National Journal of Multidisciplinary Research and Development
www.multinationaljournal.com
ISSN: 2455-9040
Received: 15-02-2026, Accepted: 16-03-2026, Published: 31-03-2026
Volume 11, Issue 1, 2026, Page No. 108-116

From interpretability to knowledge: A meta-analytical conceptual framework for explainable data mining in structured and unstructured big data

T Karthikeyan, Dr. Sashank Swami
Department of Computer Science and Engineering, Vikrant University, Gwalior, Madhya Pradesh, India

Abstract
The rapid growth of structured, unstructured, and hybrid big data has led to widespread adoption of complex machine learning and deep learning models, which often operate as black-box systems. Although explainable artificial intelligence (XAI) techniques aim to improve interpretability, existing approaches remain fragmented across data modalities and frequently fail to translate explanations into actionable knowledge. This study presents a meta-analytical and conceptual synthesis of explainability across structured, unstructured, and hybrid data mining systems. The analysis identifies key patterns and trade-offs, including the shift from intrinsic interpretability to approximation-based methods and the increasing disconnect between model performance and knowledge extraction. The findings highlight a critical gap where explanation does not necessarily imply understanding, causality, or decision relevance. To address this limitation, the paper proposes the Explainability-Driven Knowledge Discovery (EDKD) framework, which integrates data, models, explanations, and knowledge into a unified structure. The framework extends traditional data mining pipelines by explicitly incorporating knowledge as a distinct analytical layer, enabling more effective and human-centered data-driven decision-making.

Keywords: explainable artificial intelligence (XAI), data mining, big data, interpretability, knowledge discovery, meta-analysis, multimodal data, machine learning, deep learning, conceptual framework

Introduction

1. Context and Motivation

The exponential growth of digital data has fundamentally transformed modern computing and knowledge discovery. Contemporary data ecosystems generate large volumes of information across structured (e.g., relational databases and transactional records) and unstructured formats (e.g., text, images, video, and sensor streams), often coexisting within the same analytical environments (Gandomi & Haider, 2015; Han *et al*., 2011) [11, 12]. This expansion has intensified the need for advanced analytical paradigms capable of extracting meaningful insights from heterogeneous data sources (Wu *et al*., 2013; Gandomi & Haider, 2015) [11].

Traditional data mining systems, primarily designed for structured data, rely on models such as decision trees, regression techniques, and rule-based approaches that offer both computational efficiency and interpretability (Han *et al*., 2011; Zaki & Meira, 2020) [12, 34]. These models enable direct inspection of decision logic, facilitating the extraction of understandable patterns and relationships. However, the increasing prevalence of unstructured data has necessitated a shift toward more complex machine learning (ML) and deep learning (DL) models capable of processing high-dimensional and non-tabular inputs.

Deep learning architectures, including convolutional neural networks (CNNs) and transformer-based models, have demonstrated strong capability in learning complex representations from unstructured data, improving predictive performance across domains such as healthcare, finance, and natural language processing (LeCun *et al*., 2015; Vaswani *et al*., 2017; Devlin *et al*., 2019) [9, 32]. However, these models often operate as opaque systems, limiting interpretability.

The integration of structured and unstructured data has further led to hybrid and multimodal data mining systems, which leverage complementary information across modalities to enhance predictive capability (Miotto *et al*., 2016; Baltrušaitis *et al*., 2019) [4, 23]. While such systems represent a significant advancement in data analytics, they also introduce increased complexity in model design, data fusion, and interpretability. Consequently, the challenge extends beyond predictive accuracy to understanding how these systems generate insights and how such insights can be translated into actionable knowledge.

2. Interpretability Challenge

The growing reliance on complex ML and DL models has foregrounded the challenge of interpretability in data mining systems. High-performing models such as deep neural networks, ensemble methods, and transformer architectures often function as black-box systems, where the relationship between inputs and outputs is not readily transparent (Doshi-Velez & Kim, 2017; Rudin, 2019) [10, 27]. This opacity raises critical concerns regarding trust, accountability, and reliability, particularly in domains where decisions have significant real-world implications.

Explainable Artificial Intelligence (XAI) has emerged as a response to this challenge, introducing techniques aimed at making model behavior more understandable. Model-agnostic methods such as LIME (Ribeiro *et al*., 2016) and SHAP (Lundberg & Lee, 2017) [19, 26] provide post-hoc explanations by approximating local decision boundaries or attributing feature contributions.

However, existing explainability paradigms exhibit inherent limitations. Many techniques rely on approximate representations of model behavior, which may not faithfully capture the true decision-making process (Molnar, 2022) [24]. Feature attribution methods often depend on assumptions regarding feature independence and data distribution, leading to inconsistent or misleading explanations

(Lundberg & Lee, 2017; Molnar, 2022) [19, 24]. More fundamentally, explainability does not necessarily guarantee meaningful understanding, as explanations may lack domain relevance or contextual interpretability (Rudin, 2019; Marcinkevičs & Vogt, 2023) [22, 27].

3. Background: Explainability and Knowledge Discovery

Within data mining, interpretability and explainability represent complementary but distinct dimensions. Interpretability refers to the inherent transparency of a model's structure, whereas explainability involves post-hoc methods that approximate or describe the behavior of complex models (Marcinkevičs & Vogt, 2023; Burkart & Huber, 2021) [6, 22]. Accordingly, approaches can be categorized into intrinsic models and post-hoc explanation techniques.

Intrinsic models, such as decision trees and linear models, provide clear representations of decision processes. In contrast, post-hoc methods, including perturbation-based techniques, feature attribution methods, and surrogate models, seek to explain predictions generated by opaque systems (Ribeiro *et al*., 2016; Lundberg & Lee, 2017) [19, 26]. These approaches can be further distinguished into local and global explanations depending on their scope.

From a broader perspective, the primary objective of data mining extends beyond prediction to knowledge discovery, which involves identifying meaningful and actionable patterns within data (Han *et al*., 2011) [12]. However, a critical gap exists between explanation and knowledge. While explanations describe model behavior, knowledge requires contextual relevance, domain alignment, and applicability to decision-making (Rudin, 2019; Molnar, 2022) [24, 27].

4. Research Gap

Despite advancements in explainable machine learning, current paradigms remain fragmented across data modalities and lack a unified conceptual basis. Explainability techniques are often developed within specific contexts without addressing their behavior across heterogeneous data environments (Baltrušaitis *et al*., 2019; Miotto *et al*., 2016) [4, 23].

Moreover, existing approaches primarily treat interpretability as an isolated objective, without examining its role in facilitating knowledge discovery. While explanation methods provide insights into model behavior, they do not adequately address whether these insights translate into actionable knowledge. This limitation is particularly evident in hybrid systems, where multiple modalities interact, increasing complexity and reducing interpretability.

5. Research Objectives and Contributions

This study adopts a meta-analytical and conceptual framework approach to examine explainable data mining across structured, unstructured, and hybrid environments. It positions explainability within a broader pathway linking data, models, explanations, and knowledge.

The key objectives and contributions of this study are as follows

1. **Meta-Analytical Synthesis:** Identification of patterns, trade-offs, and limitations across modalities
2. **Cross-Modality Comparative Analysis:** Examination of modality-specific explainability behavior and knowledge extraction
3. **Conceptual Framework (EDKD):** A unified structure linking data, models, explanations, and knowledge
4. **Theoretical Contribution:** Bridging the gap between interpretability and knowledge discovery

Through these contributions, the study advances the understanding of explainable data mining beyond transparency, toward a more comprehensive framework for knowledge-driven analytics in heterogeneous big data systems.

Analytical Foundations and Conceptual Basis

1. Data Modalities and Problem Context

Modern data-driven systems operate within heterogeneous environments comprising structured, unstructured, and hybrid data modalities. Structured data, typically represented in tabular formats, supports classical data mining approaches with interpretable models such as decision trees and regression techniques (Han *et al*., 2011; Zaki & Meira, 2020) [12, 34].

Unstructured data, including text, images, audio, and video, requires representation learning methods, often implemented through deep learning architectures such as CNNs and transformers (LeCun *et al*., 2015; Vaswani *et al*., 2017; Devlin *et al*., 2019; Coudray *et al*., 2018) [8, 9, 32]. While these models enable high-performance pattern extraction, they introduce reduced transparency.

The integration of these paradigms has led to hybrid and multimodal systems that combine heterogeneous data sources to improve predictive capability (Miotto *et al*., 2016; Baltrušaitis *et al*., 2019) [4, 23]. However, this integration increases complexity in data representation, model behavior, and interpretability.

Accordingly, the central problem is not only predictive performance but understanding how explainability varies across modalities and how it contributes to knowledge extraction.

2. Conceptual Foundations of Explainability and Knowledge

Explainability enables machine learning models to provide interpretable justifications for their predictions (Doshi-Velez & Kim, 2017; Marcinkevičs & Vogt, 2023; Arrieta *et al*., 2020) [2, 10, 22]. It is operationalized through intrinsic and post-hoc approaches, including model-agnostic techniques such as LIME and SHAP (Ribeiro *et al*., 2016; Lundberg & Lee, 2017) [19, 26], with broader methodological perspectives discussed in recent surveys (Burkart & Huber, 2021) [6].

However, these methods often rely on approximations that may not fully capture underlying model behavior, leading to potential inconsistencies (Molnar, 2022) [24]. Feature attribution methods are also sensitive to data characteristics and interactions, raising concerns about reliability (Lundberg & Lee, 2017) [19].

From a broader perspective, data mining aims at knowledge discovery, which involves transforming data into actionable insights (Han *et al*., 2011) [12]. Explanations alone do not guarantee such knowledge; effective knowledge requires contextual relevance, domain alignment, and interpretive integration (Rudin, 2019; Molnar, 2022; Marcinkevičs & Vogt, 2023) [22, 24, 27].

3. Analytical Perspective and Synthesis Approach

This study adopts a meta-analytical and conceptual synthesis approach to examine explainable data mining

across structured, unstructured, and hybrid modalities. The focus is on identifying patterns, relationships, and trade-offs rather than summarizing individual techniques.

The analysis is structured along three dimensions

1. **Data Modality:** Structured, unstructured, and hybrid contexts influencing interpretability (Baltrušaitis *et al*., 2019; Miotto *et al*., 2016) [4, 23]
2. **Explainability Paradigm:** Intrinsic and post-hoc approaches, including model-agnostic and model-specific techniques (Ribeiro *et al*., 2016; Lundberg & Lee, 2017) [19, 26]
3. **Knowledge Extraction Capability:** The extent to which explanations support meaningful and actionable insights (Molnar, 2022; Rudin, 2019) [24, 27]

The synthesis integrates insights from foundational and recent studies to identify modality-dependent behavior, trade-offs between interpretability and complexity, and limitations in supporting knowledge discovery. This conceptual abstraction forms the basis for subsequent analysis and framework development.

Meta-Analytical Synthesis of Explainability across Data Modalities

Explainability in data mining systems evolves significantly across structured, unstructured, and hybrid data modalities, reflecting differences in data representation, model complexity, and interpretability mechanisms. Rather than constituting a uniform property, explainability operates as a modality-dependent construct, where structured systems emphasize intrinsic transparency, unstructured systems rely on post-hoc approximation, and hybrid systems introduce cross-modal complexity and interpretability fragmentation (Burkart & Huber, 2021; Linardatos *et al*., 2021) [6]. This section adopts a meta-analytical perspective to examine these variations, focusing on underlying patterns, contradictions, and trade-offs rather than individual techniques. Through this analytical abstraction, the section establishes a conceptual understanding of how explainability behaves across heterogeneous data environments and how these differences influence the transformation of model outputs into meaningful knowledge.

1. Explainability in Structured Data Systems

Structured data systems represent the traditional foundation of data mining, where data is organized into well-defined feature spaces that support direct analytical interpretation. In this context, explainability is largely embedded within model design, enabling transparent decision-making without reliance on external approximation techniques. However, a meta-analytical perspective reveals that this transparency is achieved through design constraints that limit representational flexibility (Belle & Papantonis, 2021; Burkart & Huber, 2021) [5, 6].

1.1 Intrinsic Interpretability Mechanisms

Structured systems predominantly employ intrinsically interpretable models, such as decision trees, rule-based systems, and linear models, where the relationship between inputs and outputs is explicitly defined (Han *et al*., 2011; Zaki & Meira, 2020) [12, 34]. Decision trees provide hierarchical decision paths, rule-based models encode logical relationships, and linear models quantify feature contributions through coefficients. These approaches exemplify interpretability-by-design, where transparency is inherent rather than externally imposed.

Such models eliminate the need for post-hoc explanation and reduce the risk of approximation error (Rudin, 2019; Doshi-Velez & Kim, 2017) [10, 27]. However, this transparency is achieved at the cost of limited expressive capacity, as these models typically operate within constrained hypothesis spaces. As a result, they may struggle to capture complex, non-linear relationships present in modern datasets.

2. Feature-Based Explanation Paradigms

To address the limitations of intrinsic models, structured systems increasingly incorporate feature-based explanation techniques for more complex algorithms, including ensemble methods. Post-hoc methods such as LIME and SHAP provide local and global approximations of model behavior but introduce challenges related to stability, fidelity, and interpretive consistency (Ribeiro *et al*., 2016; Lundberg & Lee, 2017; Molnar, 2022) [19, 24, 26].

Recent work has further extended these approaches by integrating global and local interpretability mechanisms to improve robustness and usability across diverse datasets (Linardatos *et al*., 2021; Marcinkevičs & Vogt, 2023) [17, 22]. While these methods extend interpretability, they introduce new challenges. Feature attribution often depends on assumptions regarding feature independence and data distribution, which may not hold in practice (Molnar, 2022) [24]. Additionally, explanation stability can vary across instances, and importance scores may not reflect causal relationships (Slack *et al*., 2020) [31]. These limitations highlight the distinction between quantitative attribution and meaningful interpretation.

These approaches can be systematically categorized based on their methodological characteristics and scope, as summarized in Table 1, which outlines intrinsic and post-hoc explainability techniques in structured contexts.

Table 1: Taxonomy of Explainability Techniques across Data Modalities

Category	Sub-Type	Method / Technique	Data Modality	Explanation Scope	Key Characteristics	References
Intrinsic (Model-Based)	Transparent Models	Decision Trees, Rule-Based Models	Structured	Global	Direct interpretability, explicit decision paths	Han *et al*. (2011); Rudin (2019) [12, 27]
	Linear Models	Linear / Logistic Regression	Structured	Global	Coefficient-based feature contribution	Molnar (2022) [24]
Post-hoc Model-Agnostic	Local Surrogate Models	LIME	Structured / Unstructured	Local	Approximation of decision boundary	Ribeiro *et al*. (2016) [26]
	Feature Attribution	SHAP	Structured /	Local + Global	Game-theoretic	Lundberg & Lee

			Unstructured		feature importance	(2017) [19]
Post-hoc Model-Specific	Tree-Based Attribution	TreeSHAP	Structured	Global	Efficient SHAP for ensembles	Lundberg *et al*. (2020) [20]
Perturbation-Based	Input Perturbation	Occlusion, Feature Masking	Unstructured	Local	Sensitivity-based explanations	Molnar (2022) [24]
Gradient-Based	Backpropagation Methods	Grad-CAM, Saliency Maps	Unstructured (Images)	Local	Visual heatmaps of feature importance	Selvaraju *et al*. (2017) [30]
Attention-Based	Attention Weights	Transformer Attention	Unstructured / Multimodal	Local	Highlights input relevance (not always causal)	Vaswani *et al*. (2017); Chefer *et al*. (2021) [7, 32]
Multimodal Explainability	Joint Attribution	Cross-Modal SHAP / Attention	Hybrid	Local + Global	Captures inter-modal interactions	Chefer *et al*. (2021) [7]
Concept-Based	High-Level Interpretations	Concept Activation Vectors	Unstructured	Global	Links features to human concepts	Marcinkevičs & Vogt (2023) [22]

1.3 Strengths and Constraints

Structured systems exhibit high interpretability, low ambiguity, and strong alignment between model behavior and explanation mechanisms. However, these advantages are offset by constraints in representational power and scalability. Increasing model complexity necessitates post-hoc explanations, reintroducing approximation and uncertainty.

Meta-analytically, structured data systems illustrate a fundamental trade-off between interpretability and expressive capacity, where transparency is achieved through simplification. This trade-off sets the baseline for understanding how explainability evolves in more complex data environments.

2. Explainability in Unstructured Data Systems

Unstructured data systems, encompassing text, images, and sequential data, introduce a fundamentally different explainability paradigm. Unlike structured systems, where interpretability is embedded, these systems rely on post-hoc approximation mechanisms due to the opacity of deep learning models. This shift reflects the transition from explicit feature representations to latent, high-dimensional embeddings, which are not directly interpretable (Samek *et al*., 2021; Linardatos *et al*., 2021) [17, 29].

2.1 Post-Hoc Explainability Approaches

Explainability in unstructured systems is dominated by model-agnostic, perturbation-based methods such as LIME and SHAP, which provide approximate insights into model behavior (Ribeiro *et al*., 2016; Lundberg & Lee, 2017) [19, 26]. However, their application introduces representation challenges, as perturbations often rely on artificial constructs that may not align with semantically meaningful features (Molnar, 2022; Samek *et al*., 2021) [24, 29]. Consequently, explanations may reflect the approximation process rather than the underlying model logic. From a meta-analytical perspective, post-hoc explainability in unstructured systems is inherently approximation-driven, involving trade-offs between fidelity, efficiency, and interpretability.

2.2 Deep Learning Opacity

The opacity of unstructured data systems arises from the use of deep learning architectures, including CNNs, RNNs, and transformers, which learn hierarchical and distributed representations (LeCun *et al*., 2015; Vaswani *et al*., 2017) [32]. These models encode information in latent feature spaces, making it difficult to map internal representations to human-understandable concepts.

While such architectures provide high predictive performance, they function as black-box systems, where internal decision processes are not directly accessible (Rudin, 2019; Burkart & Huber, 2021) [6, 27]. This opacity necessitates the use of external explanation methods, reinforcing the dependence on approximation.

2.3 Visual and Attention-Based Explanations

To improve interpretability, visual explanation techniques such as Grad-CAM and saliency maps highlight regions of input data that influence predictions, providing intuitive insights into model behavior (Selvaraju *et al*., 2017; Samek *et al*., 2021) [29, 30]. Similarly, attention mechanisms in transformer models assign weights to input elements, suggesting their relative importance (Vaswani *et al*., 2017) [32].

However, these approaches exhibit important limitations. Attention weights do not necessarily correspond to causal contributions, leading to the widely recognized issue that attention does not equate to explanation (Chefer *et al*., 2021; Marcinkevičs & Vogt, 2023) [7, 22]. Visual explanations may also be sensitive to noise and model perturbations, reducing their reliability (Adebayo *et al*., 2018) [1].

2.4 Limitations in Interpretability

Unstructured systems face significant interpretability challenges, including lack of semantic clarity, inconsistency of explanations, and absence of standardized evaluation methods. Explanations often highlight low-level features without providing higher-level understanding, limiting their usefulness for decision-making.

Additionally, different explanation methods may produce conflicting results, undermining trust and reliability (Molnar, 2022; Slack *et al*., 2020) [24, 31]. The absence of universally accepted evaluation metrics further complicates the assessment of explanation quality (Doshi-Velez & Kim, 2017) [10].

Meta-analytically, unstructured systems demonstrate that explainability becomes increasingly approximate, context-dependent, and less reliable as model complexity increases.

3. Explainability in Hybrid and Multimodal Systems

Hybrid and multimodal systems integrate structured and unstructured data to capture complementary information, representing the most complex form of data mining systems. While these systems enhance predictive capability, they

introduce cross-modal dependencies that fundamentally complicate explainability (Baltrušaitis *et al*., 2019; Kaur *et al*., 2020) [4, 15].

3.1 Fusion Complexity

Hybrid systems rely on data fusion strategies, including early fusion (combining features before modeling) and late fusion (combining outputs of separate models) (Baltrušaitis *et al*., 2019) [4]. Early fusion enables joint representation learning but introduces challenges in aligning heterogeneous data types, particularly when combining symbolic and latent representations. Late fusion preserves modality-specific structures but complicates explanation at the aggregation stage, where multiple decision pathways must be reconciled. Recent multimodal learning studies further highlight that fusion mechanisms significantly influence interpretability, as feature interactions across modalities are often non-linear and difficult to disentangle (Mai *et al*., 2022; Baltrušaitis *et al*., 2019) [4, 21]. This creates a key insight: explainability in hybrid systems is dependent on integration strategy, rather than solely on model type.

3.2 Cross-Modal Explainability Challenges

Hybrid systems face challenges related to inconsistent explanations across modalities and attribution ambiguity. Different data sources may contribute to predictions in ways that are difficult to reconcile within a unified explanation. Traditional attribution methods often fail to capture interdependencies between modalities, limiting their effectiveness.

Recent work in multimodal explainability shows that attribution methods adapted from unimodal settings struggle to provide coherent explanations when applied to heterogeneous data (Chefer *et al*., 2021; Jain *et al*., 2022) [7, 13]. This reveals a fundamental limitation: explainability techniques developed for unimodal systems do not directly extend to multimodal contexts.

3.3 Emerging Multimodal XAI Approaches

Recent multimodal XAI approaches leverage transformer-based architectures to model cross-modal interactions through attention mechanisms, enabling joint representation learning across text, image, and structured data (Vaswani *et al*., 2017; Chefer *et al*., 2021) [7, 32]. Additionally, joint feature attribution methods extend traditional explainability by capturing interdependencies between modalities, supporting more context-aware interpretations (Samek *et al*., 2021) [29]. Applications in multimodal sentiment analysis and vision-language systems further demonstrate their potential for learning aligned cross-modal representations and generating more interpretable semantic associations between modalities (Zhang *et al*., 2021) [35]. However, these approaches remain architecture-dependent and lack standardization, resulting in inconsistent and difficult-to-interpret explanations across domains (Kaur *et al*., 2020; Marcinkevičs & Vogt, 2023) [15, 22].

3.4 System-Level Limitations

Hybrid systems introduce increased computational complexity and reduced interpretability coherence. Generating explanations requires processing multiple data modalities and models, leading to scalability challenges. Additionally, explanations may vary across modalities and methods, resulting in fragmented interpretations.

Human-centered studies further indicate that users struggle to interpret multimodal explanations due to their complexity and lack of consistency, highlighting usability as a critical limitation (Kaur *et al*., 2020) [15]. Meta-analytically, hybrid systems represent the most complex explainability context, where performance gains are accompanied by significant interpretability challenges.

4. Cross-Modality Synthesis

The analysis across structured, unstructured, and hybrid data systems reveals that explainability is not a uniform property but a context-dependent construct shaped by model complexity, data representation, and interpretive mechanisms. A cross-modality synthesis highlights recurring patterns, contradictions, and trade-offs that collectively define the current state of explainable data mining. These insights form the basis for a unified conceptual understanding of explainability and its relationship to knowledge extraction (Linardatos *et al*., 2021; Burkart & Huber, 2021) [6, 17].

4.1 Common Patterns across Modalities

A fundamental pattern observed across data modalities is the progressive shift from intrinsic interpretability to post-hoc explainability. Structured systems rely on transparent, model-inherent mechanisms, whereas unstructured and hybrid systems depend increasingly on external approximation techniques to interpret complex models (Burkart & Huber, 2021; Marcinkevičs & Vogt, 2023) [6, 22]. This transition reflects the growing dominance of deep learning architectures, where interpretability is not embedded but reconstructed after model training.

Closely related is the increasing reliance on approximation-based explanations, particularly in unstructured and multimodal contexts. Techniques such as LIME and SHAP approximate model behavior through perturbations or feature attribution, but these approximations introduce uncertainty and may not faithfully represent the underlying decision process (Ribeiro *et al*., 2016; Lundberg & Lee, 2017) [19, 26]. In multimodal systems, this approximation is further compounded by cross-modal dependencies, where explanations must capture interactions across heterogeneous feature spaces (Chefer *et al*., 2021; Samek *et al*., 2021) [7, 29].

Another consistent pattern is the trade-off between model complexity and interpretability. As models become more expressive, moving from linear models to deep neural networks and multimodal transformers, their internal representations become less transparent, necessitating increasingly sophisticated explanation methods (LeCun *et al*., 2015; Samek *et al*., 2021) [29]. This trade-off underscores the inherent tension between performance and interpretability.

4.2 Contradictions and Inconsistencies

Despite advancements in explainability techniques, several contradictions emerge across modalities. One of the most prominent is the discrepancy between high predictive accuracy and potentially misleading explanations. Models may achieve strong performance while relying on spurious correlations or non-causal features, leading to explanations that do not reflect meaningful patterns (Rudin *et al*., 2022; Molnar, 2022) [24, 28].

Another critical inconsistency arises in the relationship between attention mechanisms and explanation. While

attention weights are often interpreted as indicators of feature importance, studies have shown that attention does not necessarily correspond to causal influence, leading to the well-established notion that "attention is not explanation" (Chefer *et al*., 2021; Jain & Wallace, 2019) [7, 14]. This disconnect is particularly evident in transformer-based models.

Additionally, there is a persistent mismatch between local and global explanations. Local methods provide instance-specific insights, while global methods aim to capture overall model behavior. However, these explanations are not always consistent with each other, leading to fragmented interpretations (Ribeiro *et al*., 2016; Slack *et al*., 2020) [26, 31]. This inconsistency undermines the reliability of explainability techniques.

4.3 Trade-offs across Modalities

The synthesis further reveals a set of fundamental trade-offs that characterize explainability across data modalities. The most prominent is the trade-off between accuracy and interpretability, where highly accurate models tend to be less interpretable (Balcan & Sharma, 2021; Rudin, 2019) [3, 27].

A second trade-off exists between interpretability and scalability, particularly in large-scale and multimodal systems where explanation generation introduces computational overhead (Doshi-Velez & Kim, 2017; Samek *et al*., 2021) [10, 29].

A third trade-off involves explanation fidelity versus usability. High-fidelity explanations may be too complex for human interpretation, while simplified explanations risk misrepresentation (Marcinkevičs & Vogt, 2023; Kaur *et al*., 2020) [15, 22]. These trade-offs, summarized in Table 2, highlight the inherent constraints of current explainability paradigms.

4.4 Emergent Conceptual Patterns

Beyond individual patterns and trade-offs, a set of broader conceptual insights emerges from the cross-modality synthesis. First, there is a clear fragmentation of explainability paradigms, with different methods tailored to specific data types and model architectures (Burkart & Huber, 2021) [6]. Second, there is an absence of a unified explanation structure that can operate consistently across modalities.

Third, explainability exhibits modality-dependent behavior, where effectiveness varies significantly depending on data and model complexity.

Most importantly, there is a weak linkage between explanation and knowledge. Explanations often lack semantic clarity, fail to capture causal relationships, or require domain expertise for interpretation (Molnar, 2022; Pearl, 2019) [24, 25]. As reflected in **Table 2**, this disconnect represents af undamental limitation of current explainability approaches.

Table 2: Meta-Analytical Insights on Explainability Across Data Modalities

	Structured Systems	**Unstructured Systems**	**Hybrid / Multimodal Systems**	**Key Meta-Insight**
Explainability Type	Intrinsic + Feature-based	Post-hoc dominant	Hybrid (post-hoc + joint attribution)	Shift from intrinsic → approximation-driven
Model Transparency	High	Low (black-box)	Very low (cross-modal opacity)	Transparency decreases with complexity
Explanation Mechanism	Direct rules, coefficients	Feature attribution, visualization	Cross-modal attribution, attention	Increasing abstraction in explanations
Reliability	High (stable)	Moderate (method-dependent)	Low (inconsistent across modalities)	Reliability decreases across modalities
Semantic Interpretability	High	Limited (low-level features)	Fragmented (modality-dependent)	Weak semantic alignment in complex systems
Consistency	High	Variable	Low	Explanations become unstable with complexity
Scalability Impact	Low overhead	Moderate overhead	High computational cost	Complexity increases explanation cost
Explanation Fidelity	High (exact)	Approximate	Highly approximate	Fidelity decreases with model complexity
Knowledge Extraction Capability	Moderate	Low	Very low	Explanation ≠ actionable knowledge
Key Limitation	Limited expressive power	Lack of clarity & consistency	Fragmentation & ambiguity	Core issue: no unified explanation paradigm

4.5 Synthesis Summary

The cross-modality synthesis demonstrates that explainability in data mining systems is shaped by a complex interplay of model design, data modality, and interpretive mechanisms. While existing techniques provide valuable insights into model behavior, they remain fragmented, approximation-driven, and insufficiently connected to knowledge discovery. These limitations highlight the need for a unified conceptual framework that integrates explainability with the process of transforming data into actionable knowledge. The following section addresses this need by proposing the Explainability-Driven Knowledge Discovery (EDKD) framework.

Explainability-Driven Knowledge Discovery (EDKD): A Conceptual Framework

1. Framework Motivation

The meta-analytical synthesis reveals that existing explainability approaches are fragmented across data modalities, lack cross-modal integration, and exhibit a weak connection to knowledge discovery processes (Marcinkevičs & Vogt, 2023; Molnar, 2022) [22, 24]. While current techniques provide insights into model behavior, they do not systematically support the transformation of explanations into actionable knowledge. This limitation necessitates a unified framework that explicitly links data, models,

explanations, and knowledge, enabling a structured transition from interpretability to meaningful decision support.

2. Framework Architecture and Layered Abstraction

The proposed EDKD framework conceptualizes explainable data mining as a four-layered abstraction, capturing the progression from raw data to actionable knowledge:

1. Data Layer**:** Represents structured, unstructured, and hybrid data sources, characterized by heterogeneous representations and varying levels of complexity.
2. Model Layer**:** Encompasses machine learning, deep learning, and hybrid models operating across a spectrum of complexity and representational power.
3. Explanation Layer**:** Includes intrinsic and post-hoc explainability mechanisms, spanning local and global interpretations and exhibiting modality-dependent behavior.
4. Knowledge Layer**:** Transforms explanations into interpreted, contextualized, and decision-relevant knowledge, incorporating domain understanding and human reasoning.

3. Functional Flow

The framework follows a sequential transformation:

Data → Model → Explanation → Knowledge

At each stage, information undergoes abstraction and transformation. While models extract patterns from data, explanation mechanisms attempt to interpret these patterns. However, breakdowns occur at the explanation stage, where approximations, inconsistencies, and lack of semantic clarity hinder the transition to knowledge. The knowledge layer addresses this gap by integrating explanations with contextual and domain-specific interpretation, ensuring their applicability in decision-making.

4. Theoretical Contribution

The EDKD framework establishes a conceptual bridge between explainability and knowledge discovery, positioning knowledge as a distinct and necessary layer rather than an implicit outcome of explanation. It extends the traditional data mining pipeline by introducing a structured pathway that connects model outputs to actionable insights, thereby reframing explainability as a means for knowledge generation rather than an end in itself (Rudin, 2019) [27].

5. Framework Implications

The framework provides a foundation for standardizing explainability across modalities, enabling consistent interpretation in heterogeneous data environments. It supports cross-modality integration by aligning explanation mechanisms within a unified structure and offers a basis for future research in knowledge-aware and human-centered explainable systems.

Analytical Discussion: From Interpretability to Knowledge

1. Cross-Modality Comparative Insights

The comparative analysis across structured, unstructured, and hybrid systems reveals that explainability is inherently shaped by data modality and model complexity, leading to distinct interpretability profiles. Structured systems exhibit high interpretability due to explicit feature representations and transparent models, enabling direct mapping between inputs and outputs (Zaki & Meira, 2020; Molnar, 2022) [24, 34]. In contrast, unstructured systems rely on latent representations, where explainability is achieved through post-hoc approximation and often lacks semantic clarity (Samek *et al*., 2021; Linardatos *et al*., 2021) [17, 29]. Hybrid systems further complicate this landscape by introducing cross-modal dependencies, resulting in fragmented and inconsistent explanations (Baltrušaitis *et al*., 2019; Marcinkevičs & Vogt, 2023) [4, 22].

These differences are consolidated in Table 3, which shows that interpretability decreases progressively from structured to hybrid systems, while reliance on approximation and abstraction increases. Importantly, the capacity for knowledge extraction does not scale with predictive performance, indicating that more complex models do not necessarily yield more meaningful insights (Doshi-Velez & Kim, 2017) [10]. This imbalance highlights a fundamental limitation of current explainability paradigms, which prioritize model transparency over knowledge generation.

Table 3: Cross-Modality Comparison of Explainability Characteristics

Dimension	Structured Systems	Unstructured Systems	Hybrid / Multimodal Systems
Data Representation	Explicit, tabular, well-defined features	High-dimensional, latent representations (text, images)	Heterogeneous (tabular + text/image)
Model Types	Decision trees, linear models, ensembles	CNNs, RNNs, Transformers	Multimodal deep learning, hybrid pipelines
Interpretability Level	High (intrinsic)	Low (black-box)	Very low (cross-modal complexity)
Explanation Type	Intrinsic + feature-based	Post-hoc dominant	Hybrid (post-hoc + cross-modal attribution)
Explanation Methods	Rules, coefficients, SHAP	LIME, SHAP, Grad-CAM, attention	Multimodal attention, joint attribution
Explanation Scope	Global + local (consistent)	Mostly local (approximate)	Local + fragmented global
Semantic Clarity	High (aligned with domain features)	Moderate to low (feature abstraction)	Low (modality-dependent interpretation)
Consistency of Explanations	High (stable across instances)	Variable (method-sensitive)	Low (inconsistent across modalities)
Explanation Fidelity	High (exact or near-exact)	Approximate	Highly approximate
Scalability Impact	Low	Moderate	High computational overhead
Handling Feature Interactions	Limited but interpretable	Captured but not interpretable	Complex and ambiguous
Knowledge Extraction Capability	Moderate to high	Low	Very low
Key Strength	Transparency and simplicity	High predictive performance	Rich contextual representation
Key Limitation	Limited expressive power	Lack of semantic clarity	Fragmentation and attribution ambiguity

2. Interpretability vs Knowledge Gap

A central theoretical distinction emerging from this study is that interpretability does not equate to knowledge. While explainability techniques aim to make model behavior understandable, they do not inherently provide semantic meaning, causal reasoning, or actionable insight (Lipton, 2018) [18].

First, explanation does not guarantee understanding, as techniques such as saliency maps and feature attribution highlight important inputs without conveying contextual significance (Samek *et al*., 2021) [29]. Second, explanation does not imply causality, since most methods capture correlations rather than cause–effect relationships, potentially reinforcing spurious patterns (Pearl, 2019) [25]. Third, explanation does not ensure actionability, as translating outputs into decisions requires domain knowledge and contextual interpretation (Doshi-Velez & Kim, 2017) [10].

These limitations are further exacerbated by methodological constraints. Explanation methods often exhibit instability and sensitivity to perturbations, reducing reliability (Adebayo *et al*., 2018; Slack *et al*., 2020) [1, 31]. Additionally, human factors such as cognitive bias and limited interpretive capacity influence how explanations are understood and applied (Kaur *et al*., 2020) [15].

Collectively, these issues reveal a fundamental gap: current explainability approaches primarily describe model behavior rather than enabling knowledge discovery. This gap is particularly pronounced in unstructured and hybrid systems, where explanations are increasingly abstract and detached from domain meaning.

3. Trade-off Dynamics: Accuracy–Interpretability–Knowledge

The findings suggest a triadic trade-off between accuracy, interpretability, and knowledge utility, where improvements in one dimension often come at the expense of others. Highly accurate models, particularly deep learning systems, tend to be less interpretable, while interpretable models may lack predictive power (Rudin, 2019; Samek *et al*., 2021) [27, 29]. However, neither dimension guarantees knowledge generation.

A third dimension, knowledge utility, emerges as a critical but underexplored factor. Models may be accurate and interpretable yet fail to produce actionable insights if explanations lack context or relevance (Doshi-Velez & Kim, 2017; Molnar, 2022) [10, 24]. This triadic relationship highlights the need to move beyond binary trade-offs and consider explainability within a broader knowledge-centric framework.

4. Implications for Explainable Data Mining

The analysis has important implications for both theory and practice. From a theoretical perspective, it challenges the assumption that improving interpretability inherently enhances understanding, emphasizing the need to distinguish between explanation and knowledge (Marcinkevičs & Vogt, 2023; Linardatos *et al*., 2021) [17, 22]. From a practical standpoint, it suggests that model selection should consider not only accuracy and interpretability but also the capacity to generate meaningful, domain-relevant insights, particularly in complex and multimodal environments (Baltrušaitis *et al*., 2019; Kaur *et al*., 2020) [4, 15].

5. Synthesis with EDKD Framework

The limitations identified in this discussion reinforce the necessity of the proposed EDKD framework. By introducing knowledge as a distinct layer, the framework addresses the gap between explanation and actionable insight, enabling a structured transition from model outputs to decision-relevant understanding. In this context, explainability serves as an intermediate step rather than a final objective, supporting the broader goal of knowledge-driven data mining (Burkart & Huber, 2021; Molnar, 2022) [6, 24].

Conclusion

This study presented a meta-analytical and conceptual examination of explainability across structured, unstructured, and hybrid data mining systems, demonstrating that explainability is inherently modality-dependent. The analysis revealed a clear progression from intrinsic transparency in structured systems to approximation-driven explainability in unstructured and multimodal contexts, accompanied by increasing complexity, abstraction, and fragmentation. Key patterns identified include the growing reliance on post-hoc methods, declining interpretability with model complexity, and inconsistencies in explanation reliability. Importantly, the findings highlight a fundamental disconnect between predictive performance and knowledge extraction, indicating that more accurate models do not necessarily produce more meaningful or actionable insights.

The study contributes by providing a cross-modality synthesis of explainability, identifying core trade-offs and limitations, and proposing the Explainability-Driven Knowledge Discovery (EDKD) framework. By introducing knowledge as a distinct analytical layer, the framework extends traditional data mining pipelines and establishes a structured pathway from data to actionable insight. This reframes explainability as an intermediate step rather than an end objective.

A central conclusion is that explainability alone is insufficient for knowledge discovery. Meaningful knowledge requires the integration of explanation with interpretation, contextual understanding, and domain relevance. Future research should focus on developing unified evaluation metrics, addressing multimodal explainability challenges, incorporating human-centered evaluation, and advancing knowledge-aware machine learning systems that bridge the gap between model outputs and decision-making.

References

1. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. Advances in Neural Information Processing Systems,2018:31:9505–9515.
2. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion,2020:58:82–115.
3. Balcan MF, Sharma A. Data-driven learning and explainability. Communications of the ACM,2021:64:54–63.
4. Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE

Transactions on Pattern Analysis and Machine Intelligence,2019:41:423–443.
5. Belle V, Papantonis I. Principles and practice of explainable machine learning. Frontiers in Big Data,2021:4:688969.
6. Burkart N, Huber MF. A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research,2021:70:245–317.
7. Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 782–791.
8. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Medicine,2018:24:1559–1567.
9. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, 2019, 4171–4186.
10. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.
11. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management,2015:35:137–144.
12. Han J, Kamber M, Pei J. Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann, 2011.
13. Jain DK, Rahate A, Joshi G, Walambe R, Kotecha K. Employing co-learning to evaluate the explainability of multimodal sentiment analysis. IEEE Transactions on Computational Social Systems,2022:11:4673–4680.
14. Jain S, Wallace BC. Attention is not explanation. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019.
15. Kaur H, Nori H, Jenkins S, et al. Interpreting interpretability: Understanding data scientists' use of interpretability tools. In Proceedings of the CHI Conference on Human Factors in Computing Systems, 2020, 1–14.
16. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature,2015:521:436–444.
17. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy,2021:23:18.
18. Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue,2018:16:31-57.
19. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems,2017:30:4765–4774.
20. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence,2020:2:56–67.
21. Mai S, Zeng Y, Zheng S, Hu H. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. IEEE Transactions on Affective Computing,2022:14:2276–2289.
22. Marcinkevičs R, Vogt JE. Interpretable and explainable machine learning: A methods-centric overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,2023:13:e1440.
23. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports,2016:6:26094.
24. Molnar C. Interpretable machine learning (2nd ed.). Lulu.com, 2022.
25. Pearl J. The book of why: The new science of cause and effect. Basic Books, 2019.
26. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 1135–1144.
27. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence,2019:1:206–215.
28. Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys,2022:16:1–85.
29. Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE,2021:109:247–278.
30. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, 2017, 618–626.
31. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 180–186.
32. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In Advances in Neural Information Processing Systems,2017:30:5998–6008.
33. Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering,2013:26:97–107.
34. Zaki MJ, Meira W. Data mining and machine learning: Fundamental concepts and algorithms. Cambridge University Press, 2020.
35. Zhang Y, Choi M, Han K, Liu Z. Explainable semantic space by grounding language to vision with cross-modal contrastive learning. Advances in Neural Information Processing Systems,2021:34:18513–18526.