Int. J. Appl. Math. Comput. Sci., 2026, Vol. 36, No. 2, 235–266
DOI: 10.61822/amcs-2026-0017

A REVIEW OF EXPLAINABLE SEMI–SUPERVISED METHODS IN MULTIVARIATE TIME SERIES ANALYSIS

FILIP WICHROWSKI [a,*], MARCIN OSTROWSKI [a], MARTA BORATYN [a], KATARZYNA KACZMAREK-MAJER [a]

[a] Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
e-mail: fwichrow@ibspan.waw.pl

Recent advances in information and communication technologies have led to the widespread use of smart meters, wearable sensors, and related devices in healthcare, industrial monitoring (e.g., manufacturing and energy systems), and transportation. These systems generate large volumes of sequential data, yet fully annotating them remains expensive and often impractical. Consequently, only a small fraction of the data is typically labeled, limiting the applicability of fully supervised learning. At the same time, ignoring available labels altogether, as in fully unsupervised approaches, risks discarding valuable information. This tension has motivated growing interest in semi-supervised methods that learn from both labeled and unlabeled data. However, many such approaches rely on complex black-box models, making the decision-making process opaque, particularly in high-risk domains such as medicine and finance. Explainable AI (XAI) has therefore become essential for building trust and ensuring accountability in these settings. This review surveys recent advances in explainable semi-supervised methods for multivariate time-series analysis. We introduce a taxonomy based on how explainability is integrated, categorizing the approaches as white-box (transparent), post hoc (opaque), and intermediate. Within each category, we further classify them by their most distinctive characteristics: the model class for white-box, the interpretability integration strategy for intermediate, and the explanation technique for post-hoc. We also discuss common practices and lessons learned in dealing with partial supervision and model interpretability, and highlight key challenges in sequential data analysis, such as the choice of performance metrics and explanation techniques. We find that, although interest in explainable semi-supervised time-series methods is growing, the systematic evaluation of explanations remains underdeveloped and lacks standardized evaluation practices. Nearly 70% of the reviewed works report some form of explainability validation; however, it is typically indirect, qualitative, or limited in scope. Overall, explainable semi-supervised methods represent a promising direction for future research, with potential benefits across a wide range of real-world time-series applications.

Keywords: semi-supervised learning, time series, explainable artificial intelligence, data stream analysis.

1. Introduction

In this review, we aim to synthesize the current landscape of explainable semi-supervised methods for multivariate time series. Given the increasing availability of sensor data and the practical difficulties of collecting expert-provided labels across domains (e.g., healthcare and engineering), semi-supervised approaches are gaining particular attention. However, many existing methods are tailored to tabular (static) data, thereby ignoring the evolving nature of time series.

To address this gap, we conduct an extensive literature review focusing on the following three aspects:

1. semi-supervised learning (partially labeled learning),
2. explainability (explainable approaches, interpretable, human-consistent, human-centric, human-in-the-loop),
3. multivariate time series (data streams, sensor data, temporal data, time-ordered data, sequential data).

We conducted the initial search using the Scopus database in April 2025, with a final update performed

*Corresponding author

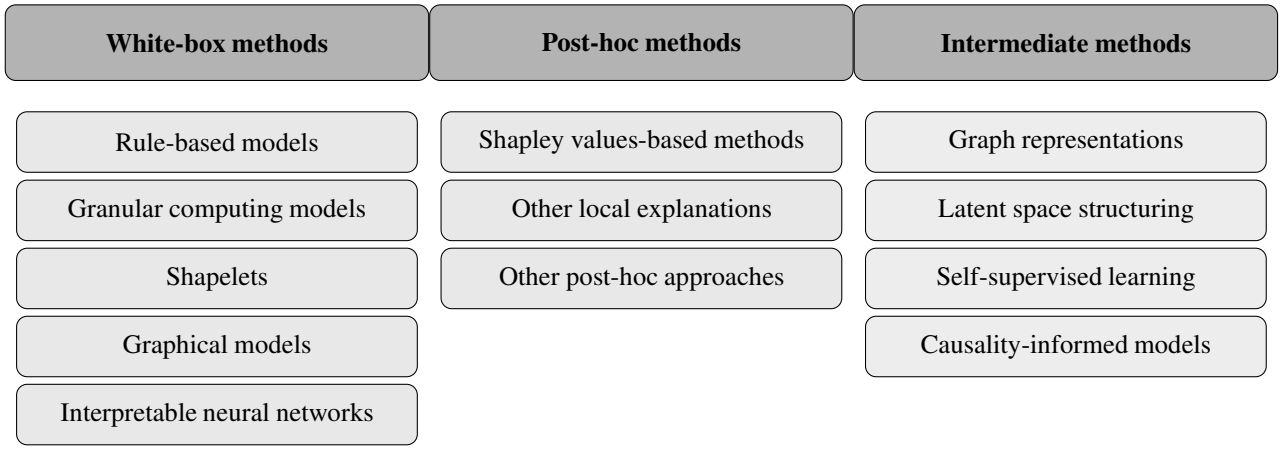

Fig. 1. Trichotomous taxonomy of semi-supervised explainable approaches for time series, categorizing methods as white-box (transparent), post-hoc (opaque), or intermediate according to how explainability is incorporated.

in January 2026. The search was limited to journal articles, book chapters, and conference proceedings published in English in the fields of computer science, mathematics, and engineering. The search yielded 123 papers, which underwent in-depth screening, and their reference lists were also examined to identify additional relevant works. At this stage, we excluded 72 papers due to misalignment with the scope of this study. In particular, we grouped the reasons for exclusion into the following categories: tailored solely for video/image data (19 papers), not designed for sequential data (13 papers), full text unavailable (3 papers), non-generalizable/highly domain specific (7 papers), no methodological novelty (4 papers), other reasons (e.g., lack of partial supervision, duplicates, unclear scope) (26 papers).

Finally, during the revision phase, we added three articles. Consequently, we included 54 papers, which we discuss in detail in this review. The flowchart is presented in Fig. 2.

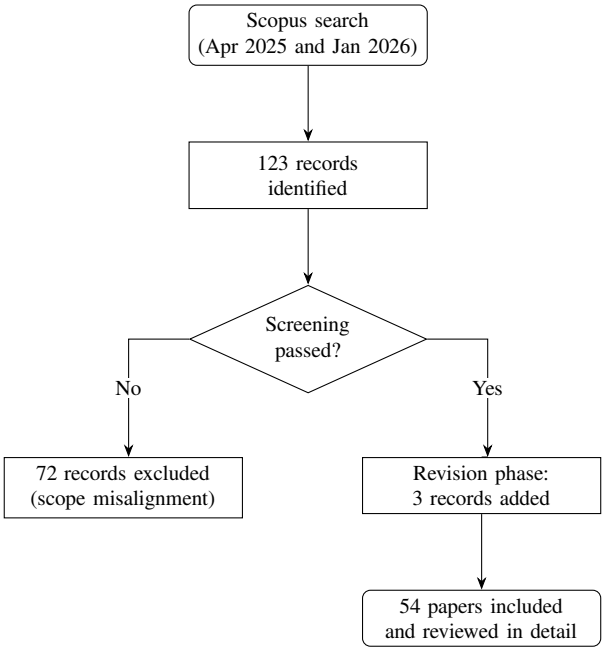

Fig. 2. Flowchart of the paper selection process for this review.

When preparing the review, we assess the quality and novelty of the proposed approaches, paying particular attention to explainability and interpretability. According to the Cambridge Dictionary of the English Language, 'explanation' is defined as *the details or reasons that someone gives to make something clear or easy to understand.* Barredo Arrieta *et al*. (2020) define explainability as the details and reasons provided by a model to make its functioning clear or easy to understand for a given audience. Closely related is the concept of interpretability, which refers to the level of understanding of how the underlying AI technology works, whereas explainability concerns understanding how an AI-based system arrived at a particular result (ISO, 2020). According to the recent ISO/IEC standard (ISO, 2025), explainability is a property of an AI system that enables a specified human audience to understand the reasons for the system's behavior, including how a particular model outcome is produced. An AI system is defined as any engineered system that generates outputs such as content, forecasts, recommendations, or decisions for human-defined objectives, including both classical models and more complex approaches such as deep neural networks (ISO, 2025).

This raises an important question: *What does explainability mean in the context of semi-supervised learning for temporal data?* While the general notion of explainability can be defined broadly, the form and scope of explanations depend on the task at hand (e.g.,

Table 1. Summary of the reviewed papers categorized into white-box (transparent) models, intermediate approaches, and post-hoc methods, further subdivided by the primary task addressed (e.g., classification, anomaly detection, representation learning).

Task	White-box	Intermediate	Post-hoc	Total
Anomaly detection	1	8	5	14
Classification	8	8	1	17
Clustering	2	1	0	3
Forecasting	0	1	0	1
Other	1	0	2	3
Regression	1	2	0	3
Representation	2	8	2	12
Segmentation	0	1	0	1
Total	15	29	10	54

classification or prediction) and on the methodological approach used (e.g., the chosen representation or model). In the following sections, we present examples of approaches that incorporate explainability or interpretability. This includes evaluating the presence of mechanisms that align with human-consistent reasoning and support human-centric understanding, such as visualizations, feature attributions, symbolic learning, and domain-relevant explanations.

We aim to provide a structured overview that summarizes existing contributions and highlights gaps and opportunities for future research at the intersection of explainability, semi-supervised learning, as well as multivariate time-series and data-stream analysis.

To the best of the authors' knowledge, no previous review has systematically examined explainability in semi-supervised time-series learning yet. Previous reviews have explored related aspects of this topic; however, a comprehensive analysis remains lacking. For example, Eldele *et al*. (2024) review strategies for time-series representation learning under scarce labeling conditions, while Zhang *et al*. (2024) provide a comprehensive survey of recent self-supervised learning methods for time series. Other surveys focus on specific model classes or problem formulations, such as few-shot learning (Li *et al*., 2025), anomaly detection (Saritha and Dhanalakshmi, 2024), and concept drift (Agrahari and Singh, 2022). We extend these works by explicitly considering explainability and interpretability in semi-supervised settings for time-series and data-stream analysis.

When discussing the explainability of the state-of-the-art works, we consider a trichotomous taxonomy (see Fig. 1) which consists of white-box (transparent) models (Section 2), post-hoc methods (Section 3) and intermediate approaches (Section 4). In the work of Ali *et al*. (2023), the intermediate category is regarded to be gray-box and defined as models that users can interpret to some degree. Within each category, we further classify methods depending on the type of model/reasoning they integrate. Table 1 summarizes the number of papers considered for each category, further divided by the main identified problem they address: classification, forecasting, anomaly detection, representation, segmentation, clustering, regression, or other. These tasks represent the primary problems tackled by the reviewed methods. We provide a detailed listing of the discussed papers, grouped by primary task and explainability type, with references in Appendix (Table A1).

Finally, in Section 5, we synthesize and discuss common practices, their advantages, limitations, and the challenges associated with handling partial supervision and model interpretability. We place particular focus on the evaluation of explainability.

2. White-box models

White-box models are designed to be inherently interpretable, with internal mechanisms and decision processes that are transparent to humans (Doshi-Velez and Kim, 2017; Rudin *et al*., 2022). Their structural design offers a key advantage: it enables direct explanations of model decisions without the need for external tools. In this review, we classify a model as being white-box when its predictive mechanism can be directly inspected and understood without resorting to external techniques, making the relationship between inputs, intermediate representations, and outputs structurally explicit (Rudin, 2018; Doshi-Velez and Kim, 2017; Lipton, 2017; Molnar *et al*., 2020). This makes white-box models particularly attractive for semi-supervised learning, as their structure often allows partially labeled data to be incorporated directly into the learning process, for example, by modifying rules, constraints, prototypes, or probabilistic parameters in response to labeled observations.

Within this category, we distinguish several modeling paradigms that realize interpretability: rule-based systems, fuzzy logic models, shapelet-based methods, graphical models, and interpretable neural architectures. Table 2 summarizes the reviewed white-box methods.

Table 2. Summary of the reviewed white-box papers. Paper identifiers follow Table A2.

ID	Partial supervision	Explainability integration/validation	Strengths/limitations
[1]	Graph-based label propagation	Symbolic representations; logic rules *Indirect validation*: predictive performance; qualitative rule examples	+ Handles irrelevant features better via improved distance metric – Runtime increases with labeled examples and logical atoms
[2]	Entropy-based selection of unlabeled data	Grammar-based description of the temporal event structure *Indirect validation*: generalization risk; accuracy	+ Effective with few labels – Heuristics; no optimality guarantees
[3]	Pseudo-labeling (self-training model)	Belief rules; domain knowledge; natural-language explanations; counterfactuals *Direct validation*: measures (e.g., feature coverage, relevance, coherence, pragmatism)	+ Interpretable rule base; improved performance via partial supervision – Validation on a single dataset; expert-dependent
[4]	Model trained on anomaly-free data	Decision tree is used after novelty detection *Indirect validation*: sensor-location consistency	+ Hybrid pipeline; physical interpretability via sensor-location – Intentional overfitting
[5]	Online active learning	Evolving fuzzy system with a shared rule base *No validation*: argued conceptually	+ Handles multi-label correlations; online – Parameter tuning needed
[6]	Rule updating using unlabeled data	Evolving fuzzy rule-based model with Gaussian membership functions; linguistic rules *No validation*: interpretability assumed	+ Online; evolving; effective with unlabeled data – Domain-specific feature engineering; heuristic thresholds
[7]	Pseudo-labeling	Interpretable prototypes and IF–THEN fuzzy rules *No validation*: predictive performance	+ Parameter-free pseudo-labeling; streaming data and concept drift – Prototype optimality not guaranteed
[8]	Dynamic incremental SS-FCM	Fuzzy linguistic summaries from semi-supervised fuzzy clustering prototypes *Indirect validation*: expert-oriented; consistency with clinical observations; usefulness	+ Interpretability; non-stationary data streams and concept drift – Requires prior domain knowledge; small sample size
[9]	Expert decisions on prototypes	Cluster prototypes; rough set decision reducts; template-based natural language descriptions *Indirect validation:* performance, cluster stability	+ Prototype-based reasoning; balance between performance and interpretability – LLMs may introduce vagueness
[10]	Constrained K-means	Shapelets combined with decision trees *Indirect validation*: prediction accuracy; qualitative domain interpretation of shapelets	+ Early online detection – Requires large simulated offline dataset; domain-specific assumptions
[11]	Pseudo-labeling; spectral regularization	Inherently interpretable subsequences *Indirect validation*: visual inspection of learned shapelets	+ Interpretable features; efficient shapelet learning – Non-convex optimization; Euclidean distance for shapelets only
[12]	Label propagation	Inherently interpretable via shapelets and distance-to-shapelet representations *Indirect validation*: ablations; visual inspection	+ Efficient candidate selection through salient subsequences – Multivariate extension assumes variable independence
[13]	Pseudo-labeling; self-supervised diffusion	Strengthened by diffusion (to resemble real subsequences) and alignment with semantic descriptions *Indirect validation*: visualizations, t-SNE; ablations	+ Language-shapelet alignment; effective for sparse labels – Univariate; limited if labels lack semantic meaning
[14]	Graph-based posterior regularization	KL-based posterior regularization via an external similarity graph *No validation:* only predictive performance	+ Theoretically justified; scalable; closed-form solutions – multiple hyperparameters
[15]	Simulated labels (physics-inspired)	Explicit mathematical equation; additive, multiplicative neurons; weights discretization *Indirect validation*: prognostic criteria commonly used for health indicators	+ Transparent; explicit equations – Training depends on simulated degradation labels; no uncertainty modeling

2.1. Rule-based methods. A simple and transparent category of white-box models comprises rule-based approaches, which rely on explicit IF–THEN rules to map inputs to outputs. Due to their logical structure and interpretability, such models are well-suited to domains where clarity, explainability, and expert involvement are important, particularly when the resulting rule set offers a simpler representation of the decision process than that of more complex models.

For example, Michelioudakis *et al*. (2023) present an online method for learning composite event definitions from partially labeled data streams, using the event calculus formalism and an inductive logic programming framework. The proposed approach infers missing labels by combining a structural similarity measure, refined through feature selection, with a mass-based dissimilarity that reflects the distribution of examples in the data. These measures are used in a graph-based label propagation method that incorporates temporal adjacency. The inferred labels are passed to a structure learner, which incrementally induces first-order logic rules for complex events. This enables accurate and scalable learning in streaming scenarios with limited supervision.

Veeraraghavan *et al*. (2007) propose a different approach based on stochastic context-free grammars. A given event is modeled as a sequence of elementary actions, each defined by a combination of spatial location and local motion; together, these sequences serve as the basis for constructing a grammar that represents the corresponding event class. The structure and parameters of each class-specific grammar are learned from data using a small set of labeled examples and a larger pool of unlabeled ones. In an iterative process, new examples with the lowest conditional classification entropy (i.e., those most confidently classified) are incorporated to refine the grammars. Once the model is learned, classification of a new sequence is performed by attempting to parse it with each grammar and selecting the best match. The authors validate their method on video recordings of traffic intersections. It demonstrates reliable performance even with limited labeled data and outperforms spectral clustering, particularly in cases involving incomplete or partially observed trajectories.

Kabir *et al*. (2025) propose another approach based on symbolic knowledge representation, introducing a self-training framework for predicting building energy consumption. The method is built on the belief rule-based expert system (BRBES), a knowledge-driven symbolic AI model designed to produce interpretable predictions. The initial BRBES is trained on labeled data using evolutionary joint optimization of its structure and parameters, and is refined until it reaches a predefined accuracy-confidence threshold. Synthetic unlabeled data are then generated through weak and strong augmentation strategies and pseudo-labeled. The model is subsequently reoptimized with both labeled and pseudo-labeled data. Its rule-based structure makes the inference process inherently explainable. Explanations are derived from the most activated rule and presented in a human-understandable language, along with counterfactuals describing how changes to the input could affect the prediction. On real energy consumption data, the method outperforms standard semi-supervised models in both accuracy and explainability.

Rule-based systems may also adopt a data-driven character. A prominent example of this approach is decision trees (Quinlan, 1986), which Movsessian *et al*. (2020) use in a two-stage method for detecting and interpreting anomalies in vibration-based structural monitoring systems. In the first stage, features are extracted from the covariance matrix of sensor responses and reduced in dimension using principal component analysis. Novelty detection is then applied using the Mahalanobis distance, and anomalies are identified as observations whose distance exceeded a statistical threshold. In the second stage, decision trees are used as a supervised model to classify these anomalies. The hierarchical tree structure enables identification of the features and sensors responsible for each classification, allowing approximate localization of simulated damage and differentiation between damage levels.

2.2. Granular computing-based methods. A particularly promising subclass of rule-based systems comprises fuzzy rule-based ones (Zadeh, 1965). Unlike traditional binary logic systems that operate on strict true/false values, fuzzy rule-based ones allow rules to be evaluated to partial degrees of truth, making these systems especially effective in environments where input data is uncertain, vague, or imprecise.

Lughofer (2022) proposes a new evolving Takagi–Sugeno type fuzzy classifier for online multi-label classification. The model employs a shared fuzzy rule base in which each rule contains separate regression consequents for each label, learned online via weighted least squares with optional L1 regularization. This formulation enables direct estimation of continuous output values, which are then thresholded to obtain a prediction of label presence or absence (0 or 1). Rule conditions (antecedents) are represented as fuzzy ellipsoids and are created incrementally by evolving sample clustering in an expanded space containing both input features and labels. This approach enables the creation of rules specific to local regions of the feature space and accounts for the co-occurrence of specific labels, which is crucial in multi-label problems. In addition, the authors introduce local label correlation learning; rule weights are adjusted to reflect interdependencies among predicted labels, thereby improving prediction accuracy when labels are

not independent. Another important innovation is the inclusion of an online active learning module, which allows selective labeling of only those samples that are new relative to existing rules (outside all ellipsoids), uncertain (generate predictions with low confidence, close to the threshold), and statistically informative. For validation, the authors used five MULAN datasets. The method outperforms evolving one-versus-rest and classifier chaining baselines in terms of partial accuracy and average precision, while maintaining strong performance even with 90% fewer labeled samples due to the active learning strategy.

Leite *et al*. (2020) propose an evolving fuzzy classifier based on fuzzy decision rules that form its knowledge base. Each rule describes the relationship between input values and the class assigned to them. Such a rule can be exemplified as follows: *If the input features have certain properties, then the object belongs to the class*. These conditions are defined by Gaussian membership functions, each describing one rule condition and determining how strongly a given input value fits into a fuzzy set. The model operates in a streaming environment and handles partially labeled data. A labeled sample creates a new rule or updates an existing one with the corresponding class. Unlabeled samples update rule parameters if they sufficiently activate a rule. If a labeled sample later activates an unlabeled rule, it is assigned the appropriate class. Validation is conducted on synthetically generated voltage waveforms simulating five disturbance classes under IEEE standards, with varying noise levels (20–60 dB) and window lengths (1, 4, 10 cycles). The model is tested in both fully supervised and semi-supervised scenarios, showing high accuracy and robustness to noise and unlabeled data.

Gu (2022) also considers data streams and proposes S^3OFIS+, a semi-supervised extension of the SOFIS+ model for streaming data with indefinite label delays. The method uses a C-nearest prototypes pseudo-labeling strategy, assigning pseudo-labels only when a majority of the C closest prototypes (one per class) agree on the label. First, the system is trained on a small labeled dataset in chunks (priming) to create class prototypes and IF–THEN fuzzy rules. After priming, unlabeled data are processed sequentially. For each sample in a new chunk, distances to existing prototypes are computed and the C closest prototypes *vote* for their classes. Finally, the system retrains using only pseudotagged samples that satisfy the confidence criterion. As in the priming stage, it identifies new prototypes representing emerging patterns and updates the fuzzy rule base. Repeating this cycle for each data chunk allows the system to evolve over time without requiring additional labeled data. The method is validated on 20 benchmark datasets, including stationary, nonstationary, and image data. It matches or outperforms state-of-the-art semi-supervised methods in accuracy and efficiency and shows strong robustness to noise, high dimensionality, and concept drift.

Another approach employs fuzzy logic combined with linguistic summarization to provide interpretable descriptions, as proposed by Kaczmarek-Majer *et al*. (2022). Their method builds on the approach introduced by Casalino *et al*. (2019), which combines dynamic, incremental, semi-supervised fuzzy clustering with linguistic summarization. Clustering is performed iteratively on consecutive data chunks, and the resulting clusters are mapped to classes represented by prototypes. These prototypes are then used to construct fuzzy membership functions corresponding to linguistic expressions, enabling the generation of linguistic summaries. The main contribution is the use of evolving cluster prototypes to improve the generation of linguistic summaries. The proposed method is validated for a particular applied scenario, that is, sensor-based remote mental health monitoring of bipolar disorder patients.

Another example of granular computing for reasoning with imperfect time series is presented by Grzegorowski *et al*. (2025) in the sales domain. The authors propose prototype-based clustering for segments with aggregated historical sales and the use of rough set theory to explain the clusters obtained from numerical time-series representations. Similar historical sales patterns form granules, and the cluster prototypes are then considered for decision-making. Experiments show improved performance, even though individual time series are not modeled directly. Instead, granulation enables variance reduction rather than fitting unstable models to individual series. Experiments confirmed that this granular computing approach is particularly helpful for imperfect time series (short, sparse, noisy, seasonal). Unlike many approaches that rely on feature importance rankings, the paper considers minimal feature subsets derived from rough sets, leading to more structural rather than heuristic interpretations.

2.3. Shapelets. An important direction in explainable time-series analysis emerged with the introduction of shapelets (Ye and Keogh, 2009), i.e., interpretable and discriminative subsequences used for classification, which offer both strong predictive performance and interpretability. However, because classical shapelet discovery methods rely on supervised learning, they typically require large amounts of labeled data to identify informative subsequences. This limits their applicability in real-world scenarios where annotations are scarce or costly. Consequently, recent work has focused on adapting shapelet-based approaches to semi-supervised and unsupervised settings in order to leverage unlabeled data, which is often more readily available.

Zhu *et al*. (2016) present one of the earliest attempts to leverage semi-supervised learning for shapelet

discovery, addressing the lack of fully labeled data in online short-term voltage stability assessment for power systems. A subset of observations that can be confidently labeled using domain knowledge is used to define must-link and cannot-link constraints, guiding a constrained K-means clustering process that assigns labels to the remaining data. To improve the efficiency of shapelet discovery, two strategies are employed. One is sampling search, in which shapelets are extracted from randomly selected subsets of the dataset to reduce computational cost. The other is adaptive variable-step search, in which the shapelet window advances across segments defined by a piecewise linear approximation, rather than moving point by point. Distances between each selected shapelet and the full set of time-series samples are computed and used to transform the data into a conventional feature space. A decision tree classifier is then applied to this representation, offering both accurate classification and interpretability. Validation on the real data demonstrates both high predictive accuracy and computational efficiency, confirming the method's suitability for real-time voltage stability assessment. While Zhu *et al*. (2016) employ partial supervision as a preprocessing step to generate labels, Wang *et al*. (2019) introduce a more integrated method, semi-supervised shapelets learning (SSSL), in which unlabeled data are incorporated through pseudo-labels that are iteratively updated. The approach formulates a joint optimization problem with an objective function that combines a regularized least-squares loss for labeled and unlabeled data, a spectral regularization term to preserve the local structure among unlabeled series, and a shapelet similarity regularization term to discourage redundant or highly similar shapelets. Shapelets, pseudo-labels, and classifier weights are learned and optimized using a coordinate descent algorithm, as the optimization problem is non-convex. Validation on 13 real-world UCR datasets indicates strong performance and superiority over other state-of-the-art methods for time-series classification.

Cai *et al*. (2023) also explore improved methods for shapelet discovery. The authors propose SE-shapelets, a method to address the prevalence of uninformative subsequences in time-series data. It leverages a small number of labeled time series and propagates pseudo-labels to nearby unlabeled instances, thereby guiding the discovery of representative shapelets. A salient subsequence chain (SSC) is used to extract salient subsequences, defined as those that differ most from their neighbors according to Euclidean distance, thereby reducing the candidate pool. Then, K-means clustering is applied to group similar candidates, with cluster centers forming the final candidate shapelets. Further, a linear discriminant selection (LDS) procedure selects the most representative shapelets by integrating a binary selection matrix with linear discriminant analysis to identify shapelets that best separate time series from different classes for clustering. The authors' validation includes 85 datasets from the UCR archive (Dau *et al*., 2018), on which their method outperforms constraint-based, label-based, and unsupervised shapelet-based clustering methods, achieving the highest average Rand index and rank. Ablation studies demonstrate that both the SSC and LDS have a positive impact on clustering performance.

A method more grounded in deep learning is DiffShape (Liu *et al*., 2024). The first core component is a self-supervised diffusion learning mechanism in which shapelets are learned from time-series data using a convolutional layer. The most similar real subsequences are then identified and used as conditions in a diffusion model to guide the generation of new shapelets that resemble real patterns. Second, a classifier trained on labeled data assigns pseudo-labels to the unlabeled data, enabling the construction of natural language descriptions for those samples. In the contrastive language-shapelet learning stage, both real and generated shapelets are embedded and aligned with their corresponding text embeddings via a contrastive loss, thereby enhancing their discriminability. The authors perform a thorough validation on 106 datasets from the UCR time-series archive, comparing DiffShape against various semi-supervised classifiers and showing that it outperforms the baselines in both classification accuracy and interpretability, particularly in low-label regimes.

2.4. Graphical models. Another subclass comprises graphical models, such as Bayesian networks, where the conditional dependence structure among random variables is represented by a graph. These models can incorporate prior or domain knowledge by explicitly specifying relationships among variables through the graph structure. Their interpretability largely depends on the graph's complexity and sparsity, with simpler and sparser graphs generally promoting interpretability, robustness, and computational efficiency in high-dimensional graphical models (Friedman *et al*., 2008).

Motivated by these advantages, Vinzamuri *et al*. (2020) propose an anomaly detection framework for sensor data that leverages sparse Gaussian graphical models (SGGMs) to capture only the most informative conditional dependencies among variables. The method employs partial supervision by leveraging known failure points to segment the time series into normal and pre-failure regimes. Separate graphical models are then learned for each regime, and statistical distance measures, such as the Kullback–Leibler divergence and the Frobenius norm, are used to quantify differences between them, where large deviations may indicate anomalies. A key contribution is the incorporation of domain semantics, represented through semantic graphs, into model construction to improve interpretability and

predictive performance. Validation includes benchmarks on several public tabular datasets, where the method either outperforms or closely competes with state-of-the-art baselines. A case study demonstrates high precision and recall across all turbines, as well as improved anomaly separation after incorporating semantic knowledge. The system also leverages modules such as IBM OpenScale to provide contrastive explanations, supporting diagnostics for reliability engineers. However, their transparency is not fully confirmed in the paper.

Libbrecht *et al*. (2015) introduce the entropic graph-based posterior regularization (EGPR) method, a graph-based posterior regularization technique that encourages similar data points to have similar posterior label distributions in probabilistic models. Notably, EGPR does not introduce a new probabilistic model per se; rather, it acts as a regularization mechanism that can be integrated into existing models, potentially improving the consistency and interpretability of their predictions. The novelty lies in applying posterior regularization based on KL divergence, using an arbitrary regularization graph rather than directly modifying the model's parameters. Inference and learning are performed with a three-way alternating optimization (with closed-form updates) that involves smoothing posteriors over the graph, performing standard inference, and updating parameters. Validation on synthetic and genomic data shows that incorporation of EGPR improves clustering accuracy in simulated settings and reduces the root mean square error (RMSE) in genome annotation compared to baseline models.

Although graphical models are inherently transparent and appealing for illustrating complex relationships and semantics, in real-world semi-supervised time-series contexts they are often insufficient on their own to address the full problem. At the same time, approaches that use graphical representations alongside other learning mechanisms have proven effective. We will cover these aspects under the category of intermediate explainable approaches and discuss them in more detail in Section 4.1.

2.5. Interpretable neural networks. Although deep neural networks have dominated recent developments in machine learning, shallow architectures still appear in contexts where interpretability is important. Moradi *et al*. (2024) propose an interpretable neural network to construct health indicators for aircraft engines, aiding in remaining useful life prediction. They evaluate it using monotonicity, prognosability, and trendability. Unlike conventional deep learning models, which often lack physical interpretability, the proposed architecture includes additive and multiplicative layers to address this issue. The integration of these layers within the model facilitates combining system characteristics through both multiplicative and divisive processes. Consequently, the model is designed to capture the system's properties and behaviors more accurately. To enhance interpretability, the model uses a ternary weight discretization method that converts continuous weights into three discrete categories. This approach reduces model complexity by lowering parameter count and simplifying equations. The study employs a semi-supervised learning approach, using simulated labels inspired by the physics of progressive damage. The approach enables the implicit incorporation of non-differentiable criteria into the learning process. The methodology is tailored to commercial turbofan engines, producing interpretable health indicators while balancing model complexity and explainability.

2.6. Summary. Table 3 provides a qualitative comparison of white-box model families discussed in this section, highlighting their core design principles, explainability strengths, typical limitations, and mechanisms for addressing partial supervision. White-box models offer a clear and trustworthy alternative to complex black-box approaches; however, this benefit often comes at a cost, as they may exhibit lower predictive performance than black-box models (Vernon *et al*., 2024; Fernandez-Delgado *et al*., 2014; Krizhevsky *et al*., 2017). Yet the extent to which this trade-off holds universally remains an open question (Bell *et al*., 2022). Recent studies suggest that the trade-off between accuracy and explainability is less pronounced and more nuanced than previously assumed (Kruschel *et al*., 2024; Bell *et al*., 2022).

White-box models are inherently interpretable and do not require external explanation tools that may reduce clarity or be unreliable in some contexts. This property makes them particularly suitable for high-stakes applications where trust, accountability, and human oversight are essential. However, even interpretable models can become opaque when embedded in complex preprocessing pipelines or frameworks that involve extensive data transformations, obscuring the original meaning of input variables.

3. Post-hoc methods

Designing interpretable models often involves imposing structural constraints (e.g., sparsity or monotonicity), which can increase modeling complexity and make optimization more challenging (Rudin, 2018). As a result, many machine learning models are not designed to be inherently explainable and provide only limited insight into the mechanisms underlying their predictions. This is particularly common for complex black-box models such as deep neural networks or ensemble methods. In such cases, external techniques are required to obtain insight into the model's decision-making mechanisms (Doshi-Velez and Kim, 2017). These techniques are

Table 3. Comparison of white-box model families discussed in Section 2, showing their core ideas, strengths, limitations, and how partial supervision is employed.

Model family	Core idea	Strengths (XAI)	Typical limitations	Partial supervision
Rule-based methods	Explicit IF-THEN logic	Fully transparent; human-readable rules	Rule explosion; limited scalability; vulnerable to noise	Entropy-based selection; label propagation; pseudo-labeling
Granular computing-based methods	Soft rules; soft clustering	Handles uncertainty; linguistic explanations	Rule-base growth; parameter tuning	Pseudo-labeling; active learning; evolving rules, constraint-based clustering
Shapelets	Discriminative subsequences	Local; time-aware explanations	High computational cost	Constraint-based clustering; label propagation, pseudo-labeling
Graphical models	Probabilistic dependency graphs	Causal and relational interpretability	Scalability; structure learning complexity	Co-training; failure-guided partitioning
Interpretable neural networks	Shallow or constrained architectures	Interpretable weights and equations	Limited complexity	Simulated labels

referred to as post-hoc methods; they are typically not integrated into the model's training process, but instead analyze a trained model's outputs to infer which features influence predictions, either locally or globally.

Post-hoc explainability methods have become a common practical approach for interpreting complex models. Among these techniques, feature-attribution methods are particularly widespread in the literature. In this section, we categorize the reviewed articles according to the type of post-hoc explanation technique employed, distinguishing among SHAP, LIME, and other dedicated approaches. A summary of the analyzed post-hoc methods, together with their main characteristics, is provided in Table 4.

3.1. Shapley values. One of the most widely used post-hoc interpretability approaches is the SHAP (Shapley additive explanations) framework (Lundberg and Lee, 2017), which is based on cooperative game theory. SHAP assigns each variable a value representing its contribution to the model's prediction. The method relies on Shapley values (Shapley, 1953), which quantify each feature's contribution by averaging its marginal contribution over all possible feature subsets. SHAP is widely used as a post-hoc explainer across a range of practical domains and frequently appears in studies comparing post-hoc explainability methods (Givisis *et al*., 2025; Noah and Pum, 2024; Salih *et al*., 2025; Nguyen *et al*., 2021).

Rao and Wang (2024) propose an explainable semi-supervised learning framework for fault detection and diagnosis in chemical processes. The method integrates three components: a backbone network that combines multiple pattern representation, multi-head self-attention, and temporal convolutional networks with dense connections (DTFCN), a semi-supervised learning framework based on MixMatch that uses consistency regularization and entropy minimization, and a post-hoc explanation module based on SHAP. Here SHAP is applied to pattern-level representations of windowed time-series data, and the resulting attributions are aggregated across temporal and pattern dimensions to obtain per-variable relevance scores. The approach is validated on two benchmark processes under supervised, semi-supervised, and imbalanced conditions. In the supervised setting, the DTFCN achieves superior mean fault detection rates compared to other deep learning models, while the proposed method offers a competitive performance in semi-supervised and data-imbalance settings.

SHAP values are also employed as part of an integrated framework for detecting sensor faults in structural health monitoring systems. To increase interpretability, Martakis *et al*. (2022) introduce the concept of decision trajectories, a visualization obtained by interpolating Shapley attribution values that shows how successive features affect the final classification of a sample. The main idea of the presented framework is to use only data from properly functioning sensors (healthy sensors). In the first stage, the support vector machine (SVM) algorithm is used as a one-class classifier to identify data that deviates from the normal pattern. Data points that fall between the 90th and the 99th percentile thresholds are considered uncertain, while predictions above the 99th percentile threshold indicate abnormal behavior. The predictions of this model are then used as labels ("normal", "uncertain", and "abnormal") in the second stage, where the XGBoost classifier is trained. Detected anomalies with identical characteristics are expected to yield similar decision trajectories. To compare these quantitatively, the authors introduce a measure that evaluates their consistency against reference benchmarks. They argue that this modification may become a powerful tool for fault detection.

Fouzi *et al*. (2024) and Harrou *et al*. (2024) provide

Table 4. Summary of reviewed post-hoc papers. Paper identifiers follow Table A2.

ID	Partial Supervision	Explainability Integration/validation	Strengths/limitations
[16]	MixMatch-style learning; pseudo-labeling	SHAP feature attribution over process variables *Indirect validation*: agreement with known fault mechanisms	+ Strong empirical performance; broad applicability – High training cost; explanations are correlational, not causal
[17]	Self-labeling; training on anomaly-free data	Decision trajectories built upon accumulated SHAP values *Indirect validation*: visual and quantitative consistency of decision trajectories within fault classes	+ Does not require fault labels; application-agnostic – Limited distinction between similar fault types; faults must be observable
[18]	Model trained on normal (no-occupancy) data	SHAP values on a surrogate model, not the detector itself *No validation*	+ Strong empirical performance; handles non-Gaussian multivariate data – SHAP applied to a secondary model, not the anomaly detection model
[19]	Model trained on normal (no-occupancy) data	SHAP values on a surrogate model, not the detector itself *No validation*	+ Strong empirical performance; handles non-Gaussian multivariate data – SHAP explains another, supervised model: explanations are not faithful; single dataset for validation
[20]	Contrastive pretraining; supervised fine-tuning	SHAP values *No validation*	+ Effective use of unlabeled data – Limited evaluation; high computational complexity
[21]	Semi-supervised cross-modal representation learning	LIME identifies features influencing predictions for visual and audio inputs *Indirect validation*: visual inspection of LIME masks	+ Reduced sensor requirements and operational cost – Sequential structure is largely discarded; domain-specific assumptions
[22]	Pseudo-labeling with double verification	Agreement between local and global interpreters *Indirect validation*: proxy metrics (fidelity, stability, robustness, efficiency, AOPC)	+ Model-agnostic; empirical gains in performance – Domain specific; hyperparameters chosen empirically
[23]	SVM-based semi-supervised model	Five explicit, human-interpretable metrics that characterize embeddings properties *Indirect validation*: consistency across datasets	+ Multivariate evaluation; visualization of explicit temporal metric – Limited interpretability scope (embedding geometry)
[24]	Pairwise constraints in COBRAS	LinC explains relations in COBRAS by connecting instances through chains of constraint- and similarity-based links *Indirect validation*: visualization; consistency with ground-truth labels	+ Clear visual explanation of clustering decisions; helps with stopping decision – Model-specific; depends on quality of distance metric
[25]	GraphSAGE; contrastive pretraining	GNN-Explainer; image-level interpretability *Indirect validation*: subgraphs checked against known biological pathways; reconstruction quality	+ Leverages graph topology; handles multi-instance data – Performance may degrade with very few known interactions; limited datasets

additional examples of the use of Shapley values by proposing complementary approaches based on statistical and distance-based anomaly detection methods. These methods employ independent component analysis (ICA), the Kantorovich distance, and double exponentially weighted moving averages (DEWMAs) to detect changes in occupancy and driving behavior. First, ICA is used to extract independent latent components from multivariate signals. The Kantorovich distance quantifies the difference in the distributions of the baseline and

test signals. DEWMAs then tracks these differences over time to identify gradual or sudden anomalies. A nonparametric, kernel-based thresholding strategy enables detection under weak supervision. Yet, to apply SHAP, the authors ignore these advanced representations and instead explain a completely separate XGBoost model trained on raw inputs. This creates a disconnect between the detection model and the explainability analysis, limiting the relevance of the insights derived from SHAP. However, the analysis highlights the significance of different environmental factors.

Sun *et al*. (2025) propose a semi-supervised learning framework that combines self-supervised contrastive learning with a transformer encoder for lithology classification. The approach first uses unlabeled sequences in a self-supervised pretraining phase, where two augmented views of the same sequence are passed through a shared encoder. A contrastive objective based on negative cosine similarity, without explicit negative samples, is employed, with gradient stopping to prevent collapse. After pretraining, the encoder is fine-tuned using a limited amount of labeled core data. For interpretability, the trained model is analyzed with SHAP, which provides feature-level attributions for individual lithology predictions and feature interaction effects. The method is thoroughly validated on two real-world well-logging datasets, where it demonstrates strong performance, better than that of state-of-the-art models in accuracy and other metrics.

3.2. Other local explanations. Another popular post-hoc explanation method is local interpretable model-agnostic explanations (known as LIME) (Ribeiro *et al*., 2016). LIME approximates the behavior of a complex model around a particular observation by fitting a simple local surrogate model (typically linear regression), enabling the identification of features that most strongly influence the prediction.

Xie *et al*. (2025) propose cross-modality knowledge transfer (CMKT) for in-situ defect detection using audio and visual sensor data, with explanations generated with LIME. The aim is to maintain high predictive performance using only data from a single sensor, thereby reducing costs. To this end, three CMKT methods are proposed: semantic alignment, fully supervised mapping, and semi-supervised mapping. Semantic alignment uses a shared encoder and classifier across modalities, trained with losses that encourage alignment of same-class samples from different modalities while separating samples from different classes. Fully and semi-supervised mapping approaches learn to map features from one modality to another by training a model using either known labels (fully supervised) or autoencoder-based unsupervised representations (semi-supervised), followed by supervised classification on the mapped features. In a case study, the authors compare the proposed methods against single-modal and multimodal fusion baselines. Semantic alignment CMKT achieves the highest performance with a single modality at test time, slightly outperforming full multimodal fusion. LIME reveals that these models focus on more relevant, shared features while discarding modality-specific noise.

3.3. Other. Beyond standard post-hoc explanation methods such as SHAP and LIME, some studies propose dedicated explainability approaches designed to better align with the specific structure or context of the application.

Yuan *et al*. (2024) focus on improving the reliability of anomaly detection explanations by introducing SADDE, a framework designed to provide reliable interpretations for anomaly detection and generate high-confidence pseudo-labels in low-label settings. This method combines two key components: the global-local knowledge association mechanism and the two-stage semi-supervised learning system. The former improves the trustworthiness of model interpretations by combining insights from two different interpretability methods: a local interpreter (e.g., DeepLift (Shrikumar *et al*., 2019)), which identifies key features influencing the prediction for a single instance, and a global interpreter, which identifies important features at the cluster level. A similarity score is computed between the key features identified by the local and global interpreters. If it exceeds a predefined threshold and both models assign the same label, the pseudo-label is accepted as reliable and used for retraining. This double verification mechanism filters out unreliable pseudo-labels, reducing false positives and missed detections. The authors validate SADDE on two public network anomaly detection datasets using only a small fraction of labeled data for pre-training. The framework is shown to outperform eight state-of-the-art interpreters (in fidelity, stability, robustness, and efficiency) and five semi-supervised learning baselines. These results demonstrate SADDE's effectiveness in producing interpretable and accurate anomaly detection under limited supervision. An essential aspect of model interpretability is evaluating the quality of the explanations produced.

Atitey *et al*. (2023) propose MIBCOVIS, a benchmarking framework for evaluating dimensionality-reduction techniques used for dynamic or spatial visualization and interpretability when ground truth is unknown. The framework integrates five evaluation metrics: the occupation index, which measures projection space coverage; the uniformity index, which quantifies the spatial uniformity of the embedding; the gradient boosting classifier index, which assesses classification accuracy in the reduced space; the separability index, which assesses cluster separation based on nearest-neighbor consistency;

and the time order structure index, which evaluates preservation of temporal or spatial ordering. These metrics are jointly modeled in a hierarchical Bayesian framework to compare and rank dimensionality reduction methods. The framework is particularly suited for analyzing large-scale biological data, such as single-cell analysis.

Lin *et al*. (2025) propose LinC, a local post-hoc explanation method for the constraint-based semi-supervised clustering algorithm COBRAS (Van Craenendonck *et al.*, 2018a) and its time-series variant COBRAS-TS (Van Craenendonck *et al.*, 2018b), designed to help users assess clustering quality and decide when to stop the interactive process. Instead of visualizing clusters by overlaying all time series, LinC explains the relationship between two selected instances by constructing a chain of intermediate instances linked by either a user-specified constraint or a high degree of similarity. The method is qualitatively validated on three labeled time-series datasets.

Another post-hoc method is MIGGRI (Huang *et al*., 2023), a two-stage framework for gene regulatory networks inference from spatial gene expression images. First, a neural network is trained with supervised contrastive loss to learn image embeddings using known interacting and non-interacting gene pairs. Second, a multi-instance GraphSAGE model aggregates per-gene image embeddings and performs semi-supervised link prediction using partial GRN topology. The framework leverages the post-hoc explainability tool GNN-Explainer (Ying *et al*., 2019), which identifies a compact subgraph and a small subset of node features that are most influential for a specific prediction. The explanation is formulated as an optimization problem that maximizes mutual information between the model's prediction and the selected subgraph and features. Experiments on *Drosophila* developmental datasets demonstrate improved GRN reconstruction performance compared to image-only baselines.

3.4. Summary. Some authors propose custom explainability tools and frameworks that introduce specialized interpretability approaches, such as SADDE (Yuan *et al*., 2024) and quality measures within the MIBCOVIS framework (Atitey *et al*., 2023), often tailored to domain-specific tasks. These methods combine multiple sources of information and incorporate additional mechanisms to assess the reliability of explanations in complex analytical settings. However, most post-hoc approaches discussed in this review rely primarily on SHAP or LIME for interpretation (Rao and Wang, 2024; Martakis *et al*., 2022; Harrou *et al*., 2024; Fouzi *et al*., 2024; Xie *et al*., 2025).

This is most likely due to the popularity and ease of interpretation of these approaches, as they are model-agnostic and offer valuable insights into model decisions. While SHAP values are often considered more robust across models (Schlegel *et al*., 2019), LIME's reliance on perturbation sampling may generate out-of-distribution instances, leading to unstable explanations (Meng *et al*., 2024). However, the applicability of these frameworks to sequential data remains contentious. We further investigate this aspect in Section 5.3.

Ultimately, choosing between white-box and black-box models depends on the specific goals and constraints of a given task. No single approach is universally optimal (Wolpert and Macready, 1997), and model choice should be guided by the application domain and the nature of the available data (Loyola-Gonzalez, 2019). Despite the significant and growing research interest in white-box approaches (Vilone and Longo, 2020; Finzel, 2025; Long *et al*., 2025; Di Marino *et al*., 2025), post-hoc methods remain the most prevalent class of explainability techniques in the literature, accounting for a substantial fraction (approximately half) of all explainability approaches (Nauta *et al*., 2023). These methods are frequently applied to deep neural networks, which achieve state-of-the-art performance on many time-series tasks (Turbé *et al*., 2023). However, their opacity often necessitates post-hoc explanation tools that may introduce additional risks in sensitive applications where trust, insight, and accountability are essential (Rudin, 2018; Linardatos *et al*., 2020).

4. Intermediate approaches: From regularization to embeddings

The distinction between inherently transparent white-box models and purely post-hoc explanations is often overly simplistic. While white-box models provide structural clarity, they may lack representational flexibility for complex multivariate time series. Conversely, post-hoc explanations applied to opaque models can suffer from approximation errors, instability, and faithfulness issues (Rudin, 2018; Slack *et al*., 2020).

Intermediate approaches aim to bridge this gap. Rather than restricting the hypothesis space to fully transparent forms or treating explanations as external artifacts, they integrate interpretability directly into the learning process. This is achieved through structured embeddings, sparsity-inducing regularization, graph-based modeling, disentanglement objectives, prototype learning, or domain-informed constraints (Doshi-Velez and Kim, 2017; Rudin *et al*., 2022).

The interpretability of such systems relies on the assumption that the learned representations are semantically meaningful. Consequently, representation learning and feature construction are not merely auxiliary

Table 5. Summary of the reviewed intermediate papers. Paper identifiers (ID) follow Table A2.

ID	Partial supervision	Explainability integration/validation	Strengths/limitations
[26]	Coarse labels, sub-maneuver discovery	And–or graph and spatio-temporal LSTM *No validation*: argued conceptually and structurally	+ Sensitive to motion changes; interpretable hierarchical representation – Feature engineering required, high data demands
[27]	Pseudo-labeling	Interpretable multi-view state-transition graphs *No validation*: conceptual, not empirically assessed	+ Interpretable representation – Discretization sensitivity; domain-specific
[28]	One-class learning	Regularized interpretable spatial-temporal GCN *Indirect validation*: alignment with manual inspection; case studies	+ Interpretable architecture – Offline training only
[29]	Self-supervision	Attention weights in the graph attention network *Indirect validation*: illustrated qualitatively; case studies; heatmaps	+ Personalized, adaptive thresholding; intuitive graph representations – Scalability concerns with large numbers of sensors
[30]	Label propagation	Graph-based learning as a linear, shift-invariant graph filter *No validation*: theoretical and analytical	+ Scalable; unique solution; simple, closed-form expression – Unclear role of parameter σ; analysis limited to doubly stochastic graphs
[31]	Physical model constraint	Physical model; initial model constraints; sparsity regularization *Indirect validation*: geological consistency	+ Physics-guided explainability, works well with limited labeled data – Domain-dependent; fine-tuning
[32]	Graph-regularized SVM; transductive SVM	L1-regularized linear SVM *Indirect validation*: non-zero weights correspond to specific time points with known behavior	+ Interpretability via sparsity; temporal feature selection – Limited performance of TSVM; inconsistent performance improvement
[33]	Failure-guided data partitioning	Sparse GGM + semantic knowledge graph; post-hoc contrastive explanations (IBM OpenScale) *Indirect validation*: expert confirmation by reliability engineers	+ Interpretable probabilistic structure; context integration – Semantic models require expert knowledge
[34]	Medical labels guide specific latents	Interpretable latent temporal processes and disentanglement guided by medical knowledge *Indirect validation*: latent trajectory visualization, clustering, and disentanglement comparison	+ General; supports clustering, forecasting, and uncertainty quantification – Weaker than dedicated supervised models; requires domain labels
[35]	Cross-entropy regularizes the latent space	Joint training of a classifier and VAE: cluster separation with semantic meaning *Indirect validation*: expert evaluation on patients; controlled simulations with known parameters	+ Continuous diagnostic map; clinically aligned latent space – Classifier regularization harms generative quality
[36]	Graph-based label propagation	Latent space structuring, enforced via compact clustering *Indirect validation*: t-SNE visualizations of the latent space; cluster purity	+ Robustness to perturbations; improvement over supervised baseline – No interpretability in the input space; evaluation on a single aviation dataset
[37]	Label-aligned supervised latents and orthogonal residual latents	Supervised latents reconstruct labels; unsupervised latents are orthogonal and encouraged to be disentangled *Indirect validation*: visual latent traversals; label reconstruction; interpretable downstream tasks	+ Combines expert and learned features; improves downstream tasks – Manual, costly hyperparameter tuning; dependence on labeled pose features

[38]	Centroid-based regularization	Interpretable embedding space and spatial interaction modeling *Indirect validation*: t-SNE visualizations	+ Spatio-temporal dynamics; visual interpretability of the training dynamics – No explainability in the input space
[39]	Manifold regularization on unlabeled data	Explicit feature construction and projection matrices *Indirect validation*: feature separability visualization (t-SNE)	+ HDC improves performance over raw sensor features – HDC representation lacks semantic interpretability
[40]	Self-supervised pretraining; supervised triplet construction	Space structuring with a triplet loss: clusters correspond to fault types *Indirect validation*: t-SNE visualization; Calinsk–Harabasz index	+ No predefined number of clusters; low labeling cost – Not suitable for real-time inference; explainability is geometric
[41]	Labels used for classification thresholds	Most non-Gaussian components treated as most discriminative *No validation*	+ Improved stability of ICA – Limited usage of labeled data
[42]	Labels structure the latent space	Label-aligned latent variables encoding interpretable signal parameters *Indirect validation*: signal parameter recovery; visual inspection of localized reconstruction errors	+ Interpretable latents; anomaly localization – Interpretability is limited to a few predefined latent variables
[43]	Label-guided encoder fine-tuning	Mapping graph representation to discriminative functional connections and associated brain regions *No validation*	+ Domain grounded representations; robust feature learning – No instance level interpretability
[44]	Labels for fine-tuning with regularization	Temporal progress embeddings: spatial attention heatmaps aligned with report semantics *Indirect validation*: comparison of attention maps and clinically meaningful report terms	+ Alignment with radiology report semantics; no dense labels required – No instance-level explanations; reliance on attention mechanisms
[45]	Contrastive pretraining and gradual unfreezing	Latent representations comparison using protected attribute conditioned similarity *Indirect validation*: correlating representation similarity gaps with performance and fairness gaps	+ Reducing fairness disparities across protected attributes – Descriptive representation similarity
[46]	Labeled data for dictionary generation	Interpretable signal-processing features; linear predictive coding *No validation*	+ Dictionary-based behavior modeling; adaptive discovery of novel events – Manual thresholds, feature engineering
[47]	Ground-truth for hyperparameter tuning	Matrix-based semantic segmentation curves; optimal transport barycenters *Indirect validation*: visual alignment of segmentation curves with known ground-truth events	+ Handles concept drift and anomalies jointly; robust multi-sensor pooling – Multiple preprocessing steps and hyperparameters
[48]	Weakly supervised non contrastive learning	CNN-based spatial self-attention weight extraction *Indirect validation*: visual inspection of attention weights vs. threshold baseline	+ Cause-effect link; improved recall over threshold-based methods – Heuristic thresholds; no ground-truth fall labels
[49]	Reconstruction-based framework	Confidence-aware imputation and discriminator-based uncertainty *Indirect validation*: confidence and trajectory support analysis; alignment with historical means	+ Robust to sparse observations; more stable than GAN-based imputation; scalability – Confidence-only interpretability; no feature/instance rationale
[50]	Unsupervised discovery; semi-supervised tuning	Feature construction based on domain knowledge and gray-box modeling *Indirect validation*: expert inspection and qualitative alignment with known and simulated faults	+ Expert knowledge incorporation; detection of multiple anomaly types – Tedious feature engineering; false positives risk; limited scalability

[51]	Pairwise must/cannot-link constraints	Pairwise constraints reflecting user intent; interpretable cluster prototypes *No validation*	+ No predefined number of clusters; handles separated components – Computational cost; prone to noise
[52]	Constrained cluster expansion	Interpretation of linear regression coefficients *No validation*	+ Transferability; physically meaningful parameters – Heuristic parameters; domain assumptions
[53]	Must/cannot-link constraints	Constrained DTW-preserving shapelets; shapelet cluster explanation process *No validation*	+ Integrates expert knowledge; constraints generalize to new data – Dependence on DTW quality
[54]	Domain adaptation	Explicit Granger-causal graphs *Indirect validation*: visualization; ablations	+ Causal interpretability; theoretical guarantees – Predefined time lag; causal assumption

components but central elements of explainability. The intermediate approaches examined in this section can therefore be viewed as mechanisms that constrain representation learning so that the resulting features align with domain-relevant structures rather than arbitrary latent abstractions.

By embedding interpretability into the architecture or objective function, these methods often preserve much of the predictive power of deep models while improving transparency (Bell *et al*., 2022; Kruschel *et al*., 2024). However, the faithfulness and robustness of such explanations vary considerably and remain active areas of research (Linardatos *et al*., 2020).

In this section, we examine how interpretability is embedded into the learning process. Specifically, we analyze how structural choices at the level of representation, objective function, or domain constraints shape the transparency of semi-supervised models. Table 5 summarizes the reviewed intermediate methods individually.

4.1. Graph and structured representations. A common strategy in intermediate approaches is to embed structural priors into the model representation. Rather than relying exclusively on abstract latent features, these models use graphs, hierarchical structures, or physically grounded constraints to encode relational, spatial, or temporal dependencies. Interpretability is enhanced by aligning the internal representation with the known data topology or domain structure, facilitating a correspondence between learned patterns and real-world relationships. Within this family of approaches, three recurring design principles emerge: (i) hierarchical decomposition of behavior into structured subcomponents, (ii) graph-based encoding of temporal or spatial dependencies, and (iii) explicit regularization grounded in physical or topological priors.

Dai *et al*. (2022) propose a hybrid framework that integrates spatio-temporal long short-term memory (LSTM) for trajectory encoding with an and–or graph probabilistic model representing hierarchical maneuvering logic. The spatio-temporal LSTM captures continuous motion dynamics from raw trajectory sequences, while the semi-supervised and–or graph decomposes coarse maneuvers into structured sub-maneuvers governed by rule-based transitions. The and–or graph enforces compositional constraints over possible maneuver sequences, allowing inference to remain consistent with human-understandable driving logic even under limited supervision. Similarly, Chen *et al*. (2021) propose Semi-Traj2Graph, which transforms GPS trajectories into three complementary graph views: a difficulty graph (transition complexity), a probability graph (transition likelihood), and a time graph (average transition duration). These structured views are processed jointly by a graph convolutional network under a multi-task learning objective, while pseudo-labeling enables semi-supervised refinement. Interpretability is achieved through the explicit relational encoding of driving states, in which classification decisions can be traced to graph-level transition patterns rather than to opaque feature activations.

Xu and Li (2024) represent system logs as spatio-temporal graphs, where nodes correspond to log events and edges encode temporal or semantic dependencies. A graph convolutional network aggregates contextual information across event neighborhoods, while an attention mechanism assigns adaptive weights to nodes and edges during message passing. Anomalies are detected by deviations in learned graph representations, and interpretability arises from inspecting attention weights and high-impact subgraphs, which reveal the event interactions that contributed most to a decision.

Temporal graphs are also applied in the biomedical domain, as demonstrated by Bijlani *et al*. (2024), where a Graph Barlow Twins architecture models daily household activities using temporal graphs derived from wearable and ambient sensors. Self-supervised contrastive learning enforces invariance between augmented graph views, producing stable graph-level embeddings. Daily anomaly scores are calculated using embedding deviations relative to learned activity distributions. These scores are then translated into actionable decisions by clinician-defined alert thresholds. In this approach, interpretability is

grounded in structural saliency (node/edge influence) and transparent thresholding, rather than symbolic rule extraction.

In contrast to neural architectures, several works incorporate graph priors in a more explicit manner. Girault *et al*. (2014) formulate semi-supervised learning as a graph signal reconstruction problem, where labels are treated as signals defined over graph nodes. A Wiener filtering interpretation is derived by minimizing the reconstruction error under a graph-Laplacian smoothness prior. This leads to a closed-form solution that propagates labels according to the graph's spectral properties. Interpretability is a direct consequence of the assumption that connected nodes should exhibit similar signal values, which makes label diffusion traceable through graph frequencies. Similarly, Flamary *et al*. (2015) incorporate graph-based regularization into support vector machines for land-cover classification from satellite image time series. A Laplacian regularizer enforces temporal smoothness across yearly satellite observations, coupling classifiers across time and preventing abrupt label transitions. The graph structure explicitly encodes temporal dependencies, rendering classification decisions interpretable as smooth trajectories over time rather than isolated predictions.

Physics-informed extensions, such as the convolutional Bi-LSTM with physical information and L1 regularization model (Liu *et al*., 2025), further incorporate structural priors into deep architectures. The model integrates the CNN and Bi-LSTM layers with a physics-guided loss term derived from an initial physical model, while an L_1 penalty is employed to enforce sparsity on reflection coefficients. This dual constraint stabilizes inversion results and restricts the hypothesis space to physically viable solutions. In this case, interpretability is derived from both the inspection of learned weights and the enforcement of physically meaningful regularity and sparsity constraints during the optimization process. The performance of these constrained models is validated through a series of comparative experiments on both synthetic and real seismic data. These experiments demonstrate that the proposed model exhibits enhanced stability under degraded initial conditions, improved vertical resolution, and higher correlation coefficients compared with traditional model-driven inversion methods and alternative deep learning baselines.

4.2. Latent space structuring and disentanglement. A common strategy for improving interpretability in semi-supervised models is to explicitly structure the latent space. Rather than allowing representations to emerge unconstrained, these methods employ architectural constraints, disentanglement objectives, or guided supervision to align latent dimensions with semantically meaningful factors. In this setting, interpretability does not arise from explicit decision rules but instead from constraining the geometry of the representation space. Such constraints encourage task-relevant axes, spatio-temporal components, or partitioned subspaces to reflect meaningful distinctions, even when only limited labeled data are available.

Several studies structure latent spaces through clinically guided semi-supervision. For example, Trottet *et al*. (2024) use a temporal variational autoencoder to analyze multivariate clinical trajectories. This is achieved via a latent stochastic process conditioned on the patient context. In addition to the standard reconstruction and KL-divergence terms, the model incorporates a guidance network that predicts sparse clinical labels from latent variables and injects a supervised loss into the training process. This additional objective encourages the latent space to organize along medically meaningful axes (e.g., disease subtypes or progression stages), effectively anchoring generative trajectories to clinically interpretable manifolds. Costa *et al*. (2021) employ a related mechanism in their recurrent variational autoencoder, where bidirectional LSTM layers encode intracardiac time series and a classification head is trained jointly with the generative objective. A cross-entropy loss is added to the evidence lower bound, with the explicit aim of forcing separation of latent representations associated with different atrial fibrillation stages. The resulting two-dimensional latent projections serve as diagnostic maps, where cluster separation directly corresponds to clinically distinct arrhythmic patterns.

Memarzadeh *et al*. (2022) introduce a robust and explainable semi-supervised deep learning model for anomaly detection in aviation data. The model is trained end-to-end and consists of three components: an encoder–decoder for representation learning and reconstruction, a classifier trained on labeled data, and a graph-based compact clustering via a label propagation term that shapes the latent space. The training objective combines (i) a supervised cross-entropy loss on labeled samples, (ii) a compact clustering via label propagation loss defined as the cross-entropy between a transition matrix derived from label-propagated class posteriors and the transition matrix induced by latent-space similarities, and (iii) a reconstruction loss used to regularize the learned representations and improve robustness. Validation is performed on a real dataset, where the authors compare their model against other baselines. The evaluation covers (1) classification performance under varying sizes of the labeled set, (2) the latent-space structure, assessed via t-SNE visualizations, and cluster purity, as well as (3) robustness, measured under the fast gradient sign method adversarial perturbations scheme. This method consistently outperforms baselines in terms of accuracy when labels are scarce and yields more

compact, class-pure latent clusters.

The process of disentanglement can also be enforced structurally through architectural partitioning. Whiteway *et al*. (2021) propose the partitioned subspace variational autoencoder, which splits the latent space into supervised and unsupervised subspaces. The supervised subset is trained to linearly reconstruct user-provided pose labels, while the unsupervised one captures residual behavioral variability. Orthogonality and independence constraints between the two subspaces prevent information leaking, ensuring that known behavioral factors are separated from unexplained dynamics. This architectural split means that hidden dimensions can be directly analyzed as either labeled pose components or new behavioral features. Zuo *et al*. (2021) employ a related structural strategy in semi-supervised spatio-temporal representation learning for multivariate time series, achieving disentanglement using a dual-channel autoencoder. One channel, based on recurrent units, handles temporal dependencies, while a separate spatial modeling block captures inter-variable relationships. The multi-stage regularization procedure starts by using labeled samples to initialize class centroids, and then refines them using soft assignments of unlabeled data. This process promotes class-separable embeddings. In this work, interpretability is derived from the explicit separation of temporal and spatial factors, combined with a centroid-based structure in the embedding space.

Alternative approaches enhance interpretability by enforcing a strong geometric structure on the embedding space. For example, Chen *et al*. (2023) encode raw multivariate sensor signals using hyperdimensional computing, producing high-dimensional binary vectors. Then, they utilize algebraic operations to preserve semantic similarity. Next, a semi-supervised twin projection vector machine learns two projection directions that maximize inter-class separation, leveraging unlabeled samples to stabilize the decision boundary. Because the class structure is linked to separable projection geometry, driving styles become distinguishable via structured binary representations rather than opaque feature activations. In a similar vein, SensorDBSCAN (Ivanov *et al*., 2025) constructs a structured embedding space using a triplet-based contrastive loss, which encourages anchor-positive proximity and anchor-negative separation. The method combines density-based clustering with active sample selection, querying labels for clusters with high uncertainty. This process of refinements makes compact, easily separable clusters in latent space, making anomalies easily detectable as separate regions.

Statistical approaches offer an alternative way to interpret the latent space by explicitly stabilizing or regularizing the representation structure. Chakrabarty and Levkowitz (2020) address the instability of independent component analysis (ICA) by performing multiple ICA runs to calculate higher-order statistics (4th-order cumulants), and selecting the most non-Gaussian components across decompositions. A label-informed threshold is then learned for classification, thereby transforming unstable ICA outputs into consistent, discriminative signal components. On the other hand, Rajendran *et al*. (2019) utilize an adversarial autoencoder to optimize the reconstruction loss, adversarial regularization to shape the latent distributions, and a semi-supervised classification loss. Anomalies are detected through three interpretable signals: the reconstruction error, deviation in latent representation, and classification inconsistency. Because latent variables are limited to encoding spectrum-level characteristics, the model can localize anomalies in the time-frequency space rather than merely flagging outliers.

Collectively, these works demonstrate how architectural design, statistical regularization, and limited supervision can be used to structure latent spaces into semantically meaningful representations. Numerous approaches have been proposed to disentangle latent factors and improve interpretability in semi-supervised learning. These include clinically guided variational autoencoders (VAEs) (Trottet *et al*., 2024; Costa *et al*., 2021), spatiotemporal disentanglement (Zuo *et al*., 2021), partitioned subspaces (Whiteway *et al*., 2021), binary embeddings (Ivanov *et al*., 2025; Chen *et al*., 2023), stabilized ICA (Chakrabarty and Levkowitz, 2020), and reconstruction-based anomaly detection (Rajendran *et al*., 2019). Together, these approaches illustrate diverse strategies for structuring latent representations to improve interpretability.

4.3. Task-driven self-supervised and contrastive learning. In this class of approaches, interpretability arises from the design of the learning objectives themselves. These typically include contrastive losses, proxy tasks, and self-supervised alignment mechanisms that encourage the model to learn structured representations reflecting properties such as semantic similarity, fairness constraints, or temporal consistency.

In contrast to latent disentanglement methods, interpretability here results from alignment constraints imposed during training rather than from explicitly factorized latent dimensions.

Some of the methods combine contrastive objectives with semi-supervised refinement to produce structurally meaningful representations. For example, Wang *et al*. (2025b) convert functional magnetic resonance imaging time series into functional connectivity graphs, where brain regions form the nodes and edges encode correlation-based interactions. A diffusion augmentation module alters node and edge attributes while preserving the global graph topology, thereby creating positive pairs that retain biologically plausible connectivity patterns.

Two parameter-shared graph isomorphism networks are trained using a contrastive loss to maximize the agreement between augmented views, thereby enforcing invariance to noise while preserving structural dependencies. The pretrained encoder is then fine-tuned on a limited amount of labeled data for disorder classification. Interpretability is enabled by learned embeddings that remain anchored to the connectivity structure; the discriminative features correspond to stable interaction patterns between regions, rather than arbitrary latent correlations.

Contrastive design also supports multimodal and longitudinal learning settings. For example, Gao *et al*. (2025) propose a multimodal longitudinal representation learning pipeline that combines sequential breast MRI scans with clinical reports to predict response to neoadjuvant therapy. The model builds modality-specific encoders whose embeddings are brought together through a single-time contrastive loss, encouraging MRI and textual representations from the same visit to converge in the latent space. Moreover, a multi-time contrastive objective enforces temporal consistency by aligning embeddings across different treatment stages for the same patient. This dual alignment constrains the embedding trajectory to reflect both cross-modal coherence and disease progression dynamics. Interpretability arises because treatment response patterns become geometrically traceable in the latent space, such that changes in embedding position correspond to clinically meaningful changes in the patient's condition, rather than arbitrary representation drift.

Yfantidou *et al*. (2024) examine the significance of fairness and transparency in representation learning. The authors adapt a variant of simple framework for contrastive learning of visual representations to time-series data. They compare self-supervised pretraining against a supervised baseline by isolating encoder representations through a stepwise freezing protocol. Specifically, pretrained encoders are frozen, and only linear classifiers are trained on top, enabling attribution of downstream fairness effects to the learned representations rather than to task-specific fine-tuning. Representation similarity is further analyzed by using centered kernel alignment, while fairness is quantified via metrics such as disparate impact and false omission rate across protected attributes. The results show that self-supervised pretraining can reduce performance disparities across demographic groups while maintaining competitive predictive accuracy. In this setting, interpretability extends beyond structural transparency, as the geometry of the learned representation directly influences the ethical properties of the model.

The efficacy of these self-supervised and contrastive approaches is demonstrated by their ability to produce representations that are both robust and interpretable, thereby serving as a foundation for training objectives. As demonstrated in the existing literature, SimCLR adaptations, such as those outlined by Yfantidou *et al*. (2024), address the critical issue of fairness in representation learning. Concurrently, biomedical frameworks, including RetVes (Ekong *et al*., 2024) and GCDA (Wang *et al*., 2025b), employ contrastive learning to enhance cross-domain generalization and interpretability. Multimodal designs (e.g., Gao *et al*., 2025) demonstrate the efficacy of contrastive alignment across time and data types in facilitating meaningful, interpretable embeddings in clinical prediction tasks.

4.4. Domain-constrained and causality-informed modeling. This group of approaches achieves interpretability by incorporating external knowledge, such as physical laws, expert-defined constraints, or causal structures, into the model's assumptions or training process. Rather than relying solely on architectural design or learning objectives, these methods guide model behavior through structured priors, human feedback, or domain-aligned logic. Predictions or internal representations are considered interpretable when they can be directly linked to known processes, reliable approximations, or causal relationships. This enables robust analysis even under sparse or ambiguous supervision. These approaches can be categorized into three overarching paradigms: (i) signal-dictionary and statistical modeling grounded in interpretable descriptors, (ii) constraint-guided clustering and segmentation informed by expert logic, and (iii) causality-preserving domain adaptation frameworks.

Several models achieve interpretability by grounding anomaly detection in an explicit statistical structure rather than latent neural representations. For instance, Min and Tewfik (2011) explore the use of skewness and kurtosis of the frequency distribution, together with linear predictive coding, to detect self-injurious behavior from multichannel accelerometer data. Known behavioral patterns are encoded as autoregressive coefficients and stored in a dictionary. New segments are matched using correlation and pole similarity, while recurring unknown patterns are added to the dictionary. This progressive pattern-matching framework renders the recognition of anomalies traceable to interpretable signal descriptors in the time-frequency domain. Similarly, Cheema *et al*. (2023) introduce a structural health monitoring pipeline that transforms multivariate sensor signals into frequency-domain subsequences, computes matrix profiles to detect differences in near-neighbor similarity, and derives semantic segmentation through the semantic segmentation curve. Multi-sensor information is then integrated via optimal transport barycenters, enabling geometrically grounded comparisons across distributions. In this context, anomalies are explained through observable pattern breaks, distributional shifts,

and misalignment across sensors. This makes deviations interpretable in terms of structural dynamics rather than opaque model scores.

El Marhraoui *et al*. (2023) use a regression model based on the convolutional neural network (CNN) to predict a clinical measure of balance performance and fall risk from inertial sensor data, rather than directly classifying fall events. Interpretability is introduced through a self-attention mechanism that computes query-key similarity maps to weight temporal-spatial signal fragments according to their contribution to the regression output. Furthermore, a perturbation-based consistency module is developed to enforce representation stability under controlled signal distortions. This module facilitates the identification of signal regions that exhibit a significant alteration in prediction stability following distortion. The resulting attention weights and sensitivity patterns provide localized, physiologically meaningful explanations. Qin *et al*. (2021) enhance interpretability by explicitly estimating confidence within a reconstruction framework. A temporal graph convolutional variational autoencoder (TG-VAE) has been developed to impute missing traffic states from sparse GPS data. In addition, a mask discriminative network (MDN) has been designed to predict the observability mask and estimate reconstruction reliability. The synergy between the cooperative and competitive aspects of the TG-VAE and the MDN yields calibrated confidence scores aligned with data quality. The model goes beyond the symbolic representation of predictions by offering a more comprehensive depiction of uncertainty structures. This enables users to assess the reliability of each imputed state, facilitating informed decision-making processes.

Another approach combines human-in-the-loop clustering with a rule-guided structure. For example, Michałowska *et al*. (2021) formulate anomaly detection as a two-phase procedure that combines unsupervised feature exploration with expert-driven refinement. Initial anomalies are identified through exploratory statistics and visualization (e.g., distribution comparison and feature ranking). Domain experts then iteratively relabel false positives and adjust feature selection. Interpretability is derived from transparent feature-level reasoning and the traceable incorporation of expert corrections into subsequent model updates. In a similar vein, Van Craenendonck *et al.* (2018b) propose a semi-supervised clustering framework that integrates user-provided must-link and cannot-link constraints into an iterative cluster refinement process. The algorithm does not implement a prior cluster fixing; rather, it progressively divides and combines "super-instance" in response to constraints specified by the user. To achieve this, time-series-specific similarity metrics are used (e.g., dynamic time warping (DTW) or shape-based distances).

Constraint-based interpretability can also be embedded into clustering dynamics. Ertl *et al*. (2021) introduce an extension of the classical density-based spatial clustering of applications with noise (DBSCAN), which incorporates logic-aware expansion rules. These rules state that a point may only join a cluster if additional statistical or physical constraints are met (e.g., preserving regression stability or aligning with domain-specific guidelines). This process effectively transforms density-based clustering into semantically constrained segmentation, such that each resulting phase corresponds to a behavior that satisfies explicitly defined conditions. Interpretability stems from the fact that cluster membership is justified not only by distance but also by domain-valid logical rules. Similarly, El Amouri *et al*. (2023) incorporate domain knowledge into representation learning through shapelet-based representations that preserve dynamic time warping alignment constraints. The system's objective function enforces attraction and repulsion constraints that reflect expert intuition while preserving the DTW structure. The learned embedding supports classical clustering, and an explanation mechanism subsequently links each cluster to explicit temporal motifs, such as characteristic trends and discontinuities.

In settings with distribution shift, interpretability demands reasoning about stable mechanisms rather than only correlations. Li *et al*. (2024) introduce Granger causality alignment, which models causal graphs in both the source and target domains to address this issue. It introduces a congruence regularizer that enforces structural alignment between them. A causal reconstructor identifies temporal dependencies, while a domain-sensitive predictor adapts conditional relationships without violating the preserved causal structure. Interpretability therefore arises from identifying invariant Granger-causal relationships that remain stable across environments, enabling predictions to be linked to persistent influence structures rather than domain-specific artifacts.

4.5. Summary. The reviewed intermediate approaches demonstrate that interpretability can be integrated at various structural levels of the learning framework. Table 6 provides their qualitative comparison. Across graph-based modeling (Dai *et al*., 2022; Chen *et al*., 2021; Xu and Li, 2024), spectral regularization (Girault *et al*., 2014; Flamary *et al*., 2015), physics-informed constraints (Liu *et al*., 2025), latent space structuring (Trottet *et al*., 2024; Costa *et al*., 2021; Whiteway *et al*., 2021; Memarzadeh *et al*., 2022), geometry-driven embedding design (Chen *et al*., 2023; Ivanov *et al*., 2025), contrastive alignment (Wang *et al*., 2025b; Gao *et al*., 2025; Yfantidou *et al*., 2024), and causality-informed modeling (Li *et al*., 2024), interpretability consistently emerges through the use of structured inductive biases

Table 6. Comparison of intermediate model families discussed in Section 4, showing their core ideas, strengths, limitations and how partial supervision is employed.

Model family	Core idea	Strengths (XAI)	Typical limitations	Partial supervision
Graph and structured representations	Encode spatial/ temporal relations via graphs and priors	Transparent dependency modeling; subgraph inspection; alignment with domain topology	Limited scalability; domain knowledge requirements; interpretability relies on graph sparsity	Pseudo-labeling; graph-based label propagation; physically grounded constraints; domain priors
Latent space structuring and disentanglement	Constrain or divide the latent space to align dimensions with semantic/clinical factors	Semantically meaningful and interpretable embeddings and latents; visualizable latent clusters	Disentanglement and interpretation in the input space not guaranteed	Sparse labels guiding latent dimensions; supervised subspaces; label propagation
Task-driven self-supervised and contrastive learning	Use contrastive or self-supervised objectives to shape structured representations	Improved robustness; interpretable similarity structure	Interpretability indirect (via representation geometry); explanation often qualitative	Contrastive pretraining and supervised refinement
Domain-constrained and causality - informed modeling	Incorporation of physical laws, expert rules, or causal structure into learning objectives	Explanations grounded in domain knowledge; causal reasoning	Requires reliable prior knowledge	Constraint-based training; human-in-the-loop feedback; causal alignment; label-guided hyperparameter tuning

rather than through explicit rule extraction.

A frequently observed design principle is the restriction of representation geometry. Such constraints may be imposed through mechanisms including graph smoothness, subspace partitioning, clustering constraints, alignment objectives, and causal regularizers. These restrictions limit how internal representations can be organized. As a result, clusters, trajectories, or dependency structures often correspond to domain-recognizable patterns such as behavioral phases, disease progression, temporal motifs, or stable influence relationships. Interpretability can arise from the topology of structured representations rather than from transparent parametrization.

These observations suggest that intermediate models should be viewed not as a compromise between white-box and post-hoc paradigms, but as a distinct class of inductively constrained representation learning frameworks. Their effectiveness depends on whether the imposed structural constraints correspond to a meaningful structure in the underlying data-generating process.

5. Lessons learned and open challenges

This section summarizes the main insights of the review and discusses the advances, limitations, and open challenges of current approaches.

5.1. Validation of explainability: Limited and inconsistent. A recurring observation across the reviewed literature is that, although explainability is frequently claimed as a key contribution, it is often insufficiently validated or not explicitly addressed. This issue is highlighted by Nauta *et al*. (2023), who report that one in three papers relies exclusively on anecdotal evidence, while only one in five includes user-based evaluation.

In this work, we analyze how interpretability and explainability are empirically validated in the reviewed studies. Each article is assigned to one of three validation levels:

- *no validation*: interpretability or explainability is claimed conceptually or architecturally, but no empirical assessment of explanation quality is provided;
- *indirect validation*: interpretability is supported through downstream performance improvements, ablation studies, qualitative visualizations (e.g., heatmaps, case studies), or alignment with domain knowledge, without explicit evaluation of explanation fidelity or robustness;
- *direct validation*: explanation quality is explicitly evaluated using quantitative measures of fidelity, stability, robustness, or structured human-grounded assessment with defined evaluation criteria.

These validation levels are used in Tables 2, 4 and 5.

Figure 3 presents the distribution of reviewed articles across validation levels within each taxonomy category.

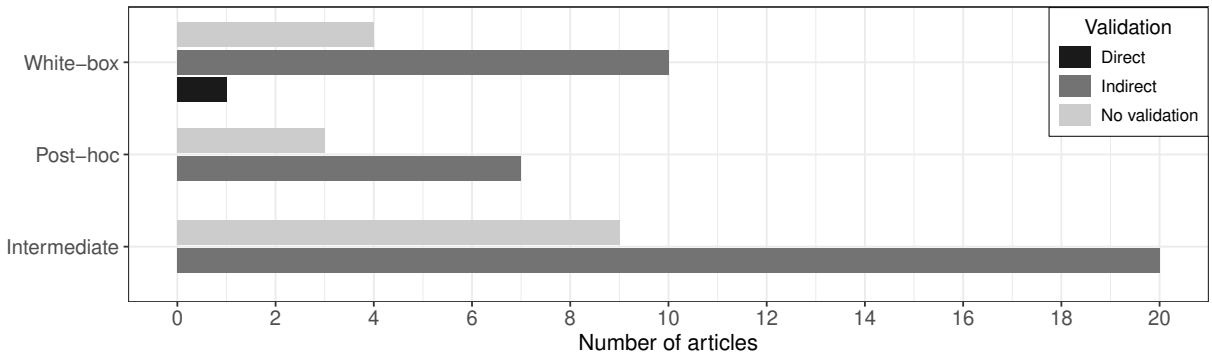

Fig. 3. Validation of explainability or interpretability: the number of articles in each category of our taxonomy.

Nearly one-third (16/54) of the reviewed articles do not report empirical validation of interpretability or explainability. While qualitative and indirect forms of support are common and diverse (37/54), ranging from simple visual illustrations to structured ablation analyses, explicit quantitative evaluation of explanation quality is reported in only one of the 54 surveyed studies (Kabir *et al*., 2025). These findings indicate that formal evaluation practices are not yet consistently integrated into current research workflows. One contributing factor may be the frequent conflation of explainability and interpretability, distinct yet often ill-defined concepts (Lipton, 2017; Murdoch *et al*., 2019; Molnar *et al*., 2020). While they serve different purposes, they are often treated as interchangeable, even though they may be complementary (Garouani *et al*., 2024).

A related challenge concerns the absence of widely accepted evaluation standards. Recent surveys on explainable AI repeatedly highlight the lack of generally accepted criteria, measures, thresholds, and evaluation protocols for explanations (Löfström *et al*., 2022; Nauta *et al*., 2023; Liao *et al*., 2022; Kim *et al*., 2024; Dembinsky *et al*., 2026). Providing an explanation alone is therefore insufficient (Dembinsky *et al*., 2026), yet existing evaluation approaches remain fragmented and often depend on researchers' methodological choices (Seth and Sankarapu, 2025). As a consequence, the quality of explanations can be sensitive to design decisions such as perturbation schemes, baselines, aggregation strategies, or similarity measures. Different evaluation setups may favor different explanation methods, which complicates reliable comparison and ranking (Wickstrøm *et al*., 2024). This is accompanied by the fact that explanations rarely include uncertainty quantification (Molnar *et al*., 2020), and, in most real-world applications, no ground-truth explanation is available against which generated explanations could be objectively assessed (Molnar *et al*., 2020; Wickstrøm *et al*., 2024). This is pronounced in post-hoc frameworks such as SHAP or LIME, where visual or feature-based attributions are typically presented without quantitative assessment of their fidelity, stability, or usefulness. The studies have shown that the outputs of these methods are sensitive to the underlying model and data characteristics, such as feature collinearity (Salih *et al*., 2024). Moreover, explanations produced by different methods, and even by the same one across different runs, frequently disagree, as demonstrated by Neely *et al*. (2021) and Krishna *et al*. (2022). This inconsistency is particularly troubling, since Krishna *et al*. (2022) also observe that practitioners tend to rely on ad hoc heuristics to choose between conflicting explanations, sometimes leading to misleading or harmful decisions.

Even if a given model is interpretable in a mathematical or structural sense, with transparent and well-understood internal mechanisms, this does not automatically translate into explanations that effectively support human understanding, decision-making, or trust in real-world settings (Nauta *et al*., 2023). This distinction is crucial: not all explanations are equally useful or informative in a given context (Doshi-Velez and Kim, 2017), and their practical value depends strongly on purpose, audience, design, and robustness (Hoffman *et al*., 2023; Parcalabescu and Frank, 2023), with simple, local explanations often better supporting end-user understanding (Herm *et al*., 2023). An explanation that appears convincing to a model developer may therefore offer limited benefit to a lay user or practitioner (Miller, 2018), which highlights the need to explicitly consider who the explanation is intended for (Adadi and Berrada, 2018). Moreover, as shown by Buçinca *et al*. (2020), user preference or increased trust does not necessarily translate into improved task performance, and human ratings of explanations may not reliably reflect their true effectiveness (Hase and Bansal, 2020). Finally, visually compelling explanations can create an impression of clarity, while their actual faithfulness to the underlying model behavior remains open to debate (Adebayo *et al*., 2020).

However, unstructured qualitative examination yields highly subjective results, as humans struggle to judge the value of XAI explanations

As pointed out by Leblanc and Germain (2023), the literature largely divides into approaches that explain black-box models post-hoc and inherently interpretable models that often disregard explainability tools altogether. However, integrating these perspectives, for example by ensuring that interpretable models remain transparent in practice, can lead to more effective outcomes (Garouani *et al*., 2024).

For the aforementioned reasons, there is a pressing need to introduce objective measures that enable explanations to be compared. For example, a methodology for computing a standardized evaluation measure that enables quantitative comparison and ranking of explanation methods is proposed by Nguyen *et al*. (2024). The main goal is to identify the most informative explainer for a given time-series classification dataset, and the evaluation is based on the idea that a good explanation should identify data regions that significantly influence the model's predictions. This is assessed by gradually perturbing increasing percentages of the most salient points in the time series, as indicated by the explanation. A larger drop in classifier accuracy after such perturbations suggests a more informative explanation. The method uses multiple classifiers and several perturbation strategies to ensure robustness. The explanation AUC (area under the accuracy curve during perturbation) is a key metric, where lower values reflect higher explanatory quality. A quantitative evaluation framework for post-hoc interpretability methods for neural networks in time-series classification is also proposed by Turbé *et al*. (2023). The authors introduce two metrics to assess how well a method identifies the most and least important time steps without relying on human judgment, retraining, or causing distribution shifts. A new synthetic dataset with known time-dependent features is created for controlled testing. Across three neural network types and three datasets, Shapley value sampling consistently performs best.

Overall, the validation of explainability in semi-supervised time-series models remains fragmented and inconsistent. This can be attributed to the absence of a standardized operational definition of explainability, the diversity of application contexts, and the inherent difficulty of defining canonical, ground-truth explanations against which generated explanations can be compared. Therefore, establishing rigorous and task-aware evaluation frameworks remains a critical open problem in the field.

5.2. Ablation studies: Isolating component effects. Ablation studies systematically remove or disable individual components of a model to assess their specific contribution to overall performance. They are particularly valuable in complex architectures with multiple interacting modules, as they help identify which components are responsible for performance gains (Lipton and Steinhardt, 2018). Despite their diagnostic value, ablation studies are not yet routine in machine learning research, partly because they often require substantial implementation effort and increased computational resources (Sheikholeslami, 2019).

In our review, ablation studies are reported in only five papers. Gao *et al*. (2025) demonstrate through an ablation study that explainability-enhancing modules, single-time vision-text alignment, and multi-time progression modeling consistently improve predictive performance over the baseline. A similar pattern is reported by Liu *et al*. (2024), in whose work removing the self-supervised diffusion and contrastive language-shapelets learning mechanisms, designed to enhance interpretability and exploit limited labeled data, results in reduced classification accuracy. Likewise, the salient subsequence chain and linear discriminant selection components enhance the discovery of representative shapelets and improve clustering performance (Cai *et al*., 2023).

The ablation study by Xu and Li (2024) evaluates the effects of removing components of the interpretable spatial-temporal graph convolutional network. Eliminating the temporal or spatial modules, responsible for identifying important events and capturing similarity-based relationships across events, significantly degrades predictive accuracy and interpretability. Removing both modules leads to the most pronounced performance drop. In contrast, excluding feature selection modules does not reduce predictive performance but increases the number of features and parameters, thereby negatively affecting interpretability.

A similar effect is observed by Li *et al*. (2024), in whose work removing interpretability-related components (e.g., sparsity penalty or causal discrepancy regularization) or forecasting-related components (e.g., a strengthened loss or domain-sensitive latent variables) impairs model performance.

Gu (2022) evaluate the effectiveness of the self-training mechanism in S3OFIS+ by comparing it with the supervised SOFIS+ on 14 benchmark datasets with varying proportions of labeled data, demonstrating its ability to leverage unlabeled data to enhance model performance.

Collectively, these findings suggest a promising direction: incorporating explainability modules and designing model architectures with interpretability in mind can preserve, and in some cases even improve, predictive performance. This challenges the common belief that accuracy and explainability necessarily involve a trade-off.

5.3. Limitations of post-hoc explainability for time series. The majority of methods described in Section 3

rely on the SHAP or LIME frameworks for post-hoc explanation. Although these methods are model-agnostic in their standard formulation, their direct application to time-series data can be problematic because they typically operate on independently perturbed input components and therefore do not explicitly account for sequential dependencies (Bento *et al*., 2021; Le Nguyen and Ifrim, 2025). Moreover, when applied to long time-series sequences, such approaches may incur substantial computational cost (Le Nguyen and Ifrim, 2025; Nayebi *et al*., 2022; Jullum *et al*., 2021; Utkin and Konstantinov, 2021).

These limitations have motivated the development of extensions specifically tailored to sequential data. We identify several such attempts among the reviewed articles. For instance, Martakis *et al*. (2022) use accumulated SHAP scores to construct decision trajectories, whereas Rao and Wang (2024) apply SHAP to windowed, pattern-based representations rather than to raw time-series data.

Beyond the methods reviewed in detail, Table 7 provides an overview of the prominent SHAP and LIME extensions for time-series data that could be beneficial for future XAI studies, compared with direct applications of the standard SHAP and LIME. For example, instead of explaining a single input vector, TimeSHAP (Bento *et al*., 2021) computes Shapley values over entire sequences, assigning attribution to both features and time steps by perturbing inputs across time, not just across features, and includes pruning strategies to handle long sequences efficiently. In principle, the method operates in a supervised learning setting; however, mechanisms such as pseudo-labeling enable its applicability in a semi-supervised setting.

Similarly, WindowSHAP (Nayebi *et al*., 2022) extends SHAP to time-series data by grouping consecutive time steps into windows and computing Shapley values over these aggregated segments. Instead of assigning importance to individual time points, it explains model predictions by attributing importance to entire windows. TSHAP (Le Nguyen and Ifrim, 2025) builds on the same windowing idea and considers two variants: one which computes time-step attributions from exact sliding-window Shapley values, and one, which leverages these window attributions to identify important regions within the time series. Compared to existing window-based methods, TSHAP substantially reduces runtime while achieving high attribution quality, showing strong agreement with ground-truth attributions on synthetic data and high faithfulness on real datasets.

Another example is C-SHAP (Jutte *et al*., 2025), a concept-based extension of SHAP designed for time-series data. The method explains model predictions using high-level temporal concepts defined as human-interpretable patterns, which are treated as features in the explanation process. In the illustrative example, time-series decomposition is used to define the concepts, including trend, daily seasonality, weekly seasonality, and a residual other component.

ShaTS (de la Peña *et al.*, 2025) is a Shapley-based, model-agnostic explainer that performs prior feature grouping (e.g., by time steps, sensors/actuators, or processes), which helps preserve temporal dependencies and yields more actionable explanations. The method identifies critical time steps and factors contributing to anomalies while being more computationally efficient than KernelSHAP.

As a LIME extension, B-LIME (Abdullah *et al*., 2023) uses bootstrapping to improve stability and segments the input signal into meaningful chunks (such as heartbeats) rather than perturbing individual time points, thereby respecting the time structure. In LIMESegment (Sivill and Flach, 2022), the time series is divided into meaningful segments instead of treating each time point as a separate feature. To generate perturbations, it replaces entire segments with realistic alternatives from other sequences, avoiding the unnatural noise used in the standard LIME. It also employs time-series-aware distance measures, such as dynamic time warping, to determine which perturbed samples are considered local. Wang *et al*. (2025a) introduce TF-LIME, which extends LIME to the time-frequency domain. It transforms signals using the short-time Fourier transform (STFT) and segments the resulting time-frequency matrix into homogeneous regions corresponding to coherent frequency patterns over time. These regions are perturbed, reconstructed via an inverse STFT, and used to fit a local surrogate model, whose weights indicate the importance of each region. The method performs well, even under noisy conditions.

Another related approach is LEFTIST (Guillemé *et al*., 2019), which proposes a model-agnostic local explanation framework for time-series classification, building on the principles of LIME and SHAP. The method adapts these frameworks to time series by defining suitable segmentation and reconstruction functions. A time series is represented as a set of interpretable segments, and a local linear surrogate model is learned by perturbing segments through transformations such as random background replacement, linear interpolation, or constant substitution. Experimental results indicate that LEFTIST achieves high-fidelity local approximations and produces explanations that are often understandable to human users.

Overall, these findings suggest that post-hoc explainability methods, although widely adopted, remain challenging to be applied reliably in sequential settings. Their sensitivity to perturbation strategies, temporal dependencies, and model-specific properties raises concerns about the fidelity and stability of the generated

Table 7. Overview of prominent extensions to the SHAP/LIME approaches proposed in the literature for time-series data. Please note that semi-supervision is not directly incorporated in some of them.

Method	Temporal extension	Core idea
SHAP	TimeSHAP (Bento *et al*., 2021)	Shapley values computed over sequences and attributions assigned to both features and time steps; perturbing inputs across time, pruning strategies
	WindowSHAP (Nayebi *et al*., 2022)	Windowing time steps and computing Shapley values on the aggregated segments
	C-SHAP (Jutte *et al*., 2025)	Construction of high-level temporal, human-interpretable concepts and treating them as features
	TSHAP (Le Nguyen and Ifrim, 2025)	Sliding-window grouping of time steps with exact window-level Shapley values; efficient and faithful; classification and regression
	ShaTS (de la Peña *et al.*, 2025)	A priori feature grouping (by time instants, sensors/actuators, or processes), preserving temporal dependencies
LIME	B-LIME (Abdullah *et al*., 2023)	Bootstrapping for improved stability; input signal segmentation into meaningful chunks (e.g., heartbeats), rather than perturbation of individual time points
	LIMESegment (Sivill and Flach, 2022)	Dividing data into meaningful segments; replacing segments with realistic alternatives from other sequences to generate perturbations (noise reduction)
	TF-LIME (Wang *et al*., 2025a)	LIME extension to time-frequency domain; uses the STFT as well as its inverse, segments the time-frequency matrix into homogeneous regions that are perturbed, reconstructed and used to fit a local surrogate model
Mixed	LEFTIST (Guillemé *et al*., 2019)	Model-agnostic, LIME/SHAP-style local explainer that operates on interpretable segments

explanations. This highlights the need for techniques specifically tailored to time-series data that can provide more reliable insights.

5.4. Partial supervision as structural guidance for explainability. The approaches reviewed in this article confirm that partial supervision is usually structurally integrated into the learning process and often closely tied to the model's explainability, rather than acting merely as a source of pseudo-labeled examples.

In many white-box and intermediate models, partial supervision provides sparse semantic guidance. Labeled data anchor specific components of the representation, such as latent dimensions, prototypes, or cluster structures, to semantically meaningful concepts, whereas unlabeled data shape the global geometry of the representation space (Trottet *et al*., 2024; Costa *et al*., 2021; Liu *et al*., 2015). This guidance encourages the representations to remain interpretable.

A similar mechanism arises when partial supervision is incorporated through external constraints associated with domain-specific knowledge. By embedding prior knowledge directly into the learning objective, such constraints preserve structural assumptions and make the influence of supervision explicitly traceable (Grzegorowski *et al*., 2025; Libbrecht *et al*., 2015; El Amouri *et al*., 2023). This mechanism is particularly evident in human-in-the-loop approaches, where experts iteratively inspect and adjust learned representations (Van Craenendonck *et al.*, 2018b; Michałowska *et al*., 2021). In this setting, partial supervision enables direct and transparent intervention in the learning process, linking the representation structure to domain reasoning.

These observations suggest that the explainability of semi-supervised time-series approaches depends not only on the task under consideration and the model type, but also on how supervision is incorporated. Different strategies for incorporating supervision can yield fundamentally different representation structures and, consequently, varying degrees of model transparency.

6. Conclusions

In this article, we reviewed recent advancements in explainable semi-supervised methods for multivariate time-series analysis. The rapid growth of this field has led to a wide range of developments across domains including healthcare, transportation, industrial monitoring, and behavioral analysis.

Despite these advances, several challenges remain to be addressed. A key issue is the inconsistent treatment of the data's spatiotemporal nature. Specifically, some models are evaluated using metrics that do not adequately capture the temporal dynamics essential for time-series analysis. In many cases, authors fail to provide sufficient justification for their choice of evaluation metrics. Additionally, they often neglect to consider how well these metrics align with the problem structure or the intended use case. This oversight raises concerns about the comparability and interpretability of reported results, which could lead to misleading conclusions if not addressed.

The assessment of explainability adds another layer of complexity. Although post-hoc tools like SHAP and LIME have gained popularity, their use is often inconsistent, and many proposed methods lack standardized evaluation protocols. This issue is even more pronounced in the context of semi-supervised learning, where there is a lack of clear ground truth labels and easily interpretable rationales. As a result, evaluating explainability in these models remains a challenging and underdeveloped area.

In the future, developing domain-specific and temporally aware evaluation metrics will be essential for assessing both model performance and interpretability. It is also important to establish best practices for integrating and evaluating explainability within semi-supervised learning pipelines. Addressing these challenges is crucial for ensuring transparency and building trust in this field. Additionally, these efforts will enable reproducibility and facilitate meaningful comparisons as the field continues to evolve rapidly.

Acknowledgment

The authors acknowledge funding from the project *ExplainMe: Explainable Artificial Intelligence for Monitoring Acoustic Features extracted from Speech* (FENG.02.02-IP.05-0302/23), carried out within the First Team programme of the Foundation for Polish Science co-financed by the European Union under the European Funds for Smart Economy 2021–2027 (FENG).

References

Abdullah, T.A.A., Zahid, M.S.M., Ali, W. and Hassan, S.U. (2023). B-LIME: An improvement of LIME for interpretable deep learning classification of cardiac arrhythmia from ECG signals, *Processes* **11**(2): 595.

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), *IEEE Access* **6**: 52138–52160.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. and Kim, B. (2020). Sanity checks for saliency maps, https://arxiv.org/abs/1810.03292.

Agrahari, S. and Singh, A.K. (2022). Concept drift detection in data stream mining: A literature review, *Journal of King Saud University—Computer and Information Sciences* **34**(10): 9523–9540.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Ser, J.D., Díaz-Rodríguez, N. and Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, *Information Fusion* **99**: 101805, DOI: 10.1016/j.inffus.2023.101805.

Atitey, K., Motsinger-Reif, A.A. and Anchang, B. (2023). Model-based evaluation of spatiotemporal data reduction methods with unknown ground truth through optimal visualization and interpretability metrics, *Briefings in Bioinformatics* **25**(1), DOI: 10.1093/bib/bbad455.

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

Bell, A., Solano-Kamaiko, I., Nov, O. and Stoyanovich, J. (2022). It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy, *2022 ACM Conference on Fairness Accountability and Transparency, Seoul, Republic of Korea*, pp. 248–266.

Bento, J., Saleiro, P., Cruz, A.F., Figueiredo, M.A. and Bizarro, P. (2021). TimeSHAP: Explaining recurrent models through sequence perturbations, *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, New York, USA*, pp. 2565–2573.

Bijlani, N., Maldonado, O.M., Nilforooshan, R., Barnaghi, P. and Kouchaki, S. (2024). Utilizing graph neural networks for adverse health detection and personalized decision making in sensor-based remote monitoring for dementia care, *Computers in Biology and Medicine* **183**(C): 109287.

Buçinca, Z., Lin, P., Gajos, K.Z. and Glassman, E.L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, *Proceedings of the 25th International Conference on Intelligent User Interfaces, New York, USA*, p. 454–464, DOI: 10.1145/3377325.3377498.

Cai, B., Huang, G., Yang, S., Xiang, Y. and Chi, C.-H. (2023). Se-shapelets: Semi-supervised clustering of time series using representative shapelets, *Expert Systems with Applications* **240**(C): 0957–4174, DOI: 10.1016/j.eswa.2023.122584.

Casalino, G., Castellano, G. and Mencar, C. (2019). Data stream classification by dynamic incremental semi-supervised fuzzy clustering, *International Journal on Artificial Intelligence Tools* **28**(08): 1960009.

Chakrabarty, S. and Levkowitz, H. (2020). A new algorithm using independent components for classification and prediction of high dimensional data, *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Valletta, Malta*, pp. 265–272.

Cheema, P., Alamdari, M.M., Vio, G., Azizi, L. and Luo, S. (2023). On the use of matrix profiles and optimal transport theory for multivariate time series anomaly detection within structural health monitoring, *Mechanical Systems and Signal Processing* **204**: 110797.

Chen, C., Liu, Q., Wang, X., Liao, C. and Zhang, D. (2021). SEMI-TRAJ2GRAPH: Identifying fine-grained driving style with GPS trajectory data via multi-task learning, *IEEE Transactions on Big Data* **8**(6): 1550–1565, DOI: 10.1109/TBDATA.2021.3063048.

Chen, X., Gao, Y., Yu, H., Wang, H. and Cai, Y. (2023). Driving style feature extraction and recognition based on hyperdimensional computing and semi-supervised twin projection vector machine, *IEEE Transactions on Intelligent Transportation Systems* **24**(12): 13976–13988.

Costa, N., Sanchez, L. and Couso, I. (2021). Semi-supervised recurrent variational autoencoder approach for visual diagnosis of atrial fibrillation, *IEEE Access* **9**: 40227–40239.

Dai, S., Li, Z., Li, L., Zheng, N. and Wang, S. (2022). A flexible and explainable vehicle motion prediction and inference framework combining semi-supervised AOG and ST-LSTM, *IEEE Transactions on Intelligent Transportation Systems* **23**(2): 840–860.

Dau, H.A., Keogh, E., Kamgar, K., Yeh, C.-C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G. and Hexagon-ML (2018). The UCR time series classification archive, https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

de la Peña, M.F., Ángel Luis Perales Gómez and Maimó, L.F. (2025). ShaTS: A Shapley-based explainability method for time series artificial intelligence models applied to anomaly detection in industrial Internet of things, https://arxiv.org/abs/2506.01450.

Dembinsky, D., Lucieri, A., Frolov, S., Najjar, H., Watanabe, K. and Dengel, A. (2026). Unifying VXAI: A systematic review and framework for the evaluation of explainable AI, https://arxiv.org/abs/2506.15408.

Di Marino, A., Bevilacqua, V., Ciaramella, A., De Falco, I. and Sannino, G. (2025). Ante-HOC methods for interpretable deep models: A survey, *ACM Computing Surveys* **57**(10): 262:1–262:36, DOI: 10.1145/3728637.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning, https://arxiv.org/abs/1702.08608.

Ekong, F., Yu, Y., Patamia, R.A., Sarpong, K., Ukwuoma, C.C., Ukot, A.R. and Cai, J. (2024). Retves segmentation: A pseudo-labeling and feature knowledge distillation optimization technique for retinal vessel channel enhancement, *Computers in Biology and Medicine* **182**: 109150.

El Amouri, H., Lampert, T., Gançarski, P. and Mallet, C. (2023). Constrained DTW preserving shapelets for explainable time-series clustering, *Pattern Recognition* **143**(C): 109804.

El Marhraoui, Y., Bouilland, S., Boukallel, M., Anastassova, M. and Ammi, M. (2023). CNN-based self-attention weight extraction for fall event prediction using balance test score, *Sensors* **23**(22): 9194.

Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwoh, C.-K. and Li, X. (2024). Label-efficient time series representation learning: A review, *IEEE Transactions on Artificial Intelligence* **5**(12): 6027–6042.

Ertl, B., Schneider, M., Diekmann, C., Meyer, J. and Streit, A. (2021). *A Semi-supervised Approach for Trajectory Segmentation to Identify Different Moisture Processes in the Atmosphere*, Springer International Publishing, Berlin/Heidelberg, pp. 264–277.

Fernandez-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?, *Journal of Machine Learning Research* **15**(1): 3133–3181.

Finzel, B. (2025). Current methods in explainable artificial intelligence and future prospects for integrative physiology, *Pflügers Archiv—European Journal of Physiology* **477**(4): 513–529, DOI: 10.1007/s00424-025-03067-7.

Flamary, R., Fauvel, M., Dalla Mura, M. and Valero, S. (2015). Analysis of multitemporal classification techniques for forecasting image time series, *IEEE Geoscience and Remote Sensing Letters* **12**(5): 953–957.

Fouzi, H., Ramakrishna, K.K., Madakyaru, M. and Ying, S. (2024). Efficient data-driven occupancy detection in office environments and feature impact analysis, *International Journal of Information Technology* **17**(5): 3013–3023.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* **9**(3): 432–441.

Gao, Y., Tan, T., Wang, X., Beets-Tan, R., Zhang, T., Han, L., Portaluri, A., Lu, C., Liang, X., Teuwen, J., Zhou, H.-Y. and Mann, R. (2025). Multi-modal longitudinal representation learning for predicting neoadjuvant therapy response in breast cancer treatment, *IEEE Journal of Biomedical and Health Informatics* **29**(12): 9041–9050.

Garouani, M., Mothe, J., Barhrhouj, A. and Aligon, J. (2024). Investigating the duality of interpretability and explainability in machine learning, *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI), Herndon, USA*, p. 861–867, DOI: 10.1109/ICTAI62512.2024.00125.

Girault, B., Goncalves, P., Fleury, E. and Mor, A.S. (2014). Semi-supervised learning for graph to signal mapping: A graph signal Wiener filter interpretation, *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy*, pp. 1115–1119.

Givisis, I., Kalatzis, D., Christakis, C. and Kiouvrekis, Y. (2025). Comparing explainable AI models: SHAP, LIME, and their role in electric field strength prediction over urban areas, *Electronics* **14**(23): 4766.

Grzegorowski, M., Janusz, A., Marcinowski, Ł., Skowron, A., Ślezak, D. and Śliwa, G. (2025). On explainability of cluster prototypes with rough sets: A case study in the FMCG market, *International Journal of Applied Mathematics and Computer Science* **35**(1): 19–31, DOI: 10.61822/amcs-2025-0002.

Gu, X. (2022). An explainable semi-supervised self-organizing fuzzy inference system for streaming data classification, *Information Sciences* **583**: 364–385.

Guillemé, M., Masson, V., Rozé, L. and Termier, A. (2019). Agnostic local explanation for time series classification, *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA*, pp. 432–439.

Harrou, F., Kini, K.R., Madakyaru, M. and Sun, Y. (2024). A semi-supervised anomaly detection strategy for drunk driving detection: A feasibility study, *Frontiers in Sensors* **5**: 1375034.

Hase, P. and Bansal, M. (2020). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? https://arxiv.org/abs/2005.01831.

Herm, L.-V., Heinrich, K., Wanner, J. and Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability, *International Journal of Information Management* **69**(C): 102538.

Hoffman, R.R., Mueller, S.T., Klein, G. and Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance, *Frontiers in Computer Science* **5**: 1096257.

Huang, Y., Yu, G. and Yang, Y. (2023). MIGGRI: A multi-instance graph neural network model for inferring gene regulatory networks for *Drosophila* from spatial expression images, *PLOS Computational Biology* **19**(11): e1011623.

ISO, I.T. (2020). Software and systems engineering: Software testing. Part 11: Guidelines on the testing of AI-based systems, *ISO/IEC TR 29119-11:2020*, International Organization for Standarization, Geneva, https://www.iso.org/standard/79016.html.

ISO, I.T. (2025). Objectives and approaches for explainability and interpretability of ML models and AI systems, International Organization for Standardization, *ISO/IEC TS 6254:2025*, Geneva, https://www.iso.org/standard/82148.html.

Ivanov, P., Shtark, M., Kozhevnikov, A., Golyadkin, M., Botov, D. and Makarov, I. (2025). SensorDBSCAN: Semi-supervised active learning powered method for anomaly detection and diagnosis, *IEEE Access* **13**: 25186–25197.

Jullum, M., Redelmeier, A. and Aas, K. (2021). GroupShapley: Efficient prediction explanation with Shapley values for feature groups, https://arxiv.org/abs/2106.12228.

Jutte, A., Ahmed, F., Linssen, J. and van Keulen, M. (2025). C-SHAP for time series: An approach to high-level temporal explanations, https://arxiv.org/abs/2504.11159

Kabir, S., Hossain, M.S. and Andersson, K. (2025). A semi-supervised-learning-aided explainable belief rule-based approach to predict the energy consumption of buildings, *Algorithms* **18**(6): 305.

Kaczmarek-Majer, K., Casalino, G., Castellano, G., Hryniewicz, O. and Dominiak, M. (2022). Explaining smartphone-based acoustic data in bipolar disorder: SEMI-supervised fuzzy clustering and relative linguistic summaries, *Information Sciences* **588**(C): 174–195.

Kim, J., Maathuis, H. and Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review, *Frontiers in Artificial Intelligence* **7**: 1456486.

Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S. and Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective, *OpenReview.net*, https://openreview.net/forum?id=jESY2WTZCe.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks, *Communications of the ACM* **60**(6): 84–90.

Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M. and Zschech, P. (2024). Challenging the performance-interpretability trade-off: An evaluation of interpretable machine learning models, *Business & Information Systems Engineering* **68**(1): 159–183

Le Nguyen, T. and Ifrim, G. (2025). TSHAP: Fast and exact SHAP for explaining time series classification and regression, *ECML PKDD 2025, Porto, Portugal*, pp. 60–77.

Leblanc, B. and Germain, P. (2023). On the relationship between interpretability and explainability in machine learning, DOI: 10.48550/arXiv.2311.11491.

Leite, D., Decker, L., Santana, M. and Souza, P. (2020). EGFC: Evolving Gaussian fuzzy classifier from never-ending semi-supervised data streams with application to power quality disturbance detection and classification, *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK*, pp. 1–9.

Li, C., Denison, T. and Zhu, T. (2025). A survey of few-shot learning for biomedical time series, *IEEE Reviews in Biomedical Engineering* **18**: 192–210.

Li, Z., Cai, R., Fu, T.Z.J., Hao, Z. and Zhang, K. (2024). Transferable time-series forecasting under causal conditional shift, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(4): 1932–1949.

Liao, Q.V., Zhang, Y., Luss, R., Doshi-Velez, F. and Dhurandhar, A. (2022). Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable AI, https://arxiv.org/abs/2206.10847.

Libbrecht, M., Hoffman, M., Bilmes, J. and Noble, W. (2015). Entropic graph-based posterior regularization, *Proceedings of the 32nd International Conference on Machine Learning, Lille, France*, pp. 1992–2000.

Lin, S., Yurtman, A., Soenen, J. and Blockeel, H. (2025). LinC: Explaining time series clusterings with user-provided constraints, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Turin,* pp. 37–52.

Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods, *Entropy* **23**(1): 18.

Lipton, Z.C. (2017). The mythos of model interpretability, https://arxiv.org/abs/1606.03490.

Lipton, Z.C. and Steinhardt, J. (2018). Troubling trends in machine learning scholarship, https://arxiv.org/abs/1807.03341.

Liu, C., Wang, F., Hu, J. and Xiong, H. (2015). Temporal phenotyping from longitudinal electronic health records: A graph based framework, *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA*, pp. 705–714.

Liu, Z., Liu, D., Sacchi, M.D., Li, J., Chen, X., Wu, Y. and Liu, G. (2025). Physically guided high-resolution acoustic impedance inversion based on hybrid networks, *IEEE Transactions on Geoscience and Remote Sensing* **63**: 1–10.

Liu, Z., Pei, W., Lan, D. and Ma, Q. (2024). Diffusion language-shapelets for semi-supervised time-series classification, *AAAI-24 Technical Track* **38**(13): 14079–14087.

Long, B., Liu, E., Qiu, R. and Duan, Y. (2025). Explainable AI the latest advancements and new trends, https://arxiv.org/abs/2505.07005.

Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view, *IEEE Access* **7**: 154096–154113.

Lughofer, E. (2022). Evolving multi-label fuzzy classifier, *Information Sciences* **597**: 1–23.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions, DOI: 10.1016/j.ins.2022.03.045.

Löfström, H., Hammar, K. and Johansson, U. (2022). A meta survey of quality evaluation criteria in explanation methods, *in* J. De Weerdt and A. Polyvyanyy (Eds), *Intelligent Information Systems*, Lecture Notes in Business Information Processing, Vol. 452, pp. 55–63, DOI: 10.1007/978-3-031-07481-3_7.

Martakis, P., Movsessian, A., Reuland, Y., Pai, S.G., Quqa, S., Cava, D., Tcherniak, D. and Chatzi, E. (2022). A semi-supervised interpretable machine learning framework for sensor fault detection, *Smart Structures and Systems* **29**(1): 251–266, DOI: 10.12989/sss.2022.29.1.251.

Memarzadeh, M., Akbari Asanjan, A. and Matthews, B. (2022). Robust and explainable semi-supervised deep learning model for anomaly detection in aviation, *Aerospace* **9**(8): 437.

Meng, H., Wagner, C. and Triguero, I. (2024). SEGAL time series classification—Stable explanations using a generative model and an adaptive weighting method for LIME, *Neural Networks* **176**(C): 106345.

Michałowska, K., Riemer-Sørensen, S., Sterud, C. and Hjellset, O.M. (2021). Anomaly detection with unknown anomalies: Application to maritime machinery, *IFAC-PapersOnLine* **54**(16): 105–111.

Michelioudakis, E., Artikis, A. and Paliouras, G. (2023). Online semi-supervised learning of composite event rules by combining structure and mass-based predicate similarity, *Machine Learning* **113**(3): 1445–1481.

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences, https://arxiv.org/abs/1706.07269.

Min, C.-H. and Tewfik, A.H. (2011). Semi-supervised event detection using higher order statistics for multidimensional time series accelerometer data, *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, USA*, pp. 365–368.

Molnar, C., Casalicchio, G. and Bischl, B. (2020). Interpretable machine learning—A brief history, state-of-the-art and challenges, *ECML PKDD 2020* (online), p. 417–431.

Moradi, M., Komninos, P. and Zarouchas, D. (2024). Constructing explainable health indicators for aircraft engines by developing an interpretable neural network with discretized weights, *Applied Intelligence* **55**(2): 143.

Movsessian, A., Garcia Cava, D., Tcherniak, D. and Janeliukstis, R. (2020). A methodology on interpretable novelty detection, *XI International Conference on Structural Dynamics*, *EURODYN 2020, Athens, Greece*, pp. 922–935.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning, *Proceedings of the National Academy of Sciences* **116**(44): 22071–22080, DOI: 10.1073/pnas.1900654116.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M. and Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI, *ACM Computing Surveys* **55**(13s): 1–42.

Nayebi, A., Tipirneni, S., Reddy, C.K., Foreman, B. and Subbian, V. (2022). WindowSHAP: An efficient framework for explaining time-series classifiers based on Shapley values, *Journal of Biomedical Informatics* **144**: 104438, DOI: 10.1016/j.jbi.2023.104438.

Neely, M., Schouten, S.F., Bleeker, M.J.R. and Lucic, A. (2021). Order in the court: Explainable AI methods prone to disagreement, https://arxiv.org/abs/2105.03287.

Nguyen, H., Cao, H., Nguyen, V. and Pham, D. (2021). Evaluation of explainable artificial intelligence: SHAP, LIME, and CAM, *Proceedings of the FPT AI Conference*, pp: 1–6.

Nguyen, T.T., Le Nguyen, T. and Ifrim, G. (2024). Robust explainer recommendation for time series classification, *Data Mining and Knowledge Discovery* **38**(6): 3372–3413.

Noah, O. and Pum, M. (2024). Evaluating explainability in AI models: Comparing SHAP, LIME, and other techniques, https://www.researchgate.net/publication/389357147.

Parcalabescu, L. and Frank, A. (2023). On measuring faithfulness or self-consistency of natural language explanations, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand*, pp. 6048–6089.

Qin, H., Zhan, X., Li, Y., Yang, X. and Zheng, Y. (2021). Network-wide traffic states imputation using self-interested coalitional learning, *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD 21, New York, USA*, pp. 1370–1378.

Quinlan, J.R. (1986). Induction of decision trees, *Machine Learning* **1**(1): 81–106.

Rajendran, S., Meert, W., Lenders, V. and Pollin, S. (2019). Unsupervised wireless spectrum anomaly detection with interpretable features, *IEEE Transactions on Cognitive Communications and Networking* **5**(3): 637–647.

Rao, S. and Wang, J. (2024). A comprehensive fault detection and diagnosis method for chemical processes, *Chemical Engineering Science* **300**: 120565.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 16, San Francisco, USA*, pp. 1135–1144.

Rudin, C. (2018). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence,* **1**(5): 206–215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L. and Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges, *Statistics Surveys* **16**: 1–85.

Salih, A.M., Raisi-Estabragh, Z., Galazzo, I.B., Radeva, P., Petersen, S.E., Lekadir, K. and Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME, *Advanced Intelligent Systems* **7**(1): 2400304.

Salih, A.M., Raisi-Estabragh, Z., Galazzo, I.B., Radeva, P., Petersen, S.E., Lekadir, K. and Menegaz, G. (2024). A perspective on explainable artificial intelligence methods: SHAP and LIME, *Advanced Intelligent Systems* **7**(1): 1–8.

Saritha, A. and Dhanalakshmi, M. (2024). A survey of advancements in anomaly detection for multivariate time series data, *in* C. Chakraborty *et al.* (Eds), *Multi-faceted Approaches for Data Acquisition, Processing & Communication*, Boca Raton, pp. 208–215, DOI: 10.1201/9781003470939-27.

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D. and Keim, D.A. (2019). Towards a rigorous evaluation of XAI methods on time series, *2019 ICCV Workshop on Interpreting and Explaining Visual Artificial Intelligence Models, Seoul, South Korea*, pp. 4197–4201.

Seth, P. and Sankarapu, V.K. (2025). Bridging the gap in XAI—Why reliable metrics matter for explainability and compliance, https://arxiv.org/abs/2502.04695.

Shapley, L.S. (1953). 17. A value for n-person games, *in* H. W. Kuhn, and A.W. Tucker (Eds), *Contributions to the Theory of Games II*, Princeton University Press, Princeton, pp. 307–318.

Sheikholeslami, S. (2019). *Ablation Programming for Machine Learning*, Master's thesis, KTH Royal Institute of Technology, Stockholm, p. 52.

Shrikumar, A., Greenside, P. and Kundaje, A. (2019). Learning important features through propagating activation differences, https://arxiv.org/abs/1704.02685.

Sivill, T. and Flach, P. (2022). LIMESegment: Meaningful, realistic time series explanations, *PMLR* **151**: 3418–3433.

Slack, D., Hilgard, S., Jia, E., Singh, S. and Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods, https://arxiv.org/abs/1911.02508.

Sun, Y., Pang, S., Li, H., Qiao, S. and Zhang, Y. (2025). Enhanced lithology classification using an interpretable SHAP model integrating semi-supervised contrastive learning and transformer with well logging data, *Natural Resources Research* **34**(2): 785–813.

Trottet, C., Schürch, M., Allam, A., Barua, I., Petelytska, L., Launay, D., Airó, P., Bečvář, R., Denton, C., Radic, M., Distler, O., Hoffmann-Vold, A.-M., Krauthammer, M. and EUSTAR (2024). SEMI-supervised generative models for disease trajectories: A case study on systemic sclerosis, https://arxiv.org/abs/2407.11427.

Turbé, H., Bjelogrlic, M., Lovis, C. and Mengaldo, G. (2023). Evaluation of POST-HOC interpretability methods in time-series classification, *Nature Machine Intelligence* **5**(3): 250–260.

Utkin, L.V. and Konstantinov, A.V. (2021). Ensembles of random SHAPS, https://arxiv.org/abs/2103.03302.

Van Craenendonck, T., Dumancic, S. and Blockeel, H. (2018a). COBRA: A fast and simple method for active clustering with pairwise constraints, *Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia,* pp. 2871–2877.

Van Craenendonck, T., Meert, W., Dumancic, S. and Blockeel, H. (2018b). COBRAS-TS: A new approach to semi-supervised clustering of time series, *in* L. Soldatova *et al.* (Eds), *Discovery Science: DS 2018*, DOI: 10.1007/978-3-030-01771-2_12.

Veeraraghavan, H., Papanikolopoulos, N. and Schrater, P. (2007). Learning dynamic event descriptions in image sequences, *2007 IEEE Conference on Computer Vision and Pattern Recognition, Los Alamitos, USA*, pp. 1–6.

Vernon, E.M., Masuyama, N. and Nojima, Y. (2024). Integrating white and black box techniques for interpretable machine learning, *Proceedings of 9th International Congress on Information and Communication Technology, Singapore*, pp. 639–649.

Vilone, G. and Longo, L. (2020). Explainable artificial intelligence: A systematic review, https://arxiv.org/abs/2006.00093.

Vinzamuri, B., Khabiri, E., Bhamidipaty, A., Mckim, G. and Gandhi, B. (2020). An end-to-end context aware anomaly detection system, *2020 IEEE International Conference on Big Data (IEEE Big Data 2020), Atlanta, USA*, pp. 1689–1698.

Wang, H., Zhang, Q., Wu, J., Pan, S. and Chen, Y. (2019). Time series feature learning with labeled and unlabeled data, *Pattern Recognition* **89**: 55–66.

Wang, J., Zhang, R. and Li, Q. (2025a). TF-LIME: Interpretation method for time-series models based on time-frequency features, *Sensors* **25**(9): 2845.

Wang, X., Fang, Y., Wang, Q., Yap, P.-T., Zhu, H. and Liu, M. (2025b). Self-supervised graph contrastive learning with diffusion augmentation for functional MRI analysis and brain disorder detection, *Medical Image Analysis* **101**: 103403.

Whiteway, M.R., Biderman, D., Friedman, Y., Dipoppa, M., Buchanan, E.K., Wu, A., Zhou, J., Bonacchi, N., Miska, N.J., Noel, J.-P., Rodriguez, E., Schartner, M., Socha, K., Urai, A.E., Salzman, C.D., Cunningham, J.P. and Paninski, L. (2021). Partitioning variability in animal behavioral videos using semi-supervised variational autoencoders, *PLOS Computational Biology* **17**(9): e1009439.

Wickstrøm, K., Höhne, M. M.-C. and Hedström, A. (2024). From flexibility to manipulation: The slippery slope of XAI evaluation, https://arxiv.org/abs/2412.05592.

Wolpert, D. and Macready, W. (1997). No free lunch theorems for optimization, *IEEE Transactions on Evolutionary Computation* **1**(1): 67–82.

Xie, J., Safdar, M., Chen, L., Moon, S.K. and Zhao, Y.F. (2025). Audio-visual cross-modality knowledge transfer for machine learning-based in-situ monitoring in laser additive manufacturing, *Additive Manufacturing* **101**: 104692.

Xu, R. and Li, Y. (2024). Interpretable spatial-temporal graph convolutional network for system log anomaly detection, *Advanced Engineering Informatics* **62**(C): 102803.

Ye, L. and Keogh, E. (2009). Time series shapelets: A new primitive for data mining, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD09, Paris, France*, pp. 947–956.

Yfantidou, S., Spathis, D., Constantinides, M., Vakali, A., Quercia, D. and Kawsar, F. (2024). Using self-supervised learning can improve model fairness, *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 24, Barcelona, Spain*, pp. 3942–3953.

Ying, R., Bourgeois, D., You, J., Zitnik, M. and Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks, https://arxiv.org/abs/1903.03894.

Yuan, Y., Huang, Y., Yuan, Y. and Wang, J. (2024). SADDE: Semi-supervised anomaly detection with dependable explanations.

Zadeh, L. (1965). Fuzzy sets, *Information and Control* **8**(3): 338–353.

Zhang, K., Wen, Q., Zhang, C., Cai, R., Jin, M., Liu, Y., Zhang, J.Y., Liang, Y., Pang, G., Song, D. and Pan, S. (2024). Self-supervised learning for time series analysis: Taxonomy, progress, and prospects, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(10): 6775–6794.

Zhu, L., Lu, C. and Sun, Y. (2016). Time series shapelet classification based online short-term voltage stability assessment, *IEEE Transactions on Power Systems* **31**(2): 1430–1439.

Zuo, J., Zeitouni, K. and Taher, Y. (2021). SMATE: Semi-supervised spatio-temporal representation learning on multivariate time series, *2021 IEEE International Conference on Data Mining (ICDM), Los Alamitos, USA*, pp. 1565–1570.

Filip Wichrowski is a PhD student and a research assistant at the Systems Research Institute Polish Academy of Sciences, Warsaw, Poland. He holds a master's degree in mathematical statistics and data analysis from the Warsaw University of Technology. He has so far contributed to several research projects, primarily in medical applications. His main research interests include stochastic processes, time-series analysis, as well as mathematical and computational statistics.

Marcin Ostrowski is an assistant in the ExplainMe project and a PhD student at the Systems Research Institute of the Polish Academy of Sciences in Warsaw, Poland. He holds an MSc in mathematics and data analysis from the Warsaw University of Technology, Poland. His research focuses on artificial intelligence, fuzzy sets, data and time-series analysis, and the application of mathematical methods in medicine, with a strong interest in developing trustworthy AI systems.

Marta Boratyn is a graduate with a master's degree in mathematics and data analysis. During her studies, she developed a strong interest in machine learning, fuzzy clustering, and survival analysis, with a focus on their applications to medical problems.

Katarzyna Kaczmarek-Majer holds an MSc in mathematics and an MSc in computer science from the University of Poznań, Poland. She earned her PhD in computer science with distinction from the Systems Research Institute of the Polish Academy of Sciences in 2015. She is now an associate professor there. Her areas of expertise include soft computing, time-series and data-stream analysis, and human-centered AI. She effectively combines her theoretical research with involvement in scientific projects, with applications mainly in medicine and healthcare. She has co-authored 50+ scientific publications, and is the vice-president of the European Society for Fuzzy Logic and Technology (EUSFLAT).

Appendix

Table A1. Detailed listing of reviewed papers grouped by primary task and explainability type. The column n indicates the number of papers in each task–type subgroup, while N indicates the total number of papers for the corresponding task.

Task	Type	References	n	N
Anomaly detection	White-box	Movsessian *et al*. (2020)	1	14
	Post-hoc	Rao and Wang (2024), Martakis *et al*. (2022), Fouzi *et al*. (2024), Harrou *et al*. (2024), Yuan *et al*. (2024)	5	
	Intermediate	Xu and Li (2024), Bijlani *et al*. (2024), Vinzamuri *et al*. (2020), Ivanov *et al*. (2025), Rajendran *et al*. (2019), Min and Tewfik (2011), Cheema *et al*. (2023), Michałowska *et al*. (2021)	8	
Classification	White-box	Veeraraghavan *et al*. (2007), Lughofer (2022), Leite *et al*. (2020), Gu (2022), Kaczmarek-Majer *et al*. (2022), Zhu *et al*. (2016), Wang *et al*. (2019), Liu *et al*. (2024)	8	17
	Post-hoc	Sun *et al*. (2025)	1	
	Intermediate	Dai *et al*. (2022), Chen *et al*. (2021), Girault *et al*. (2014), Flamary *et al*. (2015), Memarzadeh *et al*. (2022), Chen *et al*. (2023), Chakrabarty and Levkowitz (2020), El Marhraoui *et al*. (2023)	8	
Clustering	White-box	Cai *et al*. (2023), Grzegorowski *et al*. (2025)	2	3
	Post-hoc	–	0	
	Intermediate	Van Craenendonck *et al.* (2018b)	1	
Forecasting	White-box	–	0	1
	Post-hoc	–	0	
	Intermediate	Li *et al*. (2024)	1	
Other	White-box	Michelioudakis *et al*. (2023)	1	3
	Post-hoc	Xie *et al*. (2025), Lin *et al*. (2025)	2	
	Intermediate	–	0	
Regression	White-box	Kabir *et al*. (2025)	1	3
	Post-hoc	–	0	
	Intermediate	Liu *et al*. (2025), Qin *et al*. (2021)	2	
Representation	White-box	Libbrecht *et al*. (2015), Moradi *et al*. (2024)	2	12
	Post-hoc	Atitey *et al*. (2023), Huang *et al*. (2023)	2	
	Intermediate	Trottet *et al*. (2024), Costa *et al*. (2021), Whiteway *et al*. (2021), Zuo *et al*. (2021), Wang *et al*. (2025b), Gao *et al*. (2025), Yfantidou *et al*. (2024), El Amouri *et al*. (2023)	8	
Segmentation	White-box	–	0	1
	Post-hoc	–	0	
	Intermediate	Ertl *et al*. (2021)	1	

Table A2. Mapping between paper identifiers (ID column) used in Tables 2, 4, and 5 and the corresponding references.

Explainability type	ID to reference mapping	
White-box	[1] Michelioudakis *et al*. (2023)	[9] Grzegorowski *et al*. (2025)
	[2] Veeraraghavan *et al*. (2007)	[10] Zhu *et al*. (2016)
	[3] Kabir *et al*. (2025)	[11] Wang *et al*. (2019)
	[4] Movsessian *et al*. (2020)	[12] Cai *et al*. (2023)
	[5] Lughofer (2022)	[13] Liu *et al*. (2024)
	[6] Leite *et al*. (2020)	[14] Libbrecht *et al*. (2015)
	[7] Gu (2022)	[15] Moradi *et al*. (2024)
	[8] Kaczmarek-Majer *et al*. (2022)	
Post-hoc	[16] Rao and Wang (2024)	[21] Xie *et al*. (2025)
	[17] Martakis *et al*. (2022)	[22] Yuan *et al*. (2024)
	[18] Fouzi *et al*. (2024)	[23] Atitey *et al*. (2023)
	[19] Harrou *et al*. (2024)	[24] Lin *et al*. (2025)
	[20] Sun *et al*. (2025)	[25] Huang *et al*. (2023)
Intermediate	[26] Dai *et al*. (2022)	[41] Chakrabarty and Levkowitz (2020)
	[27] Chen *et al*. (2021)	[42] Rajendran *et al*. (2019)
	[28] Xu and Li (2024)	[43] Wang *et al*. (2025b)
	[29] Bijlani *et al*. (2024)	[44] Gao *et al*. (2025)
	[30] Girault *et al*. (2014)	[45] Yfantidou *et al*. (2024)
	[31] Liu *et al*. (2025)	[46] Min and Tewfik (2011)
	[32] Flamary *et al*. (2015)	[47] Cheema *et al*. (2023)
	[33] Vinzamuri *et al*. (2020)	[48] El Marhraoui *et al*. (2023)
	[34] Trottet *et al*. (2024)	[49] Qin *et al*. (2021)
	[35] Costa *et al*. (2021)	[50] Michałowska *et al*. (2021)
	[36] Memarzadeh *et al*. (2022)	[51] Van Craenendonck *et al.* (2018b)
	[37] Whiteway *et al*. (2021)	[52] Ertl *et al*. (2021)
	[38] Zuo *et al*. (2021)	[53] El Amouri *et al*. (2023)
	[39] Chen *et al*. (2023)	[54] Li *et al*. (2024)
	[40] Ivanov *et al*. (2025)	

Received: 31 July 2025
Revised: 27 February 2026
Accepted: 30 March 2026