K. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," 2017, *arXiv:1708.08296*.

[20] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 1721–1730.

[21] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[22] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.

[23] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019.

[24] EU Regulation. (Apr. 2023). *2016/679 of the European Parliament and of the council of 27 April 2016 on the General Data Protection Regulation*. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj

[25] E. Amparore, A. Perotti, and P. Bajardi, "To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods," *PeerJ Comput. Sci.*, vol. 7, p. e479, Apr. 2021.

[26] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surveys*, vol. 51, no. 5, pp. 1–42, Sep. 2019.

[27] FDA. (Feb. 2024). *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)—Discussion Paper and Request for Feedback*. [Online]. Available: https://www.regulations.gov/document?D=FDA-2019-N-1185-0001

[28] L. Maier-Hein, M. Eisenmann, D. Sarikaya, K. März, T. Collins, A. Malpani, J. Fallert, H. Feussner, S. Giannarou, and P. Mascagni, "Surgical data science—From concepts toward clinical translation," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102306.

[29] WHO. (2021). *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. World Health Organization, Geneva, Switzerland. Accessed: Feb. 2024. [Online]. Available: https://www.who.int/publications/i/item/9789240029200

[30] State of California. (2021). *California Consumer Privacy Act (CCPA)*. State of California, Department of Justice, Office of the Attorney General. Accessed: Jul. 2023. [Online]. Available: https://oag.ca.gov/privacy/ccpa

[31] 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.

[32] 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.

[33] S. Tan, G. Hooker, P. Koch, A. Gordo, and R. Caruana, "Considerations when learning additive explanations for black-box models," *Mach. Learn.*, vol. 112, no. 9, pp. 3333–3359, Sep. 2023.

[34] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Frontiers Big Data*, vol. 4, p. 39, Jul. 2021.

[35] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, Jan. 2021.

[36] P. Hase and M. Bansal, "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?" in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5540–5552.

[37] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 11, nos. 3–4, pp. 1–45, Dec. 2021.

[38] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, vol. 76, pp. 89–106, Dec. 2021.

[39] F. Vitali, "A survey on methods and metrics for the assessment of explainability under the proposed AI act," in *Legal Knowledge and Information Systems*, vol. 346. Amsterdam, The Netherlands: IOS Press, 2022, p. 235.

[40] C. Molnar, *Interpretable Machine Learning*. Durham, NC, USA: Lulu Press, 2019.

[41] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proc. ACM Conf. Fairness, Accountability, Transparency*. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 2239–2250.

[42] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[43] J. Petch, S. Di, and W. Nelson, "Opening the black box: The promise and limitations of explainable machine learning in cardiology," *Can. J. Cardiol.*, vol. 38, no. 2, pp. 204–213, Feb. 2022.

[44] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," *Diagnostics*, vol. 12, no. 2, p. 237, Jan. 2022.

[45] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102470.

[46] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. V. Tengg-Kobligk, R. M. Summers, and R. Wiest, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiol., Artif. Intell.*, vol. 2, no. 3, May 2020, Art. no. e190043.

[47] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Inf. Fusion*, vol. 77, pp. 29–52, Jan. 2022.

[48] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Comput. Methods Programs Biomed.*, vol. 226, Nov. 2022, Art. no. 107161.

[49] G. Novakovsky, N. Dexter, M. W. Libbrecht, W. W. Wasserman, and S. Mostafavi, "Obtaining genetics insights from deep learning via explainable artificial intelligence," *Nature Rev. Genet.*, vol. 24, no. 2, pp. 125–137, Feb. 2023.

[50] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, "A unifying review of deep and shallow anomaly detection," *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, May 2021.

[51] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10142–10162, Aug. 2022.

[52] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *Int. J. Comput. Vis.*, vol. 130, no. 10, pp. 2425–2452, Oct. 2022.

[53] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022.

[54] X. Zhong, B. Gallagher, S. Liu, B. Kailkhura, A. Hiszpanski, and T. Y.-J. Han, "Explainable machine learning in materials science," *Npj Comput. Mater.*, vol. 8, no. 1, p. 204, Sep. 2022.

[55] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A survey of the state of explainable AI for natural language processing," 2020, *arXiv:2010.00711*.

[56] D. Poole, A. Mackworth, and R. Goebel, *Computational Intelligence: A Logical Approach*. Oxford, U.K.: Oxford Univ. Press, 1998.

[57] M. Garnelo and M. Shanahan, "Reconciling deep learning with symbolic artificial intelligence: Representing objects and relations," *Current Opinion Behav. Sci.*, vol. 29, pp. 17–23, Oct. 2019.

[58] S. Harnad, "The symbol grounding problem," *Phys. D, Nonlinear Phenomena*, vol. 42, nos. 1–3, pp. 335–346, Jun. 1990.

[59] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM J. Res. Develop.*, vol. 3, no. 3, pp. 210–229, Jul. 1959.

[60] J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, "An overview of machine learning," *Mach. Learn.*, vol. 1, pp. 3–23, Jan. 1983.

[61] P. Bhavsar, I. Safro, N. Bouaynaya, R. Polikar, and D. Dera, "Chapter 12—Machine learning in transportation data analytics," in *Data Analytics for Intelligent Transportation Systems*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 283–307.

[62] Asimov. (2016). *The Neural Network Zoo—The Asimov Institute*. Accessed: Apr. 2023. [Online]. Available: https://www.asimovinstitute.org/neural-network-zoo/

[63] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794.

[64] B. Ghojogh and M. Crowley, "The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial," 2019, *arXiv:1905.12787*.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[66] O. Goldreich, "Computational complexity: A conceptual perspective," *ACM Sigact News*, vol. 39, pp. 35–39, Sep. 2008.

[67] S. Arora and B. Barak, *Computational Complexity: A Modern Approach*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[68] E. Briscoe and J. Feldman, "Conceptual complexity and the bias/variance tradeoff," *Cognition*, vol. 118, no. 1, pp. 2–16, Jan. 2011.

[69] S. Fortmann-Roe. (2012). *Accurately Measuring Model Prediction Error*. [Online]. Available: http://scott.fortmann-roe.com/docs/MeasuringError.html

[70] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Lacoste-Julien, and I. Mitliagkas, "A modern take on the bias-variance tradeoff in neural networks," 2018, *arXiv:1810.08591*.

[71] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Heidelberg, Germany: Springer, 2009, vol. 2.

[72] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[73] S. Fortmann-Roe. (2012). *Understanding the Bias-Variance Tradeoff*. Accessed: Apr. 2023. [Online]. Available: http://scott.fortmann-roe.com/docs/BiasVariance.html

[74] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.*, vol. 4, no. 1, pp. 1–58, Jan. 1992.

[75] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani, "Reconstructing training data from trained neural networks," 2022, *arXiv:2206.07758*.

[76] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Commun. ACM*, vol. 64, no. 3, pp. 107–115, Mar. 2021.

[77] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 48, pp. 30063–30070, Dec. 2020.

[78] K. Wang, V. Muthukumar, and C. Thrampoulidis, "Benign overfitting in multiclass classification: All roads lead to interpolation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24164–24179.

[79] J. Jung, C. Concannon, R. Shroff, S. Goel, and D. G. Goldstein, "Simple rules for complex decisions," 2017, *arXiv:1702.04690*.

[80] J. M. Alonso, C. Castiello, and C. Mencar, "Interpretability of fuzzy systems: Current research trends and prospects," in *Springer Handbook of Computational Intelligence*. London, U.K.: Springer, 2015, pp. 219–237.

[81] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 916–954, Sep. 2008.

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

[83] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Statist.*, vol. 24, no. 1, pp. 44–65, Jan. 2015.

[84] G. Casalicchio, C. Molnar, and B. Bischl, "Visualizing the feature importance for black box models," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. London, U.K.: Springer, 2018, pp. 655–670.

[85] D. Kahneman, O. Sibony, and C. R. Sunstein, *Noise: A Flaw in Human Judgment*. New York, NY, USA: Little, Brown and Company, 2021.

[86] I. Kumar, C. Scheidegger, S. Venkatasubramanian, and S. Friedler, "Shapley residuals: Quantifying the limits of the Shapley value for explanations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–11.

[87] S. Passi and S. J. Jackson, "Trust in data science: Collaboration, translation, and accountability in corporate data science projects," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, pp. 1–28, Nov. 2018.

[88] D. Doran, S. Schulz, and T. R. Besold, "What does explainable AI really mean? A new conceptualization of perspectives," 2017, *arXiv:1710.00794*.

[89] T. Lombrozo, "The structure and function of explanations," *Trends Cognit. Sci.*, vol. 10, no. 10, pp. 464–470, Oct. 2006.

[90] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proc. Conf. Fairness, Accountability, Transparency*. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 648–657.

[91] M. Gleicher, "A framework for considering comprehensibility in modeling," *Big Data*, vol. 4, no. 2, pp. 75–88, Jun. 2016.

[92] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An Introduction to Statistical Learning: With Applications in R*, vol. 1. Heidelberg, Germany: Spinger, 2017.

[93] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[94] S. Vashishth, S. Upadhyay, G. Singh Tomar, and M. Faruqui, "Attention interpretability across NLP tasks," 2019, *arXiv:1909.11218*.

[95] S. Jain and B. C. Wallace, "Attention is not explanation," 2019, *arXiv:1902.10186*.

[96] R. Munroe. (Jun. 2023). *XKCD—A Webcomic of Romance, Sarcasm, Math, and Language*. [Online]. Available: https://xkcd.com/1838/

[97] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.

[98] P. E. Rauber, S. G. Fadel, A. X. Falcão, and A. C. Telea, "Visualizing the hidden activity of artificial neural networks," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 101–110, Jan. 2017.

[99] R. Yousefzadeh and D. P. O'Leary, "Interpreting neural networks using flip points," 2019, *arXiv:1903.08789*.

[100] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," 2019, *arXiv:1908.04626*.

[101] H. Karimi, T. Derr, and J. Tang, "Characterizing the decision boundary of deep neural networks," 2019, *arXiv:1912.11460*.

[102] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[103] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[104] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[105] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.

[106] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever Hans predictors and assessing what machines really learn," *Nature Commun.*, vol. 10, no. 1, pp. 1–8, Mar. 2019.

[107] G. Ras, M. van Gerven, and P. Haselager, "Explanation methods in deep learning: Users, values, concerns and challenges," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*. London, U.K.: Springer, 2018, pp. 19–36.

[108] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proc. CHI Conf. Human Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–52.

[109] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Long Beach, CA, USA: Curran Associates, 2017, pp. 4768–4777.

[110] M. Yang and B. Kim, "Benchmarking attribution methods with relative feature importance," 2019, *arXiv:1907.09701*.

[111] R. Kass and T. Finin, "The need for user models in generating expert system explanations," *Int. J. Exp. Syst.*, vol. 1, no. 4, pp. 1–31, 1988.

[112] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Chicago, IL, USA, 2013, pp. 623–631.

[113] H. Chen, S. Lundberg, and S.-I. Lee, "Explaining models by propagating Shapley values of local components," in *Explainable AI in Healthcare and Medicine*. New York City, NY, USA: Springer, 2021, pp. 261–270.
[114] H. Chen, S. M. Lundberg, and S.-I. Lee, "Explaining a series of models by propagating Shapley values," *Nature Commun.*, vol. 13, no. 1, pp. 1–15, Aug. 2022.
[115] M. Doumpos and C. Zopounidis, "Model combination for credit risk assessment: A stacked generalization approach," *Ann. Oper. Res.*, vol. 151, no. 1, pp. 289–306, Feb. 2007.
[116] S. P. Healey, W. B. Cohen, Z. Yang, C. K. Brewer, E. B. Brooks, N. Gorelick, A. J. Hernandez, C. Huang, M. J. Hughes, R. E. Kennedy, T. R. Loveland, G. G. Moisen, T. A. Schroeder, S. V. Stehman, J. E. Vogelmann, C. E. Woodcock, L. Yang, and Z. Zhu, "Mapping forest change using stacked generalization: An ensemble approach," *Remote Sens. Environ.*, vol. 204, pp. 717–728, Jan. 2018.
[117] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values," *Artif. Intell.*, vol. 298, Sep. 2021, Art. no. 103502.
[118] G. Yeuk-Yin Chan, E. Bertini, L. Gustavo Nonato, B. Barr, and C. T. Silva, "Melody: Generating and visualizing machine learning model summary to understand data and classifiers together," 2020, *arXiv:2007.10614*.
[119] H. Kumar and J. Chandran, "Is Shapley explanation for a model unique?" 2021, *arXiv:2111.11946*.
[120] J. Yuan, G. Yeuk-Yin Chan, B. Barr, K. Overton, K. Rees, L. Gustavo Nonato, E. Bertini, and C. T. Silva, "SUBPLEX: Towards a better understanding of black box model explanations at the subpopulation level," 2020, *arXiv:2007.10609*.
[121] M. Wojtas and K. Chen, "Feature importance ranking for deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 5105–5114.
[122] 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.
[123] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," 2018, *arXiv:1806.08049*.
[124] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing design practices for explainable AI user experiences," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Honolulu, HI, USA, Apr. 2020, pp. 1–15.
[125] T. Tam Nguyen, T. Trung Huynh, Z. Ren, T. Toan Nguyen, P. Le Nguyen, H. Yin, and Q. Viet Hung Nguyen, "A survey of privacy-preserving model explanations: Privacy risks, attacks, and countermeasures," 2024, *arXiv:2404.00673*.
[126] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," 2020, *arXiv:2006.11371*.
[127] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2022.
[128] S. Mishra, S. Dutta, J. Long, and D. Magazzeni, "A survey on the robustness of feature importance and counterfactual explanations," 2021, *arXiv:2111.00358*.
[129] S. Das, A. M. Javid, P. B. Gohain, Y. C. Eldar, and S. Chatterjee, "Neural greedy pursuit for feature selection," 2022, *arXiv:2207.09390*.
[130] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[131] L. G. Nonato and M. Aupetit, "Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 8, pp. 2650–2673, Aug. 2019.
[132] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[133] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," 2018, *arXiv:1805.10820*.
[134] C. Zednik, "Solving the black box problem: A normative framework for explainable artificial intelligence," *Philosophy Technol.*, vol. 34, no. 2, pp. 265–288, Jun. 2021.
[135] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Beijing, China, Aug. 2012, pp. 150–158.
[136] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*. Oxfordshire, U.K.: Routledge, 2017.
[137] D. Brown, C. Godfrey, N. Konz, J. Tu, and H. Kvinge, "Understanding the inner workings of language models through representation dissimilarity," 2023, *arXiv:2310.14993*.
[138] E. C. Alexandrina, E. S. Ortigossa, E. S. Lui, J. A. S. Gonçalves, N. A. Correa, L. G. Nonato, and M. L. Aguiar, "Analysis and visualization of multidimensional time series: Particulate matter (PM10) from São carlos-SP (Brazil)," *Atmos. Pollut. Res.*, vol. 10, no. 4, pp. 1299–1311, Jul. 2019.
[139] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. Chau, "ActiVis: Visual exploration of industry-scale deep neural network models," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 88–97, Jan. 2018.
[140] J.-L. Wu, P.-C. Chang, C. Wang, and K.-C. Wang, "ATICVis: A visual analytics system for asymmetric transformer models interpretation and comparison," *Appl. Sci.*, vol. 13, no. 3, p. 1595, Jan. 2023.
[141] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 8, pp. 2674–2693, Aug. 2019.
[142] W. E. Marcílio-Jr, D. M. Eler, and F. Breve, "Model-agnostic interpretation by visualization of feature perturbations," 2021, *arXiv:2101.10502*.
[143] Q. Zhao and T. Hastie, "Causal interpretations of black-box models," *J. Bus. Econ. Statist.*, vol. 39, no. 1, pp. 272–281, Jan. 2021.
[144] P. Xenopoulos, G. Chan, H. Doraiswamy, L. G. Nonato, B. Barr, and C. Silva, "GALE: Globally assessing local explanations," in *Proc. ICML*, Jul. 2022, pp. 322–331.
[145] G. Singh, F. Memoli, and G. E. Carlsson, "Topological methods for the analysis of high dimensional data sets and 3D object recognition," *PBG@Eurographics*, vol. 2, pp. 91–100, Sep. 2007.
[146] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau, "FAIRVIS: Visual analytics for discovering intersectional bias in machine learning," in *Proc. IEEE Conf. Vis. Analytics Sci. Technol. (VAST)*, Vancouver, BC, Canada, Oct. 2019, pp. 46–56.
[147] E. S. Ortigossa, F. F. Dias, and D. C. D. Nascimento, "Getting over high-dimensionality: How multidimensional projection methods can assist data science," *Appl. Sci.*, vol. 12, no. 13, p. 6799, Jul. 2022.
[148] G. D. Cantareira, E. Etemad, and F. V. Paulovich, "Exploring neural network hidden layer activity using vector fields," *Information*, vol. 11, no. 9, p. 426, Aug. 2020.
[149] J. Yuan, G. Y. Chan, B. Barr, K. Overton, K. Rees, L. G. Nonato, E. Bertini, and C. T. Silva, "SUBPLEX: A visual analytics approach to understand local model explanations at the subpopulation level," *IEEE Comput. Graph. Appl.*, vol. 42, no. 6, pp. 24–36, Nov. 2022.
[150] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.
[151] Y. Li, L. Ding, and X. Gao, "On the decision boundary of deep neural networks," 2018, *arXiv:1808.05385*.
[152] S. Guan and M. Loew, "Analysis of generalizability of deep neural networks based on the complexity of decision boundary," in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Miami, FL, USA, Dec. 2020, pp. 101–106.
[153] A. Englhardt, H. Trittenbach, D. Kottke, B. Sick, and K. Böhm, "Efficient SVDD sampling with approximation guarantees for the decision boundary," 2020, *arXiv:2009.13853*.
[154] J. Sohns, C. Garth, and H. Leitte, "Decision boundary visualization for counterfactual reasoning," *Comput. Graph. Forum*, vol. 42, no. 1, pp. 7–20, Feb. 2023.
[155] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harv. JL Tech.*, vol. 31, p. 841, Jan. 2017.
[156] A. Jacovi, S. Swayamdipta, S. Ravfogel, Y. Elazar, Y. Choi, and Y. Goldberg, "Contrastive explanations for model interpretability," 2021, *arXiv:2103.01378*.
[157] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, and C. Shah, "Counterfactual explanations and algorithmic recourses for machine learning: A review," 2020, *arXiv:2010.10596*.
[158] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "FACE: Feasible and actionable counterfactual explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 344–350.
[159] M. M. Raimundo, L. G. Nonato, and J. Poco, "Mining Pareto-optimal counterfactual antecedents with a branch-and-bound model-agnostic algorithm," *Data Mining Knowl. Discovery*, pp. 1–33, Dec. 2022.
[160] R. M. J. Byrne, "Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 6276–6282.
[161] U. Aïvodji, A. Bolot, and S. Gambs, "Model extraction from counterfactual explanations," 2020, *arXiv:2009.01884*.

[162] P. Dissanayake and S. Dutta, "Model reconstruction using counterfactual explanations: Mitigating the decision boundary shift," 2024, *arXiv:2405.05369*.

[163] S. Barocas, A. D. Selbst, and M. Raghavan, "The hidden assumptions behind counterfactual explanations and principal reasons," in *Proc. Conf. Fairness, Accountability, Transparency*. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 80–89.

[164] A. Van Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*. London, U.K.: Springer, 2021, pp. 650–665.

[165] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19620–19631.

[166] M. Jia, B. Gabrys, and K. Musial, "A network science perspective of graph convolutional networks: A survey," *IEEE Access*, vol. 11, pp. 39083–39122, 2023.

[167] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNexplainer: Generating explanations for graph neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[168] M. Vu and M. T. Thai, "PGM-Explainer: Probabilistic graphical model explanations for graph neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12225–12235.

[169] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji, "On explainability of graph neural networks via subgraph explorations," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12241–12252.

[170] L. S. Shapley, "A value for n-person games," in *Contributions to the Theory of Games (AM-28), Volume II*. Princeton, NJ, USA: Princeton Univ. Press, 1953, p. 2.

[171] H. Yuan, J. Tang, X. Hu, and S. Ji, "XGNN: Towards model-level explanations of graph neural networks," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 430–438.

[172] F. Baldassarre and H. Azizpour, "Explainability techniques for graph convolutional networks," 2019, *arXiv:1905.13686*.

[173] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5782–5799, May 2023.

[174] K. Amara, R. Ying, Z. Zhang, Z. Han, Y. Shan, U. Brandes, S. Schemm, and C. Zhang, "GraphFramEx: Towards systematic evaluation of explainability methods for graph neural networks," 2022, *arXiv:2206.09677*.

[175] L. Faber, A. K. Moghaddam, and R. Wattenhofer, "When comparing to ground truth is wrong: On evaluating GNN explanation methods," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 332–341.

[176] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," 2016, *arXiv:1602.02410*.

[177] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," 2017, *arXiv:1702.00887*.

[178] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, and E. Raff, "Pythia: A suite for analyzing large language models across training and scaling," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 2397–2430.

[179] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proc. 57th Annu. Meeting Assoc. for Comput. Linguistics: Syst. Demonstrations*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 37–42.

[180] S. Biderman, U. Sai Prashanth, L. Sutawika, H. Schoelkopf, Q. Anthony, S. Purohit, and E. Raff, "Emergent and predictable memorization in large language models," 2023, *arXiv:2304.11158*.

[181] J. Vig, "BertViz: A tool for visualizing multihead self-attention in the BERT model," in *Proc. ICLR Workshop: Debugging Mach. Learn. Models*, vol. 23, 2019, pp. 1–6.

[182] A. Garde, E. Kran, and F. Barez, "DeepDecipher: Accessing and investigating neuron activation in large language models," 2023, *arXiv:2310.01870*.

[183] A. Foote, N. Nanda, E. Kran, I. Konstas, S. Cohen, and F. Barez, "Neuron to graph: Interpreting language model neurons at scale," 2023, *arXiv:2305.19911*.

[184] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 782–791.

[185] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.

[186] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[187] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," Univ. Montreal, Tech. Rep., 1341, 2009.

[188] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision—ECCV*, Zurich, Switzerland. Springer, 2014, pp. 818–833.

[189] C. Agarwal, D. Ley, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, and H. Lakkaraju, "OpenXAI: Towards a transparent evaluation of model explanations," 2022, *arXiv:2206.11104*.

[190] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju, "The disagreement problem in explainable machine learning: A practitioner's perspective," 2022, *arXiv:2202.01602*.

[191] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018.

[192] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[193] E. S. Ortigossa, F. F. Dias, B. Barr, C. T. Silva, and L. Gustavo Nonato, "T-explainer: A model-agnostic explainability framework based on gradients," 2024, *arXiv:2404.16495*.

[194] M. Hamilton, S. Lundberg, L. Zhang, S. Fu, and W. T. Freeman, "Axiomatic explanations for visual search, retrieval, and similarity learning," 2021, *arXiv:2103.00370*.

[195] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, "Towards best practice in explaining neural network decisions with LRP," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Glasgow, U.K., Jul. 2020, pp. 1–7.

[196] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[197] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[198] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.

[199] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Trans. Image Process.*, vol. 30, pp. 5875–5888, 2021.

[200] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," 2016, *arXiv:1605.01713*.

[201] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[202] P. Sturmfels, S. Lundberg, and S.-I. Lee, "Visualizing the impact of feature attribution baselines," *Distill*, vol. 5, no. 1, p. e22, 2020.

[203] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[204] S. Hu, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E. F. Fang, W. Menpes-Smith, J. Xia, H. Ye, and G. Yang, "Weakly supervised deep learning for COVID-19 infection detection and classification from CT images," *IEEE Access*, vol. 8, pp. 118869–118883, 2020.

[205] M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2020, pp. 9269–9278.

[206] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: New computational modelling techniques for genomics," *Nature Rev. Genet.*, vol. 20, no. 7, pp. 389–403, Jul. 2019.

[207] J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, and K. N. Ramamurthy, "TreeView: Peeking into deep neural networks via feature-space partitioning," 2016, *arXiv:1611.07429*.

[208] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," 2016, *arXiv:1606.05386*.

[209] M. T. Ribeiro, S. Singh, and C. Guestrin, "Nothing else matters: Model-agnostic explanations by identifying prediction invariance," 2016, *arXiv:1611.05817*.

[210] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–9.
[211] B. Lim, S. Ö. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021.
[212] R. Yousefzadeh and D. P. O'Leary, "Investigating decision boundaries of trained neural networks," 2019, *arXiv:1908.02802*.
[213] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 3, pp. 525–541, Jun. 2021.
[214] R. Turner, "A model explanation system," in *Proc. IEEE 26th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2016, pp. 1–6.
[215] H. Lakkaraju, N. Arsov, and O. Bastani, "Robust and stable black box explanations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5628–5638.
[216] X. Situ, I. Zukerman, C. Paris, S. Maruf, and G. Haffari, "Learning to explain: Generating stable explanations fast," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5340–5355.
[217] E. Friedman and H. Moulin, "Three methods to share joint costs or surplus," *J. Econ. Theory*, vol. 87, no. 2, pp. 275–312, Aug. 1999.
[218] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 5491–5500.
[219] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach," *Appl. Stochastic Models Bus. Ind.*, vol. 17, no. 4, pp. 319–330, Oct. 2001.
[220] E. Strumbelj and I. Kononenko, "An efficient explanation of individual classifications using game theory," *J. Mach. Learn. Res.*, vol. 11, pp. 1–18, Mar. 2010.
[221] E. Strumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, Dec. 2014.
[222] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," 2018, *arXiv:1802.03888*.
[223] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.* New York, NY, USA: Association for Computing Machinery, 2020, pp. 180–186.
[224] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," 2018, *arXiv:1806.10758*.
[225] G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance," 2019, *arXiv:1905.03151*.
[226] G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance," *Statist. Comput.*, vol. 31, no. 6, pp. 1–16, Nov. 2021.
[227] D. Janzing, L. Minorics, and P. Blöbaum, "Feature relevance quantification in explainable AI: A causal problem," 2019, *arXiv:1910.13413*.
[228] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: Understanding data Scientists' use of interpretability tools for machine learning," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2020, pp. 1–14.
[229] Y. Liu, S. Khandagale, C. White, and W. Neiswanger, "Synthetic benchmarks for scientific research in explainable machine learning," 2021, *arXiv:2106.12543*.
[230] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, p. 593, Mar. 2021.
[231] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," 2021, *arXiv:2102.13076*.
[232] I. Hameed, S. Sharpe, D. Barcklow, J. Au-Yeung, S. Verma, J. Huang, B. Barr, and C. B. Bruss, "BASED-XAI: Breaking ablation studies down for explainable artificial intelligence," 2022, *arXiv:2207.05566*.
[233] B. Barr, N. Fatsi, L. Hancox-Li, P. Richter, D. Proano, and C. Mok, "The disagreement problem in faithfulness metrics," in *Proc. NIPS*, New Orleans, LA, USA, 2023, pp. 1–13.
[234] J. Haug, S. Zürn, P. El-Jiz, and G. Kasneci, "On baselines for local feature attributions," 2021, *arXiv:2101.00905*.
[235] M. Ye and Y. Sun, "Variable selection via penalized neural network: A drop-out-one loss approach," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5620–5629.
[236] H. J. P. Weerts, W. van Ipenburg, and M. Pechenizkiy, "A human-grounded evaluation of SHAP for alert processing," 2019, *arXiv:1907.03324*.
[237] S. Tsirtsis and M. Gomez Rodriguez, "Decisions, counterfactual explanations and strategic behavior," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16749–16760.
[238] A. Jeyasothy, T. Laugel, M.-J. Lesot, C. Marsala, and M. Detyniecki, "Integrating prior knowledge in post-hoc explanations," 2022, *arXiv:2204.11634*.
[239] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," 2018, *arXiv:1806.07538*.
[240] J. DeYoung, S. Jain, N. Fatema Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "ERASER: A benchmark to evaluate rationalized NLP models," 2019, *arXiv:1911.03429*.
[241] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 17, pp. 14867–14875.
[242] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*.
[243] C. Agarwal, N. Johnson, M. Pawelczyk, S. Krishna, E. Saxena, M. Zitnik, and H. Lakkaraju, "Rethinking stability for attribution-based explanations," 2022, *arXiv:2203.06877*.
[244] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, "Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond," *J. Mach. Learn. Res.*, vol. 24, no. 34, pp. 1–11, 2023.
[245] T. Han, S. Srinivas, and H. Lakkaraju, "Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 5256–5268.
[246] R. F. Barber and E. J. Candès, "Controlling the false discovery rate via knockoffs," *Ann. Statist.*, vol. 43, no. 5, pp. 2055–2085, Oct. 2015.
[247] K. Banachewicz, L. Massaron, and A. Goldbloom, *The Kaggle Book: Data Analysis and Machine Learning for Competitive Data Science*. Birmingham, U.K.: Packt Publishing, 2022.
[248] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, 1998.
[249] A. R. de Leon and K. C. Carrière, "A generalized Mahalanobis distance for mixed data," *J. Multivariate Anal.*, vol. 92, no. 1, pp. 174–185, Jan. 2005.
[250] A. Bunt, M. Lount, and C. Lauzon, "Are explanations always important: A study of deployed, low-cost intelligent interactive systems," in *Proc. ACM Int. Conf. Intell. User Interfaces*. New York, NY, USA: Association for Computing Machinery, Feb. 2012, pp. 169–178.
[251] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 11, pp. 745–750, Nov. 2021.
[252] C. Meske and E. Bunde, "Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support," in *Artificial Intelligence in HCI*, Copenhagen, Denmark. Springer, 2020, pp. 54–69.
[253] A. Zhang, L. Xing, J. Zou, and J. C. Wu, "Shifting machine learning for healthcare from development to deployment and from models to data," *Nature Biomed. Eng.*, vol. 6, no. 12, pp. 1330–1345, Jul. 2022.
[254] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.
[255] S. Qiu, M. I. Miller, P. S. Joshi, J. C. Lee, C. Xue, Y. Ni, Y. Wang, I. De Anda-Duran, P. H. Hwang, and J. A. Cramer, "Multimodal deep learning for Alzheimer's disease dementia assessment," *Nature Commun.*, vol. 13, no. 1, p. 3404, 2022.
[256] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *Npj Digit. Med.*, vol. 4, no. 1, pp. 1–5, Jan. 2021.
[257] E. R. Pfaff, A. T. Girvin, T. D. Bennett, A. Bhatia, I. M. Brooks, R. R. Deer, J. P. Dekermanjian, S. E. Jolley, M. G. Kahn, K. Kostka, and J. A. McMurry, "Identifying who has long COVID in the USA: A machine learning approach using N3C data," *Lancet Digit. Health*, vol. 4, no. 7, pp. 532–541, Jul. 2022.

[258] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays," *Comput. Methods Programs Biomed.*, vol. 196, Nov. 2020, Art. no. 105608.

[259] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Comput. Biol. Med.*, vol. 121, Jun. 2020, Art. no. 103792.

[260] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[261] H. A. Elmarakeby, J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S. H. AlDubayan, K. Salari, S. Kregel, C. Richter, T. E. Arnoff, J. Park, W. C. Hahn, and E. M. Van Allen, "Biologically informed deep neural network for prostate cancer discovery," *Nature*, vol. 598, no. 7880, pp. 348–352, Oct. 2021.

[262] R. J. Chen, M. Y. Lu, J. Wang, D. F. K. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood, "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Trans. Med. Imag.*, vol. 41, no. 4, pp. 757–770, Apr. 2022.

[263] D. Wang, C. Zhang, B. Wang, B. Li, Q. Wang, D. Liu, H. Wang, Y. Zhou, L. Shi, F. Lan, and Y. Wang, "Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning," *Nature Commun.*, vol. 10, no. 1, p. 4284, Sep. 2019.

[264] N. Bar, T. Korem, O. Weissbrod, D. Zeevi, D. Rothschild, S. Leviatan, N. Kosower, M. Lotan-Pompan, A. Weinberger, and C. I. Le Roy, "A reference map of potential determinants for the human serum metabolome," *Nature*, vol. 588, no. 7836, pp. 135–140, 2020.

[265] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley, "Effective gene expression prediction from sequence by integrating long-range interactions," *Nature Methods*, vol. 18, no. 10, pp. 1196–1203, Oct. 2021.

[266] T. Buergel, J. Steinfeldt, G. Ruyoga, M. Pietzner, D. Bizzarri, D. Vojinovic, J. U. Z. Belzen, L. Loock, P. Kittner, and L. Christmann, "Metabolomic profiles predict individual multidisease outcomes," *Nature Med.*, vol. 28, no. 11, pp. 2309–2320, Nov. 2022.

[267] A. Chklovski, D. H. Parks, B. J. Woodcroft, and G. W. Tyson, "CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning," *Nature Methods*, vol. 20, no. 8, pp. 1203–1212, Aug. 2023.

[268] S. Pratt, I. Covert, R. Liu, and A. Farhadi, "What does a platypus look like? Generating customized prompts for zero-shot image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 15691–15701.

[269] J. Sarzynska-Wawer, A. Wawer, A. Pawlak, J. Szymanowska, I. Stefaniak, M. Jarkiewicz, and L. Okruszek, "Detecting formal thought disorder by deep contextualized word representations," *Psychiatry Res.*, vol. 304, Oct. 2021, Art. no. 114135.

[270] C. W. Hong, C. Lee, K. Lee, M.-S. Ko, and K. Hur, "Explainable artificial intelligence for the remaining useful life prognosis of the turbofan engines," in *Proc. 3rd IEEE Int. Conf. Knowl. Innov. Invention (ICKII)*, Kaohsiung, Taiwan, Aug. 2020, pp. 144–147.

[271] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte, "An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery," *Mech. Syst. Signal Process.*, vol. 163, Jan. 2022, Art. no. 108105.

[272] K. Xu, J. Yuan, Y. Wang, C. Silva, and E. Bertini, "MTSeer: Interactive visual exploration of models on multivariate time-series forecast," in *Proc. CHI Conf. Human Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, 2021, pp. 1–15.

[273] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Anal. Prevention*, vol. 136, Mar. 2020, Art. no. 105405.

[274] B. Sanchez-Lengeling, J. N. Wei, B. K. Lee, R. C. Gerkin, A. Aspuru-Guzik, and A. B. Wiltschko, "Machine learning for scent: Learning generalizable perceptual representations of small molecules," 2019, *arXiv:1910.10685*.

[275] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, and T. Unterthiner, "Interpretable deep learning in drug discovery," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. London, U.K.: Springer, 2019, pp. 331–345.

[276] K. McCloskey, A. Taly, F. Monti, M. P. Brenner, and L. J. Colwell, "Using attribution to decode binding mechanism in neural network models for chemistry," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 24, pp. 11624–11629, Jun. 2019.

[277] P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, and T. Laino, "Extraction of organic chemistry grammar from unsupervised learning of chemical reactions," *Sci. Adv.*, vol. 7, no. 15, Apr. 2021, Art. no. eabe4166.

[278] J. Yang, L. Tao, J. He, J. R. McCutcheon, and Y. Li, "Machine learning enables interpretable discovery of innovative polymers for gas separation membranes," *Sci. Adv.*, vol. 8, no. 29, Jul. 2022, Art. no. eabn9545.

[279] D. Jiang, Z. Wu, C.-Y. Hsieh, G. Chen, B. Liao, Z. Wang, C. Shen, D. Cao, J. Wu, and T. Hou, "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models," *J. Cheminformatics*, vol. 13, no. 1, pp. 1–23, Feb. 2021.

[280] J. Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug discovery with explainable artificial intelligence," *Nature Mach. Intell.*, vol. 2, no. 10, pp. 573–584, Oct. 2020.

[281] A. Davies, P. Velickovic, L. Buesing, S. Blackwell, D. Zheng, N. Tomasev, R. Tanburn, P. Battaglia, C. Blundell, A. Juhász, M. Lackenby, G. Williamson, D. Hassabis, and P. Kohli, "Advancing mathematics by guiding human intuition with AI," *Nature*, vol. 600, no. 7887, pp. 70–74, Dec. 2021.

[282] C. Molnar, G. Casalicchio, and B. Bischl, "iml: An R package for interpretable machine learning," *J. Open Source Softw.*, vol. 3, no. 26, p. 786, Jun. 2018.

[283] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "InterpretML: A unified framework for machine learning interpretability," 2019, *arXiv:1909.09223*.

[284] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for PyTorch," 2020, *arXiv:2009.07896*.

[285] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "iNNvestigate neural networks!" *J. Mach. Learn. Res.*, vol. 20, no. 93, pp. 1–8, 2019.

[286] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. Vera Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques," 2019, *arXiv:1909.03012*.

[287] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "AI explainability 360 toolkit," in *Proc. 3rd ACM India Joint Int. Conf. Data Sci. Manag. Data*, Jan. 2021, pp. 376–379.

[288] J. Klaise, A. Van Looveren, G. Vacanti, and A. Coca, "Alibi explain: Algorithms for explaining machine learning models," *J. Mach. Learn. Res.*, vol. 22, no. 181, pp. 1–7, 2021.

[289] W. Yang, H. Le, T. Laud, S. Savarese, and S. C. H. Hoi, "OmniXAI: A library for explainable AI," 2022, *arXiv:2206.01612*.

[290] A. Theodorou, R. H. Wortham, and J. J. Bryson, "Designing and implementing transparency for real time inspection of autonomous robots," *Connection Sci.*, vol. 29, no. 3, pp. 230–241, Jul. 2017.

[291] A. Kucharski, "Study epidemiology of fake news," *Nature*, vol. 540, no. 7634, p. 525, Dec. 2016.

[292] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, "Stakeholders in explainable AI," 2018, *arXiv:1810.00184*.

[293] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[294] L. Deck, A. Schomäcker, T. Speith, J. Schöffer, L. Kästner, and N. Kuhl, "Mapping the potential of explainable artificial intelligence (XAI) for fairness along the AI lifecycle," 2024, *arXiv:2404.18736*.

[295] H. Shefrin, *Beyond Greed and Fear: Understanding Behavioral Finance and the Psychology of Investing*. Oxford, U.K.: Oxford Univ. Press, 2002.

[296] N. N. Taleb, *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*, vol. 1. New York City, NY, USA: Random House Incorporated, 2005.

EVANDRO S. ORTIGOSSA received the B.Sc. degree in computer science and the M.Sc. and Ph.D. degrees in computer science and computational mathematics from the Institute of Mathematics and Computer Science, University of São Paulo (ICMC-USP), São Carlos, Brazil, in 2015, 2018, and 2024, respectively, focusing his work on multidimensional time-series analysis and machine learning.

He is currently a member of the Graphics, Imaging, Visualization, and Analytics Group (GIVA), ICMC-USP, where he develops research on machine learning. His research interests include data science, machine learning, explainable artificial intelligence (XAI), information visualization (InfoVis), and image processing.

THALES GONÇALVES received the B.Sc. degree in electrical engineering from the Federal Institute of Espírito Santo (IFES), in 2014, the B.Sc. degree in mathematics from the Center of Exact Sciences, Federal University of Espírito Santo (CCE-UFES), in 2017, the M.Sc. degree in signal processing and pattern recognition from the Technology Center, CT-UFES, in 2017, and the Ph.D. degree in computer science and computational mathematics from the Institute of Mathematics and Computer Science, University of São Paulo (ICMC-USP), in 2024.

He is currently a member of the Graphics, Imaging, Visualization, and Analytics Group (GIVA), ICMC-USP. During the Ph.D. study, he was a Visiting Scholar with the Visualization, Imaging, and Data Analysis Center (VIDA), New York University (NYU) Tandon School of Engineering. His research interests include machine learning, graph neural networks, and explainable/responsible AI.

LUIS GUSTAVO NONATO (Member, IEEE) received the Ph.D. degree in applied mathematics from Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil, in 1998.

From 2008 to 2010, he was a Visiting Scholar with the SCI Institute, The University of Utah. From 2016 to 2018, he was a Visiting Professor with the Center for Data Science, New York University. He is currently a Professor with the Institute of Mathematics and Computer Science, University of São Paulo (ICMC-USP), São Carlos, Brazil. His main research interests include visual analytics, geometric computing, data science, and visualization.

Dr. Nonato served on several program committees, including IEEE SciVis, IEEE InfoVis, and EuroVis. He was an Associate Editor of the *Computer Graphics Forum* journal and IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS and the Editor-in-Chief of the *International Journal of Applied Mathematics and Computational Sciences* (SBMAC SpringerBriefs).