Annals of Social Sciences & Management Studies
ISSN: 2641-838X

Juniper Publishers
key to the Researchers

Review article
Volume 13 Issue 3 - June 2026
DOI: 10.19080/ASM.2026.13.555864

Ann Soc Sci Manage Stud

Interpretable Machine Learning: Theory, Methods, and Applications

Jiangshan Zhu*

Postdoctoral Researcher, School of Management, Hangzhou Dianzi University, China

Submission: May 18, 2026; **Published:** June 09, 2026

***Corresponding author:** Jiangshan Zhu, Postdoctoral Researcher, School of Management, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract

With the development of artificial intelligence, its inherent "black-box" nature has sparked major concerns about accountability, fairness, and trust, which are particularly pronounced in critical domains such as healthcare, finance, and criminal justice. This paper offers a systematic and in-depth review of explainable artificial intelligence (XAI), emphasizing its key role in enhancing the interpretability and transparency of complex AI systems across various application scenarios. First, we seek to define interpretability in the context of machine learning and propose the PDR framework—Predictive, Descriptive, and Relevant—to address these issues. The PDR framework specifies three overarching evaluation objectives: predictive accuracy, descriptive accuracy, and relevance, where "relevance" is assessed with respect to human audiences. Next, we classify methods according to the models to which they apply: one category comprises interpretable methods based on self-explanatory models, and the other comprises interpretable methods based on external co-explanations. The latter is further subdivided into several branches: instance-based, SHAP-based, knowledge-graph-based, deep-learning-based, and clustering-model-based approaches. This taxonomy helps us better understand the characteristics of the models on which different explainability methods depend, thereby making it easier for researchers to select appropriate models to address interpretability problems. Finally, we review mainstream XAI applications across disciplinary and cross-disciplinary fields and discuss future challenges and developmental trends in interpretable machine learning, with the aim of promoting sustained progress in this area.

Keywords: PDR framework; Self-Explanatory models; External Co-explanations; XAI

Introduction

In recent years, complex models such as deep learning have achieved remarkable success in perception, prediction, and decision-making tasks [1]. Yet their "black-box" nature has raised systemic concerns about accountability, fairness, and trust in high-stakes settings—including healthcare, finance, and public governance—and has intensified scholarly confusion over what interpretability means, how to evaluate it, and how to choose among methods [2,3]. Centering these core issues, this paper adopts the Predictive–Descriptive–Relevant (PDR) framework to define interpretability within the machine-learning context and to treat the human audience as a key dimension for designing and assessing explanations, thereby establishing clear principles for method selection and performance evaluation [4]. The framework stresses that any explanation should simultaneously attend to: the model's predictive accuracy on the task; the descriptive fidelity of the explanation to the model's internal/external behavior; and the relevant utility and usability of the explanation for the target audience [4].

Further aligning with recent engineering practice, we adopt and extend the classification of intrinsic (self-explanatory) models vs. external co-explanations [5]. Intrinsic models embed interpretability during training through structural constraints and inductive biases [6]; external co-explanations follow a separated "explainer–explainee" paradigm, applying instance-/feature-level visualization or attribution analysis to an existing black box [5,7]. The latter can be further subdivided into instance-based, SHAP-based, knowledge-graph (KG)–based, deep network structure/behavior–based, and clustering-based subclasses, each with distinct trade-offs in portability, assumptions/constraints, and readability [5,7]. For example, LIME approximates complex decision boundaries with local linear surrogates and emphasizes local fidelity, but its domain of coverage and boundary stability require additional specification [8]; SHAP centers on feature contributions and offers a clear explanatory semantics in scenarios where decisions are made "by features" [9]; KG methods are advantageous for semantic reasoning and relational learning,

but they are not directly applicable to predictive explanations for non-semantic models such as SVMs [10].

Bringing the method taxonomy under the PDR lens helps clarify the discourse on "interpretability": explanations are not merely abstract "visual outputs," but design artifacts jointly constrained by task objectives, model behavior, and audience needs [2,4]. Recent surveys repeatedly emphasize this point: on the one hand, explanation methods must balance trust-building, model auditing, and domain knowledge discovery [2]; on the other hand, evaluation frameworks remain incomplete—especially lacking consistent, standardized metrics across tasks and audiences [3]. A three-layer evaluation perspective—application-level, human-level, and functional-level—can be used to validate explanations in a stratified manner, alleviating the measurement gap surrounding what counts as a "good" explanation [3].

Meanwhile, applications of interpretable learning have rapidly expanded across multiple disciplines and cross-disciplinary settings—healthcare, financial risk control, education, cybersecurity, and scientific discovery—forming a virtuous cycle in which applications pull method evolution, and methods in turn deepen understanding of domain mechanisms [11,12]. However, as model scale and complexity grow, the latency and computational cost of generating explanations pose new challenges [13]; heterogeneity in regulatory requirements, audience literacy, and usage contexts across domains further raises the engineering bar for producing explanations that are both usable and useful [12]. Moreover, constraints on user comprehension and cognitive load suggest that the form and granularity of explanations should match the decision-maker's knowledge structure, avoiding "apparently transparent but practically unusable" information overload [14].

Against this backdrop, the contributions and organization of this paper are as follows:

i. **Concepts and Evaluation:** We restate a working definition of interpretability within the machine-learning context and employ the PDR framework as a unifying evaluation thread throughout, highlighting the centrality of the human audience in the design and assessment of explanations [4].

ii. **Methodological Taxonomy:** From the engineering perspective of intrinsic vs. external co-explanations, we provide a systematic account of each subclass's assumptions, applicability boundaries, and trade-offs, illustrated with representative methods (e.g., LIME, SHAP, KG, and visualization/probe-based techniques for deep networks) [5,7-10].

iii. **Cross-Domain Applications and Frontier Topics:** For domains such as healthcare, finance, education, and cybersecurity, we distill an alignment template—task objectives–audience roles–regulatory constraints–evaluation pathways—to summarize progress and to propose open problems and a research agenda aimed at scalability, standardization, and human–AI collaboration [11,12,14].

PDR framework

This chapter establishes a unifying lens for interpretability by adopting the PDR framework and using it as the organizing thread for concepts, methods, and evaluation [4]. We first articulate what we mean by interpretability in machine learning and why a tripartite view—predictive accuracy, descriptive accuracy, and relevance to specific human audiences—is needed to adjudicate claims about explanations across tasks and domains [4]. We then analyze each desideratum in turn, clarifying how predictive performance, faithfulness to model behavior, and stakeholder-centered usefulness interact and sometimes conflict in practice [3]. Building on this, we map methodological families onto PDR—contrasting intrinsic (self-explanatory) models with external co-explanations—and explain the implications of each choice for model selection and deployment [5]. Finally, we translate PDR into practice by specifying improvement criteria, reproducible evaluation routes for descriptiveness, empirical tests of relevance, and reporting recommendations [3,4]. Together, these elements provide a coherent scaffold for the remainder of the survey and for the comparative analysis presented in subsequent sections.

Motivation and definition

Interpretability methods have proliferated, with outputs ranging from visualizations and natural-language rationales to mathematical expressions, yet consensus remains elusive on what counts as interpretability, how it should be evaluated, and how methods should be chosen [2,3]. To unify terminology and evaluation standards, Murdoch et al. proposed the PDR framework, which treats predictive accuracy, descriptive accuracy, and relevance as three overarching desiderata for constructing and assessing explanations, and explicitly requires that relevance be judged with respect to specific human audiences [4]. The framework also distinguishes two major method families—intrinsic and external—providing a common coordinate system to guide practitioners in selecting techniques for real-world tasks [5].

Predictive Accuracy

Predictiveness refers to how faithfully a learner approximates the true relationships in the data; if the underlying model fits poorly, any information extracted from it is difficult to trust [2]. In standard supervised learning, test-set performance serves as the operational measure of predictive accuracy [2]. In explainability settings, however, predictive evaluation must be aligned with the target population and usage context: the evaluation data should represent the domain that truly matters (for example, cohorts drawn from a single hospital rarely generalize to other institutions) [3], analysts should look beyond averages to the full distribution of errors to detect whether particular subgroups suffer systematically larger mistakes [3], and predictions should be stable and robust to reasonable perturbations of the data or

How to cite this article: Jiangshan Z. Interpretable Machine Learning: Theory, Methods, and Applications. Ann Soc Sci Manage Stud. 2026; 13(2): 555864. DOI: 10.19080/ASM.2026.13.555864

the model specification—when small sample deletions induce large performance swings, explanations built on such models are hard to rely on [3,13].

Descriptive accuracy

Descriptiveness denotes the extent to which an explanation method objectively characterizes the relationships a model has actually learned; in practice, it asks whether the explanation faithfully reflects the model's internal and external behavior [4]. This is especially challenging for complex black boxes such as deep networks, where nonlinear structure is stored in highly distributed, non-explicit forms [1]. A recurring tension arises between predictive and descriptive accuracy: intrinsically interpretable, structurally constrained models often achieve higher descriptive fidelity but may lose predictive performance on complex data, whereas highly expressive models—such as deep nets for images—can deliver superior predictive accuracy while being harder to analyze, thereby reducing descriptiveness [3]. When both forms of accuracy cannot be high at the same time, credibility should be established through external validation—additional experiments or independent evidence that corroborate the claims made by the explanations [3].

Relevancy

High accuracy alone is not enough; an explanation must be useful and usable for a specific audience and a specific problem [4]. Different stakeholders—patients, clinicians, biologists, statisticians—may require explanations of the same model that are consistent in substance yet differ in form and granularity [12]. We call an explanation relevant when it delivers actionable insight to a defined audience within a defined task context [4]. In trade-offs, relevance often determines whether to favor predictive or descriptive accuracy: when the goal is fairness auditing or accountability, descriptive fidelity takes priority, whereas when explanations are used chiefly to guide feature engineering or improve models, predictive performance comes first [3,12].

Mapping between methodological categories and PDR

Intrinsic, self-explanatory models versus external co-explanations—intrinsic approaches raise descriptive fidelity during model construction by imposing structural constraints such as sparsity, compositionality, and simulability [6]. Because they modify the predictor itself, they also influence predictive accuracy and are most suitable when the underlying relationships are relatively simple or when peak predictive performance is not the primary objective [6]. External co-explanations approaches, by contrast, extract information from an already trained black box at the analysis stage—through feature attributions, prototype and counterfactual reasoning, or hierarchical visualizations—thereby improving descriptiveness without directly altering predictive accuracy [5,7-9]. They are especially valuable when complex data structure necessitates black-box models to achieve acceptable performance [1,12]. In short, both families can increase descriptive accuracy, but only intrinsic/model-based methods affect predictive accuracy by design, whereas relevance determines which explanatory form and level of granularity will be most useful for the specific problem and audience [3,12].

Operationalization Assessment and Reporting Specifications

Improvement criterion. Within the PDR framework, a claim of methodological progress should raise at least one desideratum without materially degrading the other two [4]. Among the three, gains in predictive accuracy are the easiest to quantify, whereas reliable measurement of descriptive fidelity and relevance remains challenging [3].

Reproducible routes for evaluating descriptiveness. One route is simulation studies: construct a controllable data-generating process, train sufficiently powerful models to achieve near-perfect generalization, and then test whether the explanation can recover the known structure embedded in the generator, thereby providing a semi–ground truth yardstick; this paradigm has been demonstrated in settings such as neural networks and interaction detection [1,3]. A complementary route is post hoc comparison against established truths: when a domain already offers reliable experimental findings, these can serve as partial ground truth for retrospective validation of explanations [12].

Empirical evidence of relevance. The most direct test is task-level validation: if deploying the explanation in a real domain demonstrably advances scientific discovery or improves clinical decision-making, its relevance is self-evident [12]. In parallel, carefully designed human-subject studies can assess how explanations affect trust, understanding, and decision quality; despite challenges of participant representativeness and experimental control, such studies provide essential evidence that explanations are genuinely useful to practitioners [3].

Reporting recommendations. At a minimum, studies should present side-by-side evidence and costs for all three desiderata—for example, test and out-of-distribution performance, metrics of explanation–model consistency and stability, and user/task outcomes or perceived usefulness [3]. Authors should specify the intended audience and usage context to avoid showcasing novel visualizations without clarifying what problem they solve, and they should provide quantitative or experimental support for stability analyses and external validation [3].

Summary. The PDR framework grounds "interpretability" in three measurable objectives: predictive accuracy ensures the reliability of the underlying model, descriptive fidelity ensures that explanations are faithful to the model, and relevance ensures that explanations serve human needs [4]. Coupled with the engineering perspective of intrinsic versus external co-explanations, it offers an operational guide for method selection and experimental evaluation (Figure 1), and it naturally sets the stage for the subsequent sections on representative methods and cross-domain applications [5,12].

How to cite this article: Jiangshan Z. Interpretable Machine Learning: Theory, Methods, and Applications. Ann Soc Sci Manage Stud. 2026; 13(2): 555864. DOI: 10.19080/ASM.2026.13.555864

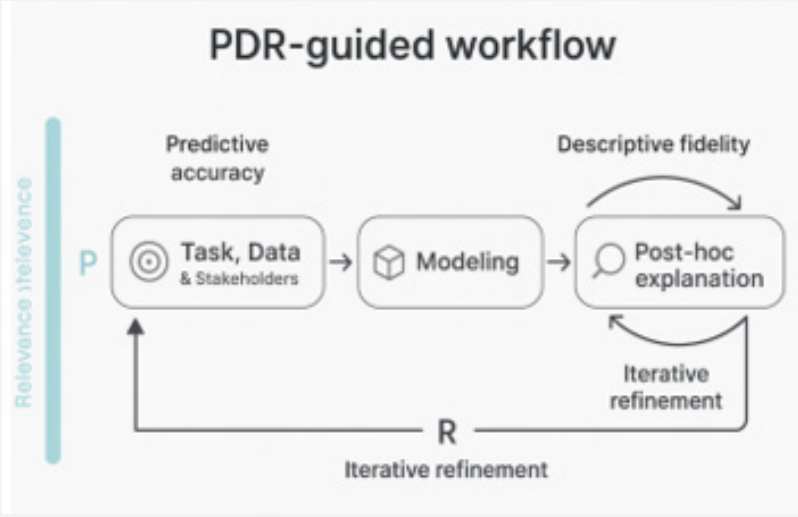

Figure 1: Overview of the cycle of PDR framework.

Interpretable Methods

Anchored in the PDR framework introduced in Chapter 2, this chapter presents two major families of interpretable methods (Figure 2)—self-explanatory models and external co-explanations—and provides operational mathematical characterizations with concise derivations. This dichotomy mirrors common engineering practice: self-explanatory models embed interpretability at training time via structural constraints and inductive biases, whereas external co-explanations analyze already-trained black boxes to extract human-readable evidence. For each family, we relate the techniques back to the PDR desiderata and indicate typical use cases.

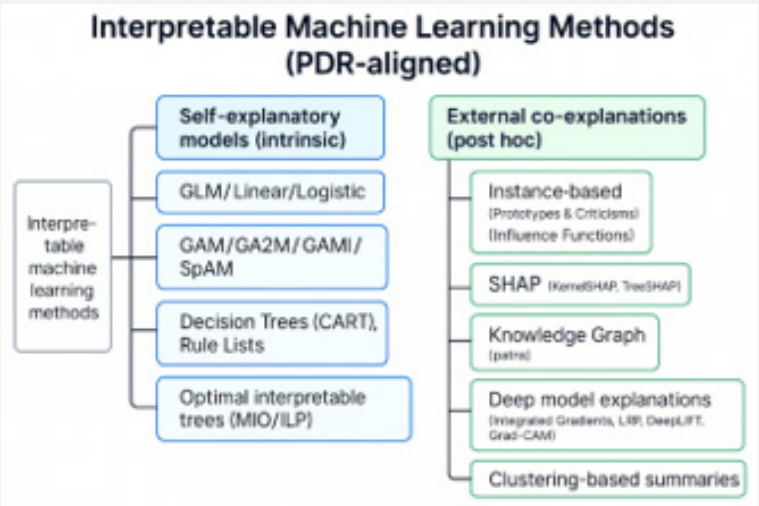

Figure 2: Overview of the method families of XAI.

Self-explanatory models

These approaches endow the predictor with built-in interpretability during model construction, typically yielding high descriptive accuracy and, under suitable conditions, competitive predictive performance. Representative classes include linear and generalized linear models, additive models, and decision trees and rule systems.

Linear/Logistic regression and generalized linear models (GLMs)

Model form and interpretive semantics.

How to cite this article: Jiangshan Z. Interpretable Machine Learning: Theory, Methods, and Applications. Ann Soc Sci Manage Stud. 2026; 13(2): 555864. DOI: 10.19080/ASM.2026.13.555864

Linear regression:

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon, \mathrm{IE}[\varepsilon] = 0.$$

Each coefficient β_j is a directly interpretable marginal linear effect of feature x_j; sign and magnitude carry semantic meaning.

Logistic regression (a GLM with logit link):

$$\Pr(y = 1 \mid x) = \sigma(\eta), \eta = \beta_0 + \sum_j \beta_j x_j \sigma(t) = \frac{1}{1 + e^{-t}}.$$

Here β_j encodes the linear contribution to the log-odds. Monotonicity constraints and sparsity penalties $(e.g., \ell_1)$ can enhance readability while controlling variance.

PDR view. When nonlinearities are weak or features have been engineered appropriately, GLMs can balance predictive and descriptive accuracy while remaining highly relevant to analyst and domain-expert audiences.

Additive models and sparse additive models (GAM/GA2M/GAMI/SpAM)

Generic GAM:

$$g\{\mathrm{IE}(y = 1 \mid x)\} = \beta_0 + \sum_{j=1}^{p} f_j(x_j).$$

with link function $g(\cdot)$ and univariate smooth components f_j. Each f_j produces an interpretable partial-effect curve.

Pairwise interactions (GA2M/GAMI):

$$g\{\mathrm{IE}(y|x)\} = \beta_0 + \sum_j f_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k).$$

usually with sparsity or heredity constraints so that only a small number of salient interactions are retained.

Sparse additive models (SpAM). With a basis expansion $f_j(x) = \sum_m \theta_{jm} \phi_{jm}(x)$, estimate

$$\min_{\{\theta_j\}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, \sum_j f_j(x_{ij})\right) + \lambda \sum_j \|\theta_j\|_2.$$

so that only a limited set of function components is selected, trading parsimony for accuracy. PDR view. Additive models provide function-level explanations (high descriptiveness) and, with appropriate capacity control or boosted variants, can retain strong predictive performance.

Decision trees and rule learning

CART backbone.

Trees are built by recursive partitioning to minimize within-node impurity, followed by cost-complexity pruning:

$$\min_T \sum_{t \in T} \sum_{i \in T} \ell(y_i, \hat{y}_t) + \alpha|T|.$$

where $|T|$ is the number of leaves. Paths from root to leaves yield human-readable if–then rules.

Bayesian/MDL rule lists.

An ordered rule list $\mathcal{R} = \{r_1 \succ r_2 \succ \cdots\}$ is learned by balancing fit and simplicity:

$$\max_R \log p(y \mid x, R) + \log p(R) \ or \ \min_R MDL(R).$$

producing compact, semantically clear rule sets.

Optimal interpretable trees (MIO/ILP). Mixed-integer optimization can directly search for trees that meet depth/leaf constraints while satisfying accuracy targets, yielding globally interpretable structures.

PDR view. Trees and rules offer strong descriptiveness and audience relevance. In highly nonlinear, high-dimensional regimes, depth control and regularization are essential to avoid undue losses in predictive accuracy.

External co-explanatory models

These methods operate post hoc on an already trained black box, aiming to improve descriptive accuracy and stakeholder relevance without modifying the predictor. Common branches include instance-based explanations, SHAP/game-theoretic attributions, knowledge-graph-based explanations, gradient/propagation-based analyses for deep networks, and clustering-based summaries.

Instance-based explanations: prototypes/criticisms and influence functions

MMD-Critic (prototypes and criticisms). Using the maximum mean discrepancy (MMD) in an RKHS H with feature map φ:

$$MMD^2(X,P) = \left\| \frac{1}{|X|} \sum_{x \in X} \phi(x) - \frac{1}{|P|} \sum_{p \in P} \phi(p) \right\|_H^2$$

select a small set of prototypes P to approximate the data and choose criticisms as points that maximally deviate from the prototype summary. Prototypes convey common patterns; criticisms expose boundary and atypical behavior.

Influence functions. Let $R_n(\theta) = \frac{1}{n} \sum_i \ell(Z_i, \theta)$, and $\hat{\theta} = \arg\min_\theta R_n(\theta)$. A small upweighting ε of a training point z yields:

How to cite this article: Jiangshan Z. Interpretable Machine Learning: Theory, Methods, and Applications. Ann Soc Sci Manage Stud. 2026; 13(2): 555864. DOI: 10.19080/ASM.2026.13.555864

$$\frac{\partial \hat{\theta}_{3}}{\partial \varepsilon}\Big|_{\varepsilon=0} = -H_{\hat{\theta}}^{-1}\nabla \ell\left(z,\hat{\theta}\right),\ H_{\hat{\theta}} = \nabla_{\theta}^{2}R_{n}\left(\hat{\theta}\right).$$

so the effect of a training point on a test loss is approximated by:

$$I\left(Z_{train},\ Z_{test}\right) \approx -\nabla_{\theta}\ell\left(Z_{test},\hat{\theta}\right)^{T}\ H_{\theta}^{-1}\nabla_{\theta}\ell\left(Z_{train},\hat{\theta}\right).$$

This identifies helpful or harmful examples and links predictions to concrete data evidence.

PDR view. These instance-level tools heighten descriptiveness and often relevance by grounding explanations in representative or influential cases, while leaving predictive accuracy unchanged.

SHAP: game-theoretic additive attributions

For feature set N = {1,...,M}, the Shapley value for feature i at instance x is:

$$\phi_{i}\left(x\right) = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!\left(M-|S|-1\right)!}{M!}\left[f_{SU\{i\}}\left(x\right) - f_{S}\left(x\right)\right].$$

satisfying efficiency, symmetry, and dummy axioms and thus yielding a unique, fair allocation. KernelSHAP uses locally weighted linear regression, while TreeSHAP provides polynomial-time exact computation for tree ensembles.

PDR view. In settings where decisions are naturally expressed in terms of feature contributions, SHAP delivers faithful, per-instance additive explanations with strong descriptiveness and good audience relevance. Care must be taken to specify the dependence model and the choice between interventional and conditional expectations, as these affect explanatory semantics.

Knowledge-graph-based explanations

A knowledge graph consists of triples (h, r, t). Path-based reasoning exposes human-readable causal or semantic chains. A simple translational scoring function such as TransE is:

$s\left(h,r,t\right) = -\left\|h+r-t\right\|_{2}$. Personalized PageRank or attention mechanisms over relation paths can be used to surface high-scoring evidence trails that justify recommendations or predictions.

PDR view. By returning entity–relation paths with scores, KG-based explanations align well with expert mental models in many domains, enhancing relevance while preserving descriptive links to model behavior.

Gradient/propagation methods and visualizations for deep models

Integrated Gradients. Given a baseline x′ and input x, the attribution on coordinate i is:

$$IG_{i}\left(x\right) = \left(x_{i}-x_{i}^{'}\right)\int_{0}^{1}\frac{\partial F\left(x^{'}+\alpha\left(x-x^{'}\right)\right)}{\partial x_{i}}d\alpha.$$

Layer-wise relevance propagation (LRP) and DeepLIFT. Scores are redistributed layer by layer to conserve total relevance, for example:

$$R_{i} = \sum_{j}\frac{a_{i}w_{ij}}{\sum_{i'}a_{i}w_{i'j}+\in sign\left(\cdot\right)}R_{j}.$$

or via differences relative to a reference activation. Grad-CAM weights convolutional feature maps by class-specific gradients to produce discriminative localization maps.

PDR view. These techniques provide instance-level evidence heatmaps and feature attributions that improve descriptiveness and, when carefully presented, relevance for practitioners working with images, text, or signals.

Clustering-based explanations

Clusters are interpretable through prototypes (medoids) and inter-cluster distances that articulate grouping rationales. One can also learn interpretable distances or fit trees/rules to the cluster assignments to obtain concise cluster descriptions. Because low-dimensional embeddings may introduce topological distortion, uncertainty and distortion measures should accompany visualizations.

PDR view. Clustering explanations summarize structure at a cohort level, often aligning well with stakeholder needs in exploratory analysis while leaving the predictive core untouched.

Summary: connecting methods to PDR

Predictive accuracy is primarily affected by self-explanatory models, which can approach black-box performance when relationships are simple or features are well engineered. Both families can raise descriptive accuracy, with external methods (SHAP, Integrated Gradients, influence functions, KG paths) particularly suited to per-instance, evidence-level explanations. Relevance, determined by audience and task, dictates the appropriate output form and granularity—rules, function curves, evidence paths, or heatmaps—and thus serves as the primary anchor for method selection and reporting (Table 1).

How to cite this article: Jiangshan Z. Interpretable Machine Learning: Theory, Methods, and Applications. Ann Soc Sci Manage Stud. 2026; 13(2): 555864. DOI: 10.19080/ASM.2026.13.555864

Table 1: Method–PDR quick map.

Family	Typical methods	Predictive (P)	Descriptive (D)	Relevance (R)	When to prefer
Self-explanatory	GLM; GAM/GA^{2}M/GAMI/SpAM; Decision Trees; Rule Lists; Optimal Interpretable Trees (MIO/ILP)	Constrained by structure; can approach black-box performance when relationships are simple or features well engineered	High (global/function/rule-level explanations)	High (well aligned with auditors and domain experts)	When relationships are relatively simple; when end-to-end interpretability, auditability, and compliance are required
External co-explanations	Prototypes/Criticisms & Influence Functions; SHAP (KernelSHAP, TreeSHAP); Knowledge-Graph Path Explanations; Integrated Gradients / LRP / DeepLIFT / Grad-CAM; Clustering Summaries	Does not change the predictor	Medium–High (mostly local/evidence-level)	Medium–High (effective for user communication)	When high-performance black boxes are needed; when instance-level evidence or path-based justification is required

XAI applications across disciplinary

This chapter examine the role and limits of XAI across disciplines. We cover environmental science, education, social media, cybersecurity, finance, law, agriculture, and healthcare using a four-part template—task and data modality, model family, explanation type.

XAI in Environmental Science

Explainable AI is rapidly being adopted across domains with clear problem-driven value. Chen and Mason [15,16] used XAI to derive a domain lexicon for classifying job postings into green versus conventional energy; combining algorithmic explanations with expert curation produced a precise, fine-grained vocabulary that reveals differences in job types, occupational structure, and geographic distribution, thereby informing workforce transitions toward green energy.

In energy and environmental science, the ubiquity of complex models makes transparency imperative. A survey of power systems notes that critical tasks—grid stability and load forecasting—are dominated by black-box models, motivating model-agnostic and standardized interpretability tools. In climate science, [17] proposes an evaluation framework spanning robustness, faithfulness, and complexity to guide method selection and strengthen credibility. Complementarily, [18] reproduces European wildfire occurrence with XAI and applies SHAP to identify key environmental drivers—such as land-surface temperature and solar radiation—with regional heterogeneity, providing mechanistic insight for effective prevention strategies.

XAI in Education

As AI systems become more intelligible to students, instructors, and administrators, XAI is emerging as a key trend in education. Explainable Artificial Intelligence in Education [19] introduces the XAI-ED framework and, through cases such as RiPPLE, FUMA, and AcaWriter, demonstrates explainable feedback and recommendation built on NLP, clustering, and rule learning. The authors note that while deep models excel at prediction, post-hoc tools—feature relevance scoring and surrogate models—can substantially increase transparency with minimal added complexity [19]. A separate study charts the rise of XAI within learning management systems (LMS) and educational data mining (EDM), emphasizing the interpretability needs of both learners and teachers, offering a taxonomy and design guidelines, and cautioning against potential misuse [20]. Additional work highlights XAI's role in personalization: LIME, SHAP, and attention mechanisms make AI decisions clearer and more engaging for students [21]. Moreover, Belief Rule Base (BRB) and SP-LIME have been used to identify factors shaping engineering students' interests, with PCA-based clustering by interest levels informing curriculum design [22]. Taken together, these studies underscore XAI's promise in enhancing the interpretability of educational tools, supporting informed decision-making, and boosting learner engagement [23-34].

XAI in Cybersecurity

XAI is improving transparency and auditability across cybersecurity tasks, including malware, intrusion, and anomaly detection. Zhang et al. [35] apply SHAP to reveal feature

How to cite this article: Jiangshan Z. Interpretable Machine Learning: Theory, Methods, and Applications. Ann Soc Sci Manage Stud. 2026; 13(2): 555864. DOI: 10.19080/ASM.2026.13.555864

contributions behind malware-classifier decisions. Hariharan et al. [35] show that permutation importance provides global explanations for intrusion detection via feature ranking. Šarčević et al. [36] use if–then decision-tree rules to produce readable, rule-based classification for complex network data. Saeed and Omlin [37] evaluate TreeSHAP in real-world malware detection; despite clear interpretability gains, usability hurdles limited analyst uptake, underscoring the need for seamless integration of XAI into security workflows. Taken together, these studies demonstrate that XAI can enhance interpretability and support decision-making in mission-critical security settings.

XAI in Finance

In high-stakes applications such as risk management, credit scoring, and fraud detection, XAI brings accountability and transparency to complex models. Weber, Carl, and Hinz [30] provide a systematic review of techniques including SHAP and LIME, detailing their use in portfolio and credit risk assessment. Awosika et al. [31] propose a federated learning (FL) plus XAI framework for financial institutions that preserves privacy while enabling collaborative fraud detection with interpretable outputs. Soundararajan and Shenbagaraman [32] integrate XAI with blockchain to create immutable, auditable records of model decisions in predictive modeling. Bussmann et al. [33] combine XGBoost with Shapley values in a peer-to-peer lending context to produce interpretable borrower profiles and reveal key risk drivers. Collectively, these studies indicate that XAI can improve transparency, strengthen trust, and support regulatory compliance, advancing responsible, data-driven decision-making in finance.

XAI in Law

XAI has emerged as a pivotal line of work in law, aimed at meeting demands for transparency, accountability, and fairness when AI influences judicial and legal decisions. As AI is integrated into contract analysis, evidence appraisal, and decision support, the need for interpretable and comprehensible models becomes acute. By exposing reasoning chains behind model outputs, XAI helps stakeholders build trust and achieve fairer outcomes. Hacker et al. [38] argue that making decision processes transparent improves legal defensibility and responsibility in complex settings such as contract and tort law. Richmond et al. [39] emphasize that XAI enables scrutiny of the reliability and relevance of AI-generated evidence, strengthening procedural justice in high-stakes decisions. Vale et al. [40] demonstrate the adaptability of post-hoc techniques (e.g., LIME), which deliver localized insights for complex legal cases. Overall, integrating XAI into legal practice can align AI applications with core legal principles, preserving transparency, fairness, and accountability while ensuring ethical and lawful deployment.

XAI in Agriculture

In agriculture, XAI is increasingly pivotal: it supports complex trade-offs under resource constraints to maximize yields and advances sustainable practices. By turning model recommendations into actionable explanations, XAI is well suited to highly variable settings shaped by seasonality and soil conditions. Study [29] shows that XAI improves the trustworthiness and usability of crop recommendation systems, enabling farmers to interpret predictions in light of their own needs. Likewise, [29] demonstrates that interpretable models such as decision trees clarify how soil quality, climate, and related factors influence crop selection—building trust and enabling location-specific decisions. Overall, XAI is powering precision agriculture and sustainability, helping farmers make clearer, more confident data-driven choices while preserving practicality and accountability.

XAI in Healthcare

In healthcare, XAI has become integral to trustworthy AI by exposing the reasoning behind complex models and reinforcing clinician confidence. SHAP and LIME are widely adopted with distinct trade-offs: SHAP provides fine-grained, consistent, per-instance attributions but incurs higher computational cost, whereas LIME is model-agnostic and broadly applicable yet may be unstable in high-dimensional settings [23,26,30]. In medical imaging, CAM and Grad-CAM highlight salient anatomical regions to aid tumor and lesion recognition, but they are largely confined to CNNs and can struggle with precise localization [25,29]. Federated transfer learning (FTL) enables privacy-preserving, multi-institutional training and safeguards patient data, though cross-site heterogeneity and consistency remain challenging [23]. Interpretable models such as decision trees and Bayesian networks offer transparent decision paths valuable for precision medicine but often face scalability limits on large, complex datasets [27]. Counterfactual explanations support causal reasoning via "what-if" scenarios, yet implementation is difficult with multimodal, highly correlated clinical data [24,29]. Attention and saliency methods can surface diagnostic cues but risk oversimplifying complex dependencies [28,43]. Overall, medical XAI must balance interpretability with predictive performance; method selection should account for clinical relevance, data complexity, and computational feasibility to ensure safe and effective integration [25,29].

Similarly, Okada, Ning, and Ong review XAI in emergency medicine, highlighting gains in triage efficiency and decision support. SHAP and LIME clarify feature importance and improve transparency, providing rationale for formerly opaque computations and thereby strengthening physician trust [42]. The authors also note the demands of real-time explainability in high-pressure contexts (e.g., resuscitation), the cognitive gap between clinicians and developers, and the need for standardized

How to cite this article: Jiangshan Z. Interpretable Machine Learning: Theory, Methods, and Applications. Ann Soc Sci Manage Stud. 2026; 13(2): 555864. DOI: 10.19080/ASM.2026.13.555864

evaluation metrics and multidisciplinary collaboration to assure usability and safety at the emergency front line [41,43].

Conclusion and Future work

Conclusion

Guided by the PDR framework, this survey turns interpretability from a slogan into an operational triad for evaluation and method selection. PDR emphasizes that explanations must remain faithful to model behavior, useful to specific human audiences, and compatible with baseline predictive performance. Placing intrinsic (self-explanatory) and post-hoc external co-explanations on the same map clarifies their complementary strengths and limits from an engineering and audience-centric perspective. Across Chapters 2–4 we move from concepts and evaluation norms to mathematical characterizations of the two method families, and then to cross-domain evidence in healthcare, finance, law, education, agriculture, cybersecurity, and environmental science. Together these results recast explanations as design artifacts constrained jointly by task goals, model behavior, and audience needs, while demonstrating—in multiple application domains—that XAI contributes not only to trust and compliance but also to discovery and action (e.g., feature engineering, counterfactuals, evidence paths).

At the application layer, our synthesis highlights domain-specific priorities and pain points: auditability and compliance in finance; legal defensibility and procedural fairness in law ; location-aware, actionable recommendations in agriculture; and robust, clinically relevant explanations in medicine. These observations reinforce a core PDR insight: the human-audience dimension determines the appropriate form and granularity of explanations and should anchor method choice and reporting practice.

Limitations and Future work

Although the PDR framework supplies a coherent semantics and evaluative axis for XAI, progress remains limited by several structural factors: relevance exhibits audience- and data-generating–process–dependent semantic drift, with few task-agnostic metrics and benchmarks; most explanations are correlational and not aligned with interventional/counterfactual semantics, inviting unwarranted causal extrapolation; explanations are brittle to perturbations, model choices, and distribution shift; deployment is constrained by HCI burdens, usability, and computational cost; and tensions persist between privacy/compliance and auditability. Addressing these gaps requires coordinated advances in methods, evaluation, and governance: establish PDR-aligned benchmarks and reporting standards that jointly document evidence and costs for predictive, descriptive, and relevance desiderata; align feature-attribution techniques with structural causal models to enable actionable, recourse-oriented explanations; develop stability, robustness, and uncertainty estimation for explanations with appropriate calibration, and define/monitor "explanation drift" for life-cycle management (including federated recalibration); for multimodal and large models, move toward concept- and evidence-level explanations to curb surface plausibility; conduct human-subject and field studies with task-level endpoints to quantify causal effects on trust and fairness; design low-cost explanations via sampling, distillation, and caching under compute/carbon constraints; and close the loop from explanation to intervention to evaluation to drive data acquisition, feature engineering, and model improvement, supported by reproducible tooling, open resources, and minimal compliance checklists.

References

1. Goodfellow I, Bengio Y, & Courville A (2016) Deep learning. MIT Press eBooks.
2. Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, et al. (2023) Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion 99: 101805.
3. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, et al. (2019) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision 128(2): 336-359.
4. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, & Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences 116(44): 22071-22080.
5. Mi J, Li A, & Zhou L (2020) Review Study of interpretation methods for future interpretable Machine Learning. IEEE Access, 8: 191969-191985.
6. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1(5): 206-215.
7. Sagi O, & Rokach L (2020) Explainable decision forest: Transforming a decision forest into an interpretable tree. Information Fusion 61: 124-138.
8. Oviedo F, Ferres JL, Buonassisi T, & Butler KT (2022) Interpretable and explainable machine learning for materials science and chemistry. Accounts of Materials Research 3(6): 597-607.
9. Gunning D, Vorm E, Wang JY, & Turek M (2021) DARPA's explainableAI(XAI) program: A retrospective. Applied AI Letters 2(4).
10. Wang X, Wang D, Xu C, He X, Cao Y, et al. (2019) Explainable Reasoning over Knowledge Graphs for Recommendation. Proceedings of the AAAI Conference on Artificial Intelligence 33(01): 5329-5336.
11. Adadi A, Berrada M (2018) Peeking Inside the Black-Box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6: 52138-52160.
12. Roscher R, Bohn B, Duarte MF, & Garcke J (2020) Explainable machine learning for scientific insights and discoveries. IEEE Access, 8: 42200-42216.
13. Mangalathu S, Karthikeyan K, Feng D, & Jeon J (2021) Machine-learning interpretability techniques for seismic performance assessment of infrastructure systems. Engineering Structures 250: 112883.
14. Miller T (2018) Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267: 1-38.

How to cite this article: Jiangshan Z. Interpretable Machine Learning: Theory, Methods, and Applications. Ann Soc Sci Manage Stud. 2026; 13(2): 555864. DOI: 10.19080/ASM.2026.13.555864

15. Chen H and Mason CM (2024) Explainable AI (XAI) for constructing a lexicon for classifying green energy jobs: A comparative analysis of occupation, industry and location composition with traditional energy jobs. IEEE Access 12: 142709-142720.

16. Machlev R, Heistrene L, Perl M, Levy KY, Belikov J, et al. (2022) Explainable artificial intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. Energy AI 9: 100169.

17. Bommer PL, Kretschmer M, Hedström A, Bareeva D, and Höhne MMC (2024) Finding the right XAI method—A guide for the evaluation and ranking of explainable AI methods in climate science. Artif Intell Earth Syst 3(3): e230074.

18. Li H, Vulova S, Rocha AD, and Kleinschmit B (2024) Spatio-temporal feature attribution of European summer wildfires with explainable artificial intelligence (XAI). Sci Total Environ 916: 170330.

19. Khosravi H, Shum SB, Chen G, Conati C, Tsai YS, et al. (2022) Explainable artificial intelligence in education. Comput Educ: Artif Intell 3: 100074.

20. Rachha A and Seyam M (2023) Explainable AI in education: Current trends, challenges, and opportunities. Proc Southeast Con Apr, pp. 232-239.

21. Shiva K, Etikani P, Bhaskar VVSR, Nuguri S, Dave A (2024) Explainable AI for personalized learning: Improving student outcomes. Int J Multidisciplinary Innov Res Methodol 2068(2): 198-207.

22. Ghosh S, Kamal MS, Chowdhury L, Neogi B, Dey N, et al. (2023) Explainable AI to understand study interest of engineering students. Educ Inf Technol 29(4): 4657-4672.

23. Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, et al. (2022) Explainable AI for healthcare 5.0: Opportunities and challenges. IEEE Access 10: 84486-84517.

24. Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, et al. (2023) The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. Comput Biol Med 166: 107555.

25. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, et al. (2021) Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. Appl Sci 11(11): 5088.

26. Ahmed S, Kaiser MS, Hossain MS, and Andersson K (2024) A comparative analysis of LIME and SHAP interpreters with explainable ML-based diabetes predictions. IEEE Access.

27. Allen B (2024) The promise of explainable AI in digital health for precision medicine: A systematic review. J Personalized Med 14(3): 277.

28. Hulsen T (2023) Explainable artificial intelligence (XAI): Concepts and challenges in healthcare. AI 4(3): 652-666.

29. Sadeghi Z, lizadehsaniR, Akif CIFCI M, Kausar S, Rehman R, et al. (2024) A review of explainable artificial intelligence in healthcare. Comput Electr Eng 118: 109370.

30. Weber P, Carl KV, and Hinz O (2023) Applications of explainable artificial intelligence in finance—A systematic review of finance, information systems, and computer science literature. Manage Rev Quart 74(2): 867-907.

31. Awosika T, Shukla RM, and Pranggono B (2024) Transparency and privacy: The role of explainable AI and federated learning in financial fraud detection. IEEE Access 12: 64551-64560.

32. Soundararajan R and Shenbagaraman DVM (2022) Enhancing financial decision-making through explainable AI and blockchain integration: Improving transparency and trust in predictive models. Educ Admin Theory Pract 30(4): 9341-9351.

33. Bussmann N, Giudici P, Marinelli D, and Papenbrock J (2020) Explainable AI in fintech risk management. Frontiers Artif Intell 3: 26.

34. Zhang Z, Hamadi HA, Damiani E, Yeun CY, and Taher F (2022) Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEE Access, 10: 93104-93139.

35. Hariharan S, Rejimol Robinson RR, Prasad RR, Thomas C, and Balakrishnan N (2022) XAI for intrusion detection system: Comparing explanations based on global and local scope. J Comput Virol Hacking Techn 19(2): 217-239.

36. Šarčević A, Pintar D, Vranić, and Krajna A (2022) Cybersecurity knowledge extraction using XAI. Appl Sci 12(17): 8669.

37. Saeed W and Omlin C (2023) Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowl-Based Syst 263: 110273.

38. Hacker P, Krestel R, Grundmann S, and Naumann F (2020) Explainable AI under contract and tort law: Legal incentives and technical challenges. Artif Intell Law 28(4): 415-439.

39. Richmond KM, Muddamsetty SM, Gammeltoft-Hansen T, Olsen HP, and Moeslund TB (2023) Explainable AI and law: An evidential survey. Digit Soc 3(1): 815-825.

40. Vale D, El-Sharif A, and Ali M (2022) Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. AI Ethics 2(4): 815-826.

41. Sivaraman V, Bukowski LA, Levin J, Kahn JM, and Perer A (2023) Ignore, trust, or negotiate: Understanding clinician acceptance of AI-based treatment recommendations in health care. Proc CHI Conf Hum Factors Comput Syst, pp. 1-18.

42. Ghassemi M, Oakden-Rayner L, and Beam AL (2021) The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 3(11): e745-e750.

43. Okada Y, Ning Y, and Ong MEH (2023) Explainable artificial intelligence in emergency medicine: An overview. Clin Experim Emergency Med 10(4): 354-362.

How to cite this article: Jiangshan Z. Interpretable Machine Learning: Theory, Methods, and Applications. Ann Soc Sci Manage Stud. 2026; 13(2): 555864. DOI: 10.19080/ASM.2026.13.555864

DOI: 10.19080/ASM.2026.13.555864

How to cite this article: Jiangshan Z. Interpretable Machine Learning: Theory, Methods, and Applications. Ann Soc Sci Manage Stud. 2026; 13(2): 555864. DOI: 10.19080/ASM.2026.13.555864