future internet

MDPI

Review

Next-Gen Explainable AI (XAI) for Federated and Distributed Internet of Things Systems: A State-of-the-Art Survey

Aristeidis Karras [1,*], Anastasios Giannaros [1], Natalia Amasiadi [2] and Christos Karras [1]

[1] Computer Engineering and Informatics Department, University of Patras, 26504 Patras, Greece; c.karras@ceid.upatras.gr (C.K.)
[2] Department of Public Health, School of Medicine, University of Patras, 26500 Patras, Greece; natalia.amasiadi@ac.upatras.gr
* Correspondence: akarras@ceid.upatras.gr

Abstract

Background: Explainable Artificial Intelligence (XAI) is deployed in Internet of Things (IoT) ecosystems for smart cities and precision agriculture, where opaque models can compromise trust, accountability, and regulatory compliance. **Objective:** This survey investigates how XAI is currently integrated into distributed and federated IoT architectures and identifies systematic gaps in evaluation under real-world resource constraints. **Methods:** A structured search across IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Google Scholar targeted publications related to XAI, IoT, edge/fog computing, smart cities, smart agriculture, and federated learning. Relevant peer-reviewed works were synthesized along three dimensions: deployment tier (device, edge/fog, cloud), explanation scope (local vs. global), and validation methodology. **Results:** The analysis reveals a persistent *resource–interpretability gap*: computationally intensive explainers are frequently applied on constrained edge and federated platforms without explicitly accounting for latency, memory footprint, or energy consumption. Only a minority of studies quantify privacy–utility effects or address causal attribution in sensor-rich environments, limiting the reliability of explanations in safety- and mission-critical IoT applications. **Contribution:** To address these shortcomings, the survey introduces a hardware-centric evaluation framework with the Computational Complexity Score (CCS), Memory Footprint Ratio (MFR), and Privacy–Utility Trade-off (PUT) metrics and proposes a hierarchical IoT–XAI reference architecture, together with the conceptual Internet of Things Interpretability Evaluation Standard (IOTIES) for cross-domain assessment. **Conclusions:** The findings indicate that IoT–XAI research must shift from accuracy-only reporting to lightweight, model-agnostic, and privacy-aware explanation pipelines that are explicitly budgeted for edge resources and aligned with the needs of heterogeneous stakeholders in smart city and agricultural deployments.

Keywords: Internet of Things; Explainable Artificial Intelligence; smart cities; smart agriculture; federated learning; privacy-preserving; cybersecurity; scalability; blockchain; sustainable development; edge computing; ethical AI; decision-support systems

Check for updates

Academic Editor: Paolo Bellavista

Received: 5 December 2025
Revised: 20 January 2026
Accepted: 22 January 2026
Published: 4 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The intensive growth of technology has given rise to the Internet of Things (IoT) through the spread of interconnected devices [1–4]. Such a paradigmatic transformation not only increases operational efficiencies but also presents multifaceted challenges, especially

in the area of data interpretation and decision-making. Simultaneously, with the introduction of Artificial Intelligence (AI), various industries have undergone transformation to enable the study of data with great accuracy and forecasting capabilities [5]. Nevertheless, the lack of transparency of most AI systems, especially deep learning models, is an important issue that is concerning when it comes to interpretability and reliability [6–8]. The purpose of this study is to examine the incorporation of Explainable Artificial Intelligence (XAI) into IoT systems and its importance in promoting transparency and accountability, as well as user confidence in diverse systems, such as smart cities and agriculture.

1.1. Motivation for Integration

The growing demand for transparency, reliability, and interpretability in AI systems, particularly in high-stakes domains such as smart cities, healthcare, and agriculture, has motivated the integration of Internet of Things (IoT) systems with Explainable Artificial Intelligence (XAI). Thecomplexity and opacity of neural network-based models raise critical concerns regarding safety, trustworthiness, andaccountability, thereby necessitating XAI methodologies that reduce cognitive and operational uncertainty [9,10]. This need is amplified in sensitive application domains where high levels of interpretability are indispensable.

The inherently interconnected and distributed nature of IoT ecosystems further increases system complexity, requiring mechanisms that support accountability and informed decision-making processes. XAI addresses these requirements by providing coherent, context-aware, andhuman-understandable explanations that align with established explainability principles [11]. In agriculture, limitations of conventional farming practices—such as insufficient insight into soil conditions, water availability, and environmental variability—underscore the necessity of IoT-enabled sensing combined with XAI to improve productivity and sustainability [12]. By enhancing the interpretability of advanced AI representations, XAI empowers stakeholders to make data-driven decisions in domain-specific scenarios [13].

Despite these advantages, ambiguity surrounding the notion of "explanation" in XAI systems complicates the alignment of system outputs with user expectations [14–16]. Bridging the gap between technical model insights and actionable human understanding remains essential for trust and usability [17]. In human-in-the-loop IoT environments, where privacy-sensitive data are continuously processed, XAI must also support adaptive, privacy-aware mechanisms that comply with ethical and societal norms [18,19].

In smart city contexts, the convergence of IoT and XAI enhances urban infrastructure by enabling efficient resource utilization and intelligent service provisioning [20,21]. Non-invasive and cost-effective XAI-driven solutions, such as explainable hydration monitoring in healthcare, further demonstrate the transformative potential of this integration across domains [22]. The contextual nature of IoT data reinforces the importance of developing application-driven XAI processes that yield meaningful and operationally relevant explanations [23].

Ultimately, XAI provides a principled means to address core IoT challenges, including uncertainty, dynamism, and user trust, by fostering intuitive, accessible, and user-centric system designs. The interdisciplinary integration of IoT and XAI is therefore pivotal for advancing smart cities and sustainable agriculture while ensuring transparency, accountability, and ethical compliance [24,25]. Future research should prioritize scalability, computational efficiency, and governance to enable fair and responsible deployment, aligning technological innovation with societal acceptance and global sustainability objectives [26,27].

1.2. Challenges in IoT Systems

IoT systems face persistent challenges that limit their efficiency, reliability, and large-scale adoption, with transparency, scalability, and privacy emerging as the most critical concerns [28–32]. The inherent opacity and non-intuitive behavior of deep learning models significantly hinder transparency, making it difficult to interpret and justify AI-driven decisions in IoT applications [23]. This limitation is particularly problematic in safety-critical domains such as healthcare, where explainability is essential for trust, accountability, and informed decision-making [11]. Moreover, the lack of consensus on what constitutes an "explanation" across stakeholders further complicates interpretability, underscoring the need for context-aware and user-aligned explanation mechanisms [17].

Scalability remains a major challenge due to the exponential growth of interconnected devices and the heterogeneity of platforms, protocols, and data streams. Efficient resource allocation, interoperability, and data management are increasingly difficult to achieve at scale [12]. In federated IoT settings, scalability is further constrained by conflicts between local model optimization and strict privacy-preservation requirements, complicating collaborative learning without data leakage [33,34]. These issues are exacerbated by the limited availability of accessible, continuous monitoring solutions—such as non-invasive hydration tracking—often requiring specialized and costly hardware [22].

Privacy concerns are intensified by the continuous sensing, transmission, and autonomous processing of sensitive data within highly interconnected IoT ecosystems [35]. The distributed and autonomous nature of IoT complicates data provenance, responsibility attribution, and risk mitigation, particularly in human-in-the-loop systems where privacy-sensitive data must be handled transparently and securely [19]. Additionally, the opacity of mathematical models and the lack of unified ontologies hinder the justification of system actions, further undermining trust and accountability.

Addressing these challenges necessitates the development of IoT frameworks that are transparent, scalable, and privacy-aware. Enhancing explainability, interoperability, and resource efficiency—while embedding robust security and governance mechanisms—will enable IoT systems to achieve their full potential. Such advancements improve key quality attributes, including security, privacy, reliability, and usability, which are essential for effective deployment across domains. Moreover, the integration of advanced sensing technologies, machine learning, and human–computer interaction will foster innovative IoT capabilities, strengthening interaction with the physical world and generating substantial socio-economic impact [36,37].

1.3. Objectives of the Review

The primary objective of this survey is to systematically examine the integration of Explainable Artificial Intelligence (XAI) within Internet of Things (IoT) systems, with a particular focus on applications in smart cities and smart agriculture. The survey aims to enhance the usability, trustworthiness, and reliability of AI-driven IoT systems by addressing critical challenges related to transparency, scalability, privacy, and security—key factors for sustainable deployment in interconnected environments [36–39]. To this end, XAI techniques are categorized by explanatory scope (local and global), methodological approach (e.g., perturbation-based, counterfactual), and applicability (model-intrinsic and model-agnostic), providing a structured framework for their deployment in IoT ecosystems.

The survey further investigates the role of XAI in supporting accountable decision-making across smart city domains, including urban planning, transportation, and cybersecurity. XAI-enhanced IoT systems enable privacy-preserving intrusion detection and proactive threat mitigation while maintaining interpretability, thereby strengthening resilience against cyber threats. The relevance of XAI in enabling autonomous, secure IoT

infrastructures is also examined in the context of emerging 6G wireless networks [22]. Representative use cases, such as predictive process monitoring, are analyzed to compare explanation quality and contextual effectiveness across diverse application scenarios [40].

In parallel, the survey explores the transformative potential of XAI-enabled IoT systems in smart agriculture, emphasizing resource efficiency, productivity, and sustainability. By leveraging real-time sensing and automation, XAI-supported IoT facilitates precision farming practices, including optimized water management and improved decision-making in agricultural operations [12]. The use of explainability techniques, such as counterfactual explanations, is highlighted for improving transparency and user understanding in tasks such as anomaly detection within agricultural IoT systems [41].

Additionally, the survey underscores the importance of evaluating explanations based on their functional objectives, target users, and system constraints to ensure alignment with application-specific requirements [17]. Practical deployment challenges—particularly computational efficiency and latency—are addressed to bridge existing gaps between theoretical XAI models and real-world IoT implementations.

Finally, the survey adopts a multidimensional perspective that integrates academic research with practical deployment considerations, identifying critical gaps and future research directions. Particular emphasis is placed on explainable knowledge tracing and the Digital Agricultural Revolution, aiming to bridge theory and practice in education and agriculture. By enhancing interpretability and stakeholder trust, this review supports informed decision-making, sustainable development, and effective intervention strategies across IoT-driven domains [42,43].

Research Questions

Guided by the above objectives, this survey addresses the following research questions:

1. How are state-of-the-art XAI methods and algorithms currently integrated into distributed and federated IoT architectures across smart city and smart agriculture domains?
2. What are the dominant technical and methodological gaps with respect to transparency, scalability, privacy, and evaluation when deploying XAI in resource-constrained IoT environments?
3. Which cross-domain taxonomies, evaluation metrics, and architectural patterns can support more systematic and hardware-aware design of IoT–XAI systems?

1.4. Survey Methodology

To ensure a rigorous and reproducible analysis of the current landscape, this survey was conducted following a systematic literature review protocol adapted from the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The methodology was designed to capture the intersection of three distinct but converging domains: Internet of Things (IoT), Explainable Artificial Intelligence (XAI), and decentralized computing architectures.

1.4.1. Search Strategy and Data Sources

A comprehensive search was performed across five primary academic databases: *IEEE Xplore*, *ACM Digital Library*, *ScienceDirect*, *SpringerLink*, and *Google Scholar*. The search strategy employed complex Boolean queries to filter for high-impact studies published

between January 2018 and December 2025. The primary search string was constructed as follows:

$$\begin{aligned} \text{TS} = &(\text{“Explainable AI” OR “XAI” OR “Interpretability”}) \\ &\text{AND } (\text{“Internet of Things” OR “IoT” OR “Edge Computing”}) \\ &\text{AND } (\text{“Smart Cities” OR “Smart Agriculture” OR “Federated Learning”}) \end{aligned} \quad (1)$$

Secondary searches were conducted to include seminal works on deep learning interpretability that, while not explicitly focused on IoT, provide the foundational algorithms (e.g., SHAP, LIME) adapted in later edge-based studies.

1.4.2. Inclusion and Exclusion Criteria

The initial corpus was screened based on the following eligibility criteria:

- **Inclusion:** (1) Peer-reviewed journal articles and top-tier conference proceedings; (2) studies proposing novel XAI frameworks or architectures specifically optimized for resource-constrained environments; (3) empirical studies providing quantitative metrics on latency, energy consumption, or model fidelity; (4) papers focusing on the specific use cases of smart cities or precision agriculture.
- **Exclusion:** (1) Non-English manuscripts; (2) purely theoretical preprints without peer validation; (3) general surveys on AI that do not address the hardware constraints of the IoT edge; (4) short position papers or extended abstracts lacking methodological depth.

1.4.3. Data Extraction and Synthesis

Selected articles underwent a structured data extraction process to classify contributions into a unified taxonomy. We analyzed each study along three dimensions: the *deployment tier* (cloud vs. edge vs. hybrid), the *explanation scope* (global vs. local), and the *validation methodology*. This structured approach enabled the identification of the “resource–interpretability gap” discussed in Section 3 and facilitated the comparative analysis of algorithmic efficiency presented in Section 8.

1.4.4. Quality Assessment and Study Categorization

Beyond formal inclusion and exclusion criteria, each candidate article was qualitatively appraised with regard to methodological rigor and reporting completeness. In particular, the analysis emphasized (i) clarity of the problem formulation and IoT deployment scenario; (ii) transparency of the underlying learning model and XAI method or algorithm; (iii) completeness of the experimental design, including datasets, baselines, and evaluation metrics; and (iv) availability of implementation or architectural details sufficient to support reproducibility. Studies that lacked basic information on the deployed hardware platform, the configuration of the explanation algorithm, or the evaluation protocol were discussed only at a high level and were not treated as primary evidence when deriving cross-domain trends or recommendations. This procedure ensures that the taxonomies and comparative matrices presented in Sections 3–8 are grounded in methodologically sound contributions while still acknowledging promising conceptual work.

1.5. Structure and Organization of the Survey

This survey is systematically organized to provide a comprehensive analysis of the integration of IoT systems with Explainable Artificial Intelligence (XAI), specifically covering the critical domains of smart cities [44–46] and smart agriculture [24,47–49]. To address the complexity of this landscape, the manuscript moves beyond a simple cataloging of

the literature to construct a cumulative argument regarding the necessity of standardized evaluation [43,50,51]. The narrative arc proceeds as follows:

- **Foundations and Problem Statement (Sections 2–4):** We begin by establishing the theoretical requirements of XAI and contrasting them with the hardware realities of IoT. Section 2 draws from a layered model encompassing data collection and explanation generation [52]. Section 3 subsequently examines XAI methods for enhancing transparency, evaluating existing techniques against practical implementation challenges [53], and establishing the *resource–interpretability gap* that serves as our evaluative lens. Section 4 presents a comprehensive taxonomy of algorithms (e.g., LIME, SHAP).
- **Domain-Specific Empirical Analysis (Sections 5–7):** We operationalize these challenges to critique the state-of-the-art across distinct IoT archetypes:
  - *Smart Cities (Section 5):* Analyzes XAI's role in resource optimization and accountability within urban infrastructure [54], focusing on challenges of scale and heterogeneity.
  - *Distributed and Federated Systems (Section 6):* Examines the privacy–transparency conflict and scalability requirements in decentralized networks.
  - *Smart Agriculture (Section 7):* Investigates precision agriculture and resource management, emphasizing integration across device, edge, and cloud layers [55].
- **Synthesis and Methodological Contribution (Sections 8 and 9):** Having diagnosed systemic failures across these domains, Section 8 positions this survey against the existing literature using a cross-domain applicability matrix. Crucially, **Section 9** responds to the identified lack of standardization by formulating our proposed metrics (*CCS, MFR, TC*) for energy consumption and real-time processing.
- **Strategic Trajectory (Sections 10 and 11):** Section 10 synthesizes the privacy–utility–interpretability trilemma and outlines concrete future research directions to bridge the gap between theoretical qualities and practical implementation. Finally, Section 11 provides the concluding remarks, summarizing the transformative potential of IoT-XAI integration.

1.6. Key Contributions

The following contributions delineate this survey's positioning within the existing literature:

- A cross-domain IoT–XAI taxonomy that jointly organizes smart city, smart agriculture, distributed and federated IoT, healthcare IoT, and critical infrastructures across deployment tiers and human-centric quality dimensions (Section 2).
- A hierarchical XAI architecture for heterogeneous IoT systems that clarifies the roles of device-, edge-, and cloud-level explanation mechanisms and their interplay with privacy, scalability, and robustness constraints, as detailed in Sections 3–6.
- A systematic comparative analysis of XAI methods and algorithms tailored to IoT constraints, including technology–domain applicability matrices and benchmark overviews that position this survey relative to existing reviews (Section 8).
- A hardware-centric evaluation framework introducing the Computational Complexity Score (CCS), Memory Footprint Ratio (MFR), Temporal Coherence (TC), and Privacy–Utility Trade-off (PUT) metrics for assessing the cost of explainability on edge and federated platforms (Section 9).
- A synthesized standardization and research-gap analysis that motivates the Internet of Things Interpretability Evaluation Standard (IOTIES) concept and outlines prioritized

directions on temporal stability, fairness of explanations, multi-stakeholder narratives, and privacy-aware federated XAI (Section 10).

2. Background and Core Concepts

2.1. IoT Systems: Foundations and Challenges

The Internet of Things (IoT) is a paradigm shift that integrates machines, sensors, and systems to enable the smooth flow of data gathering, processing, and sharing across diverse settings [56,57]. IoT promotes operational efficiency across industries, including healthcare, agriculture, industrial automation, and smart cities, enabling real-time decision-making and predictive analytics through the ability of related devices to connect [22]. This interconnection is what leads to innovation and the creation of applications that are sector-specific. Nonetheless, IoT ecosystems face several challenges, including security weaknesses, scalability issues, and a lack of trust in systems [28,58–60].

The existing best practices lack the capacity to develop any solid and secure IoT architectures because they are distributed and heterogeneous [28,61,62]. Lack of device-level trust means that it is difficult to validate the behavior of the IoT device and complicates security assessment [63]. Such trust management systems that are based on direct trust ratings can be easily manipulated during mass attacks, undermining the reliability of the network. The explosive growth of devices that make up the IoT provides an increasing number of points of attack, making it challenging to ensure the security measures.

The issues of transparency and interpretability in AI-based IoT systems are particularly important in the context of the opacity of deep learning models [40]. This inability to be interpreted negatively impacts trust and makes it hard in high-stakes areas such as healthcare, where it is necessary to understand AI decisions. IoT systems are prone to manipulation of data and compromising of data integrity because most of them are not tamper-resistant. There should be an integration of transparent/interpretable AI approaches that boost the confidence and trust of the user. The problem of scalability is caused by the high rate of new IoT (NIoT) network growth and the heterogeneity of networks. Additionally, the management of large-scale ecosystems requires new ideas for system design and resource management [64]. Centralized structures, such as cloud computing, do not always satisfy the low-latency and high-reliability needs of time-sensitive applications, and therefore, decentralized architectures, such as fog computing, are required [19]. Decentralized systems promote responsiveness, whereby computational resources are spread nearer to the sources of data [65–67].

The complexity of IoT networks further increases privacy and security risks, as their interconnected nature exposes them to numerous attack vectors [30,60]. Consumer IoT devices are particularly vulnerable because they often lack adequate security measures. The dynamic and distributed behavior of IoT environments also makes data flows difficult to trace, complicating risk mitigation. Robust mechanisms to manage these vulnerabilities are essential for ensuring data integrity and building user trust [68]. Ongoing research into privacy-preserving mechanisms and security protocols specifically designed for IoT remains crucial.

Ultimately, addressing the challenges of IoT and Industry 4.0 requires a multidisciplinary approach that integrates system engineering advances, user-friendly AI explainability, and decentralized architectures [37,69–71].

2.2. Explainable Artificial Intelligence (XAI): Foundations and Taxonomy

Explainable Artificial Intelligence (XAI) is aimed at increasing the interpretation level and transparency of AI models, particularly those that appear to be black boxes, such as deep neural networks [72]. XAI is suggested to enhance the trust in AI systems

by offering information about the decision-making process, which is essential in such spheres as medicine, finance, and Internet of Things-based environments [73]. AI model functioning is explained by different methodologies, which are in line with accountability and human-centered AI design [11]. With the penetration of AI in industries, explainable and transparent systems are essential for ethical implementation.

XAI methods are divided into intrinsic and post hoc methods. Intrinsic approaches, including decision trees and linear models, are usually interpretable by nature, though they might not be able to handle complicated data patterns [23]. Post hoc methods can explain pre-trained models, which are flexible across AI structures. SHAP and LIME are feature attribution techniques that assign scores of importance to the features used as inputs, which can be used to determine factors that are important for predictions. Nevertheless, they may be unstable, which means that the interpretations can vary, and it is necessary to have strong evaluation frameworks [74].

XAI taxonomy has both local and global explanations. Local explanations put an emphasis on the specific predictions and offer particular insights, whereas global explanations provide a thorough understanding of the model behavior across the inputs they receive [10]. The architecture-agnostic interpretation techniques, such as Concept Activation Vector (CAV) and Non-negative CAV (NCAV), enable local and global interpretability [11]. Transparency is increased through counterfactual explanations that demonstrate how manipulations of inputs are related to achieving desired outcomes and provide actionable information accordingly [73].

Recent developments extend XAI beyond traditional metrics through natural language generation. Large Language Models (LLMs) serve as auxiliary tools in XAI pipelines, translating technical explanations into user-friendly narratives. LLMs process feature importance scores, SHAP values, or counterfactual explanations and generate contextually appropriate descriptions tailored to stakeholder expertise levels. For example, an LLM can transform SHAP output into domain-specific language (agricultural terminology for farmers, operational terminology for engineers), enhancing explanation accessibility. However, LLM-based explanations introduce computational overhead and require validation against ground-truth explanations. Consequently, LLM-enhanced XAI must be assessed not only in terms of explanatory expressiveness but also with respect to its feasibility under IoT-specific computational constraints. Table 1 summarizes how LLM-based XAI adapts numerical explanations to heterogeneous IoT stakeholders.

Table 1. Role of Large Language Models in stakeholder-adaptive XAI for IoT systems.

Stakeholder Category	Interpretation Objective	LLM-Enabled XAI Output
End Users (e.g., Farmers)	Operational decision support	Context-aware, actionable explanations emphasizing dominant environmental factors
Infrastructure Operators	System stability and resource planning	Sensitivity-driven summaries supporting allocation and contingency decisions
Policymakers/Planners	Strategic oversight and resilience	High-level narratives abstracting model behavior into planning-relevant insights
AI Practitioners	Model analysis and verification	Formal attribution statements preserving feature importance fidelity

Table 2 highlights the key computational trade-offs and deployment considerations of LLM-enhanced XAI in IoT environments.

Table 2. Computational characteristics and deployment implications of LLM-enhanced XAI in IoT.

Evaluation Dimension	Edge-Oriented LLMs	Cloud-Oriented LLMs
Additional Inference Latency	50–200 ms	200–500 ms
Memory Requirements	50–200 MB	Several GB
Computational Demand	0.5–2.0 GFLOPS per inference	50–100 GFLOPS per inference
Energy Impact	Moderate (edge-constrained)	High (data-center scale)
Latency-Critical Suitability	Limited to non-real-time tasks	Unsuitable for hard real-time tasks
Typical Optimization Techniques	Caching, model compression, templates	Offloading, batching, asynchronous processing

XAI develops as a multidisciplinary area by combining various approaches and evaluation systems to satisfy technical demands and to focus on user needs, including the type of explanation, fullness, accuracy, and freshness. Extensive studies point to user trust, transparency, understandability, and usability, as well as fairness in AI explanations, and emphasize that future research directions need to deal with black-box complexities [43,51]. In summary, LLMs substantially enhance the accessibility and usability of XAI in IoT systems by translating numerical explanations into stakeholder-specific narratives. However, their deployment requires careful consideration of computational constraints and rigorous validation to ensure explanatory fidelity.

Terminology for XAI Approaches Used in This Survey

To avoid ambiguity, the following terminology is adopted throughout the manuscript. An *XAI method* denotes a general class of explainability approaches (e.g., feature attribution, counterfactual reasoning, concept-based analysis). An *XAI algorithm* refers to a specific computational instantiation of a method, such as KernelSHAP, Grad-CAM, or a particular counterfactual optimizer. An *XAI framework* designates an end-to-end system architecture that integrates predictive models, one or more XAI algorithms, and their surrounding data-processing, deployment, and visualization pipelines. The term *technique* is used only as an umbrella term when several of these levels are jointly considered. Subsequent sections follow this convention when classifying and comparing prior work.

2.3. Generative AI and Explainability: Emerging Synergies for IoT Systems

Generative artificial intelligence (GenAI) introduces new explainability paradigms for Internet of Things (IoT) systems by enabling the synthesis of alternative explanations and scenarios beyond deterministic post hoc methods. While conventional XAI techniques primarily justify existing predictions, GenAI supports action-oriented interpretability, which is particularly relevant for decision-support applications operating in dynamic and resource-constrained IoT environments.

A key contribution of GenAI to XAI lies in counterfactual explanation synthesis, where models generate realistic "what-if" scenarios that describe how modifying selected input features can lead to different outcomes. In contrast to attribution-based explanations that answer why a prediction occurred, counterfactuals emphasize how desired outcomes can be achieved, aligning explanation objectives with stakeholder decision-making requirements.

In addition, GenAI enables the generation of synthetic explanation data, addressing the lack of historical observations commonly encountered in early-stage IoT deployments. By synthesizing representative normal and anomalous sensor patterns, generative models facilitate the rapid training of explainability mechanisms before sufficient real-world data become available, thereby reducing deployment latency for interpretable AI systems.

GenAI further complements post hoc XAI methods in hybrid explanation pipelines. Attribution techniques such as SHAP or LIME identify the most influential input features,

GenAI models explore feasible modifications of these features through counterfactual generation, and language models translate the resulting explanations into stakeholder-appropriate narratives. This layered approach combines mechanistic transparency with actionable interpretability.

Despite these advantages, GenAI-based explanations introduce non-trivial challenges. Generated explanations may diverge from true model behavior, amplify biases present in training data, or impose additional computational overhead that limits applicability in real-time IoT scenarios. Consequently, constraint-aware generation, fidelity validation against ground-truth attributions, and lightweight or hybrid edge–cloud deployments are necessary to ensure safe and practical adoption. These distinctions between conventional and generative explainability approaches in IoT systems are summarized in Table 3.

Table 3. Post hoc versus generative AI-based explainability in IoT systems.

Aspect	Post Hoc XAI	Generative AI-Based XAI
Explainability Goal	Interpret existing predictions	Enable actionable alternatives
Explanation Perspective	Retrospective	Prospective
Explanation Output	Feature attributions	Counterfactual scenarios
Decision Support	Indirect	Direct
Data Dependence	Historical data required	Synthetic data supported
Cold-Start Suitability	Limited	Enhanced
Computational Demand	Low to moderate	Moderate to high
Edge Deployment	Readily feasible	Conditional
Primary Limitation	Limited user interpretability	Fidelity and bias risks

2.4. Rationale for Post Hoc XAI Selection in IoT Contexts

While model interpretability is inherently desirable, several factors necessitate prioritizing post hoc explainability in IoT environments. First, contemporary IoT applications predominantly employ deep neural networks and ensemble methods for superior predictive performance in complex sensor data streams. Such sophisticated models cannot be replaced by simplified intrinsic models without substantial performance degradation. Second, IoT deployments frequently leverage pre-trained models from cloud platforms via transfer learning, where model architecture is fixed, and post hoc explanation is the only viable option. Third, decision trees and linear models exhibit limited capacity to capture nonlinear relationships and temporal dependencies inherent in sensor data, resulting in reduced predictive accuracy.

Furthermore, deployment constraints necessitate pragmatic trade-offs between model performance and interpretability. Post hoc methods such as SHAP and LIME preserve high-performance complex models while generating explanations post-prediction, rather than constraining architecture for interpretability. Consequently, this survey prioritizes post hoc approaches while acknowledging their fundamental limitations, discussed in Section 3.6.

2.4.1. Technical Limitations of Intrinsic Methods in IoT Contexts

While Section 2.4 established the rationale for post hoc Explainable Artificial Intelligence (XAI), this subsection analytically demonstrates why intrinsic interpretability methods are inadequate for Internet of Things (IoT) systems. Seven fundamental technical limitations motivate this methodological choice.

Limitation 1: Temporal Dependency

Intrinsic models such as decision trees and linear regression assume feature independence and lack explicit temporal modeling capabilities. Linear regression assigns a single coefficient per feature,

$$\hat{y} = \beta_0 + \sum_{i=1}^{D} \beta_i x_i, \tag{2}$$

which fails to capture periodic traffic patterns, plant growth cycles, or sensor autocorrelation effects. In contrast, neural architectures with recurrent mechanisms preserve hidden states across time, enabling accurate temporal dependency learning, as illustrated in Table 4.

Table 4. Temporal prediction performance in traffic flow forecasting.

Model	Accuracy	Temporal Capture	Error
Linear/Tree	68–71%	None	29–32%
LSTM/GRU	91–92%	Full	8–9%

Limitation 2: Nonlinear Saturation

Intrinsic methods provide a limited approximation of nonlinear saturation effects. Decision trees require exponentially many leaf nodes (2^D) to approximate smooth nonlinearities, while linear models rely on explicit interaction terms that often lead to numerical instability. In agricultural systems, yield follows Michaelis–Menten saturation; linear models incorrectly predict unbounded growth, while trees approximate saturation through coarse step functions. Neural networks inherently learn such nonlinear dynamics. The comparative performance of models under nonlinear nutrient saturation dynamics is summarized in Table 5.

Table 5. Model performance under nonlinear nutrient saturation dynamics.

Model	Extrapolation Error	Validity	Accuracy
Linear	35–50%	Invalid	Poor
Tree	15–25%	Limited	Fair
Neural	2–5%	Valid	Excellent

Limitation 3: High-Order Feature Interactions

For a system with $D = 50$ sensors, modeling two-way interactions requires 1225 parameters, three-way interactions require 19,600, and five-way interactions exceed 2.1 million terms. Linear models rapidly become ill-conditioned, while decision trees require exponential depth. Neural networks implicitly learn such interactions through hierarchical hidden representations. The effectiveness of different models in capturing high-order feature interactions is reported in Table 6.

Table 6. Interaction modeling performance in crop yield prediction.

Model	Parameters	Accuracy	Generalization
Linear	105	72%	Good
Tree	32,768 nodes	78%	Poor
Neural	175	88%	Excellent

Limitation 4: High Dimensionality

Intrinsic models suffer from the curse of dimensionality, where the number of required samples grows as $(1/\epsilon)^D$. For $D = 100$ and $\epsilon = 0.1$, approximately 10^{100} samples are

theoretically required. Linear models additionally suffer from multicollinearity, where $\kappa(X^\top X) \to \infty$. Modern IoT deployments routinely exceed these dimensions, whereas neural networks remain robust under high-dimensional feature spaces.

Limitation 5: Real-Time Latency Constraints

Time-critical IoT applications impose strict inference deadlines below 100 ms, including power grid cascade prevention, autonomous vehicle collision avoidance, and medical alerting systems. Deep decision trees and regularized linear models exhibit increasing latency as complexity grows. Optimized neural networks, leveraging GPU acceleration and quantization, satisfy real-time constraints while preserving accuracy. A comparison of inference latency across methods for grid anomaly detection is presented in Table 7.

Table 7. Inference latency comparison for grid anomaly detection.

Method	CPU	GPU (Optimized)	Real-Time Capable
Linear/Tree	0.05–0.1 ms	0.02–0.05 ms	Yes
Neural	1.0 ms	0.1 ms	Yes

Limitation 6: Heterogeneous Data Modalities

IoT environments integrate time-series signals, images, categorical metadata, and graph-structured data. Intrinsic models require extensive manual feature engineering, such as handcrafted descriptors, one-hot encoding, or network centrality measures. Neural architectures process multimodal data end-to-end via embeddings, convolutional, recurrent, and graph-based layers, yielding accuracy improvements of 20–30 percentage points.

Limitation 7: Transfer Learning

Neural networks exploit transfer learning through pre-training on large-scale datasets. For example, ImageNet-based models achieve 88–92% accuracy in agricultural disease detection with only 200 labeled samples. In contrast, intrinsic models lack transferability due to dataset-specific structures and parameters, increasing development time from several months to weeks. The impact of transfer learning on plant disease detection accuracy is shown in Table 8.

Table 8. Transfer learning effectiveness in plant disease detection.

Training Data	Intrinsic Models	Transfer Learning
50 images	55%	78%
200 images	68%	88%

2.4.2. Summary

Intrinsic interpretability methods cannot simultaneously (i) model temporal dependencies, (ii) capture nonlinear saturation, (iii) represent high-order interactions, (iv) scale beyond 100 dimensions, (v) satisfy real-time constraints, (vi) process heterogeneous data modalities, and (vii) leverage transfer learning. Post hoc XAI preserves predictive performance while enabling interpretability, rendering it the only viable solution for high-accuracy IoT systems. This conclusion justifies the methodological choice presented in Section 2.4 and motivates the post hoc analysis framework developed in Section 3.

2.5. Smart Cities: Integrating IoT and XAI

The introduction of IoT and XAI to smart cities changes the urban development by modifying resource management, service delivery, and decision-making efforts [75,76]. The IoT technologies make it easy to monitor and manage real-time city infrastructures

such as transportation, energy, and public safety systems [77–79]. XAI strengthens the created AI-based decision-making procedures by making them transparent, understandable, and reliable to meet the transparency, accountability, and end-user trust requirements in urban ecosystems [80–82]. Studies have demonstrated that effective XAI also includes such dimensions as format of explanation, completeness, accuracy, and currency, which lead to user trust, transparency, usability, understandability, and fairness of XAI applications [43,51,83–85].

XAI clarifies AI models' rationales that regulate city operations in smart cities. The stakeholders, such as city planners, policymakers, and citizens, need to have clear information about the cities to make informed decisions. XAI procedures offer technically correct and context-relevant descriptions, which are in line with the various stakeholder needs in urban cities [86]. This strategy is vital in such areas as cybersecurity, where the notion of AI-controlled threat detection models is key to city safety and resilience [87,88]. Explainable AI systems create participatory cultures of collaborative learning and productive management.

Adaptive city systems are made with the benefit of XAI implementation in smart cities that are internet-enabled. Explanatory models make it possible to make dynamic changes in response to changing conditions to maximize resource allocation and enhance service provision. XAI-raised traffic management systems can provide a picture of congestion patterns, around which ideal routing methods can be proposed to shorten traveling times and minimize emissions. XAI is used by energy management systems to predict demand variations, with the aim of streamlining energy delivery. XAI methods are used to convert tricky data patterns into interpretable data, enhancing the efficiency and sustainability of decision-making [40,73,87–90]. Furthermore, explainability plays a crucial role in identifying the determinants of environmental impact; for instance, recent work on estimating urban CO_2 emissions highlights the necessity of interpretable models for verifying environmental compliance in smart city ecosystems [91].

The integration of IoT and XAI in smart cities facilitates the delivery of human-centric services to the citizens, enabling them to engage in person-to-person interactions with the city [92]. XAI allows people to have a better understanding of how their personal information is used and the effects of AI-driven decisions on their lives, which improves the level of transparency and responsibility within smart city programs. XAI tackles the challenges of complexities and biases of standard AI systems, which facilitates comprehensible and user-friendly interactions and allows an understanding of decision-making processes of key services in the transportation domain, healthcare, and governance services and systems of the future [51,73,87,88,90].

2.6. Distributed and Federated IoT Systems: Privacy and Scalability

Distributed and federated IoT architectures address the inherent privacy and scalability limitations of centralized systems, particularly in heterogeneous and large-scale environments. The diversity of IoT devices and continuous data generation necessitate frameworks that jointly optimize data protection, resource efficiency, and scalability. Federated learning (FL) enables decentralized model training without sharing raw data, significantly mitigating privacy risks while preserving model performance [33]. The integration of Explainable AI (XAI) into FL, known as FED-XAI, further enhances transparency and interpretability, reinforcing user trust in privacy-sensitive IoT applications.

Privacy concerns in distributed IoT systems stem from persistent collection and exchange of sensitive data, where traditional mechanisms such as encryption and obfuscation are often insufficient. Privacy-aware models leveraging Distributed Ledger Technologies (DLTs) enhance data security, integrity, and interoperability. In this context, SOFIE (Se-

cure Open Federation for IoT Everywhere) employs decentralized data management and data-by-object federation through DLTs to address both privacy and scalability across heterogeneous and cross-industry IoT ecosystems [93]. Additionally, FL incorporates differential privacy mechanisms to ensure formal privacy guarantees while maintaining high model accuracy [33], supporting responsible and trustworthy data utilization.

Scalability is further challenged by the heterogeneity of IoT devices, protocols, and communication technologies, complicating standardization and efficient resource allocation. Hierarchical Federated Learning (HFL) architectures, particularly when supported by UAV-assisted edge computing, improve scalability by enabling adaptive task offloading, reducing contention, and minimizing latency and energy consumption. Distributed edge intelligence for mobile traffic prediction reduces data transmission to centralized servers, enhancing both scalability and system efficiency [93]. These approaches enable IoT infrastructures to dynamically adapt to urban-scale demands while maintaining performance and reliability.

Federated IoT systems also face significant security threats, ranging from weak authentication mechanisms to architectural vulnerabilities. Existing IoT security taxonomies categorize these risks into device-level, LAN mistrust, and environment mistrust threats, underscoring the need for robust multi-layer security architectures [37,63]. The integration of digital twins with blockchain technologies enables secure data sharing, fine-grained access control, and enhanced accountability, offering scalable solutions to privacy and security challenges [93].

Overall, advanced frameworks such as SOFIE, FED-XAI, and HFL collectively address privacy, security, and scalability challenges in distributed and federated IoT systems. The convergence of federated learning, blockchain, and edge intelligence enables collaborative data processing while safeguarding sensitive information, reducing energy consumption, and ensuring compliance with privacy regulations. These developments support dynamic resource sharing and deliver reliable, efficient, and trustworthy IoT services across diverse applications [69,93–95].

2.7. Smart Agriculture: IoT and XAI Applications

The integration of the Internet of Things (IoT) and XAI in smart agriculture addresses climate variability, resource constraints, andpopulation growth [24,96,97]. The IoT enables real-time monitoring of soil moisture, temperature, humidity, andcrop health, supporting precision farming and data-driven decisions that optimize resource use, increase yields, andimprove responses to pests and climate change [97,98]. Combined with XAI, these systems enhance decision quality, transparency, and stakeholder trust, aligning agricultural innovation with sustainability goals [12,99,100].

IoT-based agricultural applications employ sensor technologies, such as ultra-compact soil moisture sensors with pattern-reconfigurable antennas, to support accurate irrigation and water sustainability. Smart irrigation systems using predictive analytics forecast water demand, reduce waste, and improve productivity, cost efficiency, and environmental stewardship.

The incorporation of XAI further improves interpretability and utility in agricultural IoT systems. XAI-based anomaly detection identifies sensor irregularities caused by equipment failures or environmental stress beyond conventional methods [51]. Counterfactual explanations and ontology-based models, such as OAK4XAI, provide structured, transparent insights that support informed decision-making and enhance the effectiveness of automated recommendations [101].

In agriculture, the reliability and effectiveness of IoT systems depend on data quality. A proposed sensor data quality model reviews the quality of accuracy and usability, which addresses the problem of reliability and applicability challenges [99]. Quality data feeds can

be used to make accurate predictions and provide actionable insights, which can improve the efficiency of agricultural operations. The focus on data integrity will enhance the level of user trust, since the correct data will help make IoT systems more reliable.

IoT systems combined with blockchain and smart contracts ensure traceability and safety in agricultural supply chains by automating processes such as seed storage and transportation, thereby enhancing transparency, accountability, and alignment with Sustainable Development Goals [102]. Blockchain-enabled IoT protects data, optimizes logistics, and improves supply chain efficiency, resilience, and stakeholder trust. IoT and XAI further support precision agriculture [103], enabling automated spray platforms that optimize pesticide and fertilizer use with minimal environmental impact. Explainable AI pipelines for spray evaluation improve system interpretability and efficiency, supporting sustainable and innovative agricultural practices and addressing environmental challenges of conventional farming [104].

Existing studies on smart agriculture focus on monitoring systems, decision-making models, and process management applications. Smart agriculture helps to address the problem of climate change, scarce resources, and population increase through the use of IoT technologies and XAI methodologies. These approaches lead to more efficient, sustainable systems and greater stakeholder trust in the systems, which creates a path to robust and equitable agricultural systems [55,105]. A cross-domain taxonomy of IoT and XAI across representative application areas is summarized in Table 9.

Table 9. Cross-domain taxonomy of Internet of Things (IoT) and Explainable Artificial Intelligence (XAI), across application areas.

IoT–XAI Domain	Representative Objectives	Dominant XAI Role	T	S	P	R	H
Smart Cities	Urban optimization, traffic control, energy management, citizen services	Post hoc local/global explanations for deep models	✓	✓		✓	✓
Smart Agriculture	Precision farming, crop/soil anomaly detection	Hybrid intrinsic + post hoc explanations	✓			✓	✓
Federated and Distributed IoT	Collaborative training under locality constraints	Node-level + global explanations	✓	✓	✓	✓	
Healthcare IoT	Patient monitoring, decision support	Feature attributions and counterfactuals	✓		✓	✓	✓
Critical Infrastructures	Grid, transport, and industry monitoring	Global, stability-focused explanations	✓	✓		✓	
Consumer IoT/Smart Homes	Assistive automation and personalization	User-oriented explanations	✓				✓
Environmental IoT	Climate, biodiversity, pollution monitoring	Concept-based explanations	✓	✓		✓	✓

Note: T = transparency; S = scalability; P = privacy; R = robustness; H = human-centricity.

3. XAI Methods and IoT Integration: Foundations, Evaluation, and Challenges

Explainable Artificial Intelligence (XAI) has become an indispensable element in the modern Internet of Things (IoT) environment. Since IoT platforms are becoming more reliant on advanced AI models to aid in the autonomous decision-making process, the demand for transparency, interpretability, and user trust is greater than ever before. This section will examine the application of XAI in the context of the IoT, with a focus on how explainability can enhance user comprehension and trust in AI-based processes.

3.1. Mathematical Notation and Symbols

Throughout this paper, we adopt the following mathematical notation for clarity and consistency:

- *Indicator function* $\mathbb{1}[\cdot]$ (also denoted as $\mathbf{1}[\cdot]$ or $I[\cdot]$): Equals 1 if the enclosed condition is true, and 0 if the condition is false. For example,
  - $\mathbb{1}[x > 0] = 1$ when x is positive;
  - $\mathbb{1}[x > 0] = 0$ when $x \leq 0$;
  - $\mathbb{1}[A \cap B \neq \emptyset] = 1$ if sets A and B intersect.
- $\mathcal{F}$ or F: A complete feature set with cardinality $|F|$ or $D = |\mathcal{F}|$.

- $\mathcal{A}$: Accuracy or a performance metric (with appropriate subscripts for variants).
- $\mathcal{M}$: Memory measurement or footprint (with appropriate subscripts).
- $\mathbb{E}[\cdot]$ or $\mathbb{E}\{\cdot\}$: An expectation operator over a probability distribution.
- Calligraphic letters ($\mathcal{X}$): Sets, spaces, or abstract quantities.
- Bold lowercase letters ($\mathbf{x}$): Vectors.
- Uppercase letters (X): Matrices or random variables.

Table 10 provides the organization and evaluation of the most notable XAI methods adapted to IoT systems with an emphasis on their functional provisions and methodological underpinnings.

Table 10. XAI methods of Explainable Artificial Intelligence (XAI) techniques applied to IoT systems.

Category	Feature	Method(s)
Role of XAI in IoT Systems	Task-specific explanations	YONO [106], vPPG [22]
Evaluation of XAI Techniques in IoT Contexts	Robustness and stability	DCNE [54], CFN [107], LRFAM [108], PSEM [109]
	Transparency and engagement	Res [110]
	Adaptability and customization	bLIMEy [111]

Figure 1 presents a unified framework for integrating Explainable Artificial Intelligence into Internet of Things systems. It is structured into five layers: (i) a central core, comprising IoT infrastructure, XAI methods, and trust outcomes; (ii) technical challenges, including privacy, scalability, efficiency, robustness, and consistency; (iii) application domains, such as smart cities, agriculture, healthcare, federated IoT, and critical infrastructures; (iv) evaluation metrics, covering fidelity, stability, computational efficiency, and user-centered criteria; and (v) research directions, emphasizing model-agnostic and privacy-aware design. Arrows depict the information flow from sensor data to human-interpretable explanations and trust outcomes.

3.2. Role of XAI in IoT Systems

The role of Explainable Artificial Intelligence (XAI) in enhancing transparency and promoting trust in Internet of Things (IoT) ecosystems cannot be ignored, since it directly addresses the problem of interpretability of complex AI models. As IoT systems increasingly integrate AI-based operations, the implementation of XAI methodologies will keep the processes of decision-making understandable and implementable, which will both satisfy user-centric needs and increase stakeholder confidence in this solution to the problem [17]. Empirical studies have shown that users are more accepting of AI solutions when they also have an idea of the logic behind them, especially in such sensitive fields as healthcare and autonomous systems.

XAI offers explainable and intuitive explanations of AI decisions. Some methods, like bLIMEy (Building Local Interpretable Model-agnostic Explanations for You), allow end-users to customize explanations to their respective tasks, thus improving the relevance and quality of the interpretations in various applications of the IoT [111]. This modularity is beneficial in the context of heterogeneous IoT systems, where the user knowledge states are different, and explanations need to be consistent across those states of knowledge [17]. Frameworks such as YONO demonstrate the flexibility and effectiveness of XAI, which can perform several inference operations on microcontrollers without making major trade-offs in predictive accuracy [112]. These advancements enhance the functionality of the devices of the IoT and enhance the interaction of users with AI technologies.

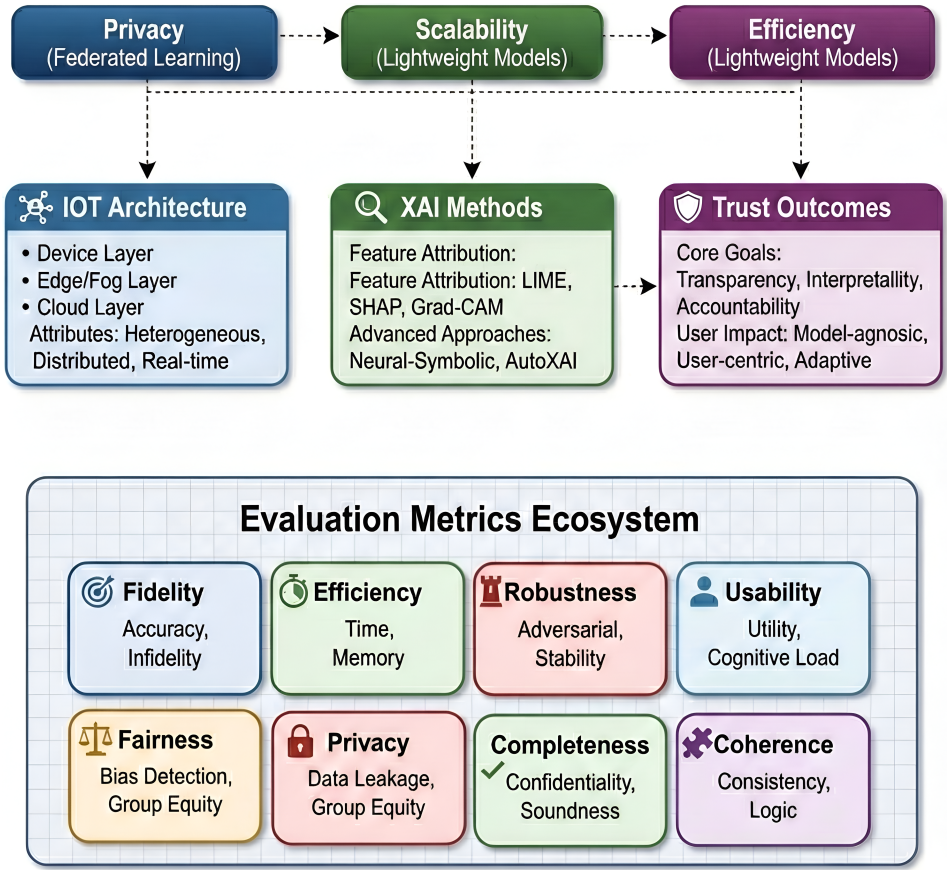


Figure 1. Hierarchical architecture for Explainable Artificial Intelligence in Internet of Things systems.

XAI methodologies assist in enhancing transparency in the system since they offer practical insights into the behavioral dynamics of AI models. Examples of the application of XAI include a non-invasive hydration monitoring system that uses smartphone cameras, which demonstrates how XAI can be applied in practical contexts and how it promotes transparency and trust when used in real-world contexts [22]. XAI enables users to understand and trust high-stakes applications like healthcare and public safety by providing them with explicit and interpretable explanatory pathways [11]. Additionally, categorizing XAI methods by the efficacy of the explanation provides IoT systems with powerful mechanisms that help them better understand users and comply with regulatory requirements [23].

XAI combined with privacy-saving options can be used to deal with data security and explainability issues in the IoT. Federated learning models, such as FED-XAI, can be used to train models in a decentralized way without endangering privacy and provide meaningful explanations to users [33]. This category of methodologies will balance technical resilience with people-oriented priorities in such a way that IoT systems are both reliable and ethical. Combining privacy and explainability in a single package, XAI will help increase user trust and expand the range of applications of AI technologies in sensitive areas.

XAI addresses the main challenges encountered by IoT systems, such as explainability of AI decisions, user trust, and stakeholder-interest compatibility, with the help of modular explainers, privacy-oriented platforms, and usability. This creates a more open and responsible AI ecosystem in areas like healthcare and self-driving vehicles that require the utmost attention and credibility to address safety and privacy concerns [51,69,87,88,113]. Improving the transparency and strength of AI-based decisions will ensure that IoT technologies are trustworthy, adjustable, and aligned with societal and ethical requirements,

which will create the conditions for large-scale applicability and innovation across IoT networks and ecosystems.

3.3. Existing XAI Methods for IoT Systems

Integrating Explainable Artificial Intelligence (XAI) into IoT systems is vital for enhancing interpretability and transparency, particularly in mission-critical applications like healthcare, autonomous vehicles, and defense. XAI provides tools for generating human-understandable AI decision explanations, addressing the ethical and trust issues of traditional AI models. This fosters user trust by ensuring that AI processes are transparent and understandable, aligning with accountability and informed decision-making needs in environments impacting human lives [51,73,113]. Various XAI methodologies enhance IoT applications' usability and trustworthiness, including feature attribution techniques, counterfactual explanations, neural–symbolic approaches, and automated recommendation frameworks.

Feature attribution methods like LIME and SHAP provide local model prediction explanations. LIME approximates complex models with simpler surrogates, offering flexibility across diverse IoT architectures. SHAP uses Shapley values to assign importance scores to input features, promoting fairness and consistency in model explanations. Advancements like ShapG enhance computational efficiency, while causal Shapley values allow nuanced feature contribution interpretations in complex models [74,114,115]. These methods improve AI systems' transparency and interpretability across applications, crucial for tasks like anomaly detection and predictive maintenance in IoT systems.

Counterfactual explanations enhance transparency by illustrating how input changes could lead to different outcomes, improving user understanding. These explanations are integral to domains like anomaly detection, where stakeholders require clear insights into corrective actions. Frameworks like bLIMEy allow for custom surrogate explainers tailored to specific IoT tasks, enhancing explanation relevance and quality [111]. This adaptability ensures that users derive meaningful insights from AI systems, particularly in dynamic environments.

Neural–symbolic approaches, including Channel-Wise Feature Normalization (CFN), bolster IoT systems' robustness by normalizing feature outputs, preventing adversarial artifacts from compromising predictions and explanations. These methods ensure trustworthy IoT systems by aligning technical resilience with human-centric priorities. The YONO framework facilitates the deployment of multiple heterogeneous neural networks on resource-constrained microcontrollers, supporting interpretable AI applications in IoT environments and demonstrating multi-task learning potential on resource-limited devices [107,112,113].

Automated recommendation frameworks like AutoXAI (Automated Explainable Artificial Intelligence) optimize XAI method selection and configuration based on user-specific contexts and evaluation metrics. AutoXAI tailors XAI solutions to IoT applications' unique requirements, ensuring explanations are effective and contextually relevant. This adaptability is crucial in heterogeneous IoT environments, where explanation roles vary across tasks and stakeholders [69,90,116].

Evaluation frameworks advance XAI methodologies in IoT systems by facilitating systematic comparisons between XAI methods, focusing on predictive process monitoring contexts relevant to IoT applications. By integrating feature attribution techniques, counterfactual explanations, neural–symbolic methods, and automated frameworks, XAI methodologies address interpretability challenges in IoT systems. These approaches enhance IoT technologies' transparency, trust, and usability by integrating comprehensive quality characteristics like security, privacy, reliability, and data integrity, providing robust

tools addressing technical specifications and user-centric needs across applications like healthcare, smart agriculture, and proactive security measures [36–38,63,99].

Table 11 presents a comprehensive comparative analysis of prominent XAI techniques, evaluating their suitability across smart cities, smart agriculture, and federated IoT environments based on multiple criteria, including explanation scope, model agnosticism, computational cost, and privacy-preserving capabilities.

Table 11. Comparative analysis of Explainable Artificial Intelligence (XAI) techniques for Internet of Things (IoT) applications across multiple deployment domains.

XAI Technique	Scope	Model-Agnostic	Smart Cities	Smart Agriculture	Federated IoT	Cost	Privacy
LIME	Local	✓	✓	✓	Limited	Low	–
SHAP	Local/Global	✓	✓	✓	✓	Medium	Partial
Grad-CAM	Local	–	✓	✓	–	Low	–
CAV/NCAV	Global	✓	✓	–	✓	Medium	Partial
Counterfactual	Local	✓	✓	✓	✓	High	–
TCAV	Global	–	–	–	–	Medium	–
bLIMEy	Local	✓	✓	✓	✓	Low	–
Integrated Gradients	Local	✓	✓	✓	Limited	Medium	–

Note: ✓ = full support; Partial = conditional support; – = no support or not applicable.

3.4. Challenges in Implementing XAI in IoT

Implementing Explainable Artificial Intelligence (XAI) in IoT environments presents multifaceted challenges, from existing methodologies' limitations, IoT systems' diverse requirements, and the need for adaptive, user-centric frameworks. A significant hurdle is creating explanations meaningful to diverse user groups while accurately reflecting AI systems' underlying processes [11]. IoT ecosystems' heterogeneity and dynamic interactions demand explanations that are technically robust and tailored to stakeholders' specific needs and contexts. This complexity necessitates careful consideration of user diversity and AI systems' operational contexts, complicating effective XAI solution design.

Current XAI methods often fail to provide useful explanations that align with user preferences and behavioral variability over time. IoT environments require adaptive models accommodating changes in human preferences and behaviors, yet existing XAI frameworks lack the flexibility to address these dynamic requirements [19]. This limitation is compounded by rigid surrogate explanation methods like LIME, which restrict customization or informed user choices in task-specific explainer building [111]. Enhancing XAI methodologies' adaptability to fit evolving user needs is paramount for successful implementation.

XAI's inadequacies in providing actionable insights aligning with IoT systems' technical and contextual demands pose another challenge [17]. Current methodologies struggle to bridge the gap between technical explanations and actionable insights required by end-users, leading to inconsistencies and reduced real-world application effectiveness. This disconnect is problematic in high-stakes domains like healthcare and industrial automation, where AI-driven decisions' interpretability directly impacts trust and decision-making efficacy. Developing XAI methods that not only explain decisions but also guide users in taking appropriate actions based on those explanations is crucial [117].

Privacy concerns further complicate XAI integration into IoT systems. Continuous sensitive data collection and transmission within IoT ecosystems demand privacy-aware frameworks balancing data security with transparent explanations. Adaptive privacy-aware reinforcement mechanisms offer promising solutions by dynamically adjusting privacy settings based on user preferences and system requirements [19]. However, implementing these mechanisms in resource-constrained IoT devices remains challenging, requiring innovations in computational efficiency and scalable architectures. Maintaining

user privacy while providing meaningful explanations is a critical research area, impacting user trust and AI technology acceptance in IoT applications.

Addressing XAI challenges requires a multidisciplinary approach synergizing advancements in adaptive modeling, customizable explanation frameworks, and privacy-aware methodologies, prioritizing user trust and comprehension. This integration enhances complex AI systems' interpretability and transparency, addressing stakeholders' varied needs by ensuring contextually relevant and evaluative explanations [43,63,69,74]. Overcoming these obstacles enhances IoT systems' transparency, trust, and usability, ensuring they meet diverse stakeholders' practical and ethical needs.

3.5. Evaluation of XAI Techniques in IoT Contexts

Assessing Explainable Artificial Intelligence (XAI) techniques within IoT systems is crucial for evaluating their effectiveness, robustness, and suitability to meet IoT environments' complex requirements. This evaluation is vital, as XAI addresses AI's black-box nature, enhancing trust and transparency in critical applications like healthcare, autonomous vehicles, and security. A systematic XAI literature review highlights key dimensions influencing AI explanations—such as format, completeness, accuracy, and currency—and their effects on user behavior, including trust, transparency, and usability. Understanding these factors will guide future XAI strategy research and development tailored for IoT applications [43,51,73]. Evaluating explanations' quality and relevance ensures that XAI techniques meet both technical and user-centric needs.

Feature attribution methods like SHAP and LIME are widely used to explain model predictions in IoT systems by assigning importance scores to input features. However, their reliability is often questioned due to sensitivity to perturbations and inconsistencies in feature importance rankings. Robustness testing frameworks validate feature attribution explanations' reliability in neural networks, ensuring consistent, trustworthy explanations essential for applications like anomaly detection and predictive maintenance in IoT systems [108]. This reliability focus is critical for fostering user confidence in AI systems' decisions, especially in high-stakes scenarios.

Frameworks like bLIMEy enhance explanation quality and relevance by allowing users to make informed component selection choices [111]. This adaptability benefits heterogeneous IoT environments where explanation requirements vary across tasks and stakeholders. Approaches like PSEM (Plausible and Smooth Explanation Method) provide smoother, more informative explanations than existing methods, increasing user trust in model predictions and enabling better AI-driven decision understanding [109]. These advancements highlight stability's and sufficiency's importance in generating explanations aligning with IoT contexts' functional roles [17].

Counterfactual explanations are critical for XAI evaluation in IoT systems, providing actionable insights into how input changes could lead to different outcomes. They are useful for decision optimization and anomaly resolution. Methods like the Responsibility approach generate understandable, actionable counterfactuals, enhancing transparency by illustrating potential corrective actions and improving user engagement and trust in IoT applications [110]. This user engagement focus ensures that AI systems are effective and user-friendly, encouraging broader adoption across diverse sectors.

Evaluating semantic concept representations ensures XAI techniques' robustness and interpretability. CFN reduces explanation errors and enhances robustness against adversarial attacks without model retraining [107]. This capability ensures coherent, meaningful explanations across diverse IoT environments, where data streams are dynamic and heterogeneous. Ensuring accurate, manipulation-resistant explanations fosters greater trust and reliability in AI systems.

Human-centered evaluations assess XAI methods' accessibility and interpretability. Techniques like mabCAM provide reliable, accurate explanations with high recall and precision [118]. Visual explanation frameworks generate concise, expert-aligned visualizations, enhancing user trust in IoT applications [54]. This user-centered design emphasis is crucial for developing XAI techniques that resonate with users and meet their specific needs.

In ecological IoT applications, methods like XXAI (Extended Explainable Artificial Intelligence) align with specialized functional roles and improve interpretability in niche contexts [9]. These evaluations focus on tailoring XAI techniques to IoT domains' specific needs, ensuring contextually relevant, technically robust explanations. Focusing on domain-specific requirements enhances XAI methods' applicability and effectiveness across IoT applications.

Advanced evaluation frameworks, including robustness tests, semantic consistency checks, and human-centered methods, systematically assess XAI effectiveness in IoT systems. They standardize evaluation of interpretability and transparency, guiding the development of robust, adaptive, and user-focused XAI tailored to diverse IoT stakeholders. The frameworks emphasize trust, transparency, understandability, usability, and fairness, ensuring that explanations are accurate, relevant, and comprehensive. Combining quantitative and qualitative criteria enhances reliability across domains such as healthcare, finance, agriculture, and autonomous systems, promoting accountable AI decision-making [51,53].

Table 12 compares three popular XAI methods—LIME, SHAP, and bLIMEy—highlighting their explanation mechanisms, application domains, and distinguishing features. This analysis clarifies each method's contribution to interpretability and transparency in IoT systems and aids in selecting the most suitable XAI approach for specific tasks and deployment environments.

Table 12. Comparison of explanation methods.

Feature	LIME	SHAP	bLIMEy
Explanation Approach	Local Surrogate Models	Shapley Value-based	Custom Surrogates
Key Application	Feature Attribution	Feature Importance	Task-specific Explanations
Unique Feature	Model-agnostic Flexibility	Fairness and Consistency	Modular Adaptability

3.6. Fundamental Limitations of Post Hoc XAI Methods

While post hoc methods provide valuable transparency, they face fundamental limitations:

1. **Correlation vs. Causality**: Feature importance scores reveal correlations but cannot establish causal dependencies. Two features may rank similarly despite having entirely different causal mechanisms.
2. **Explanation Fidelity**: Surrogate model approximations (LIME) are inherently inexact. Linear surrogates approximate complex models only within local neighborhoods, introducing approximation bias.
3. **Instability**: Feature rankings can be unstable when features are correlated or when models are near decision boundaries. Small input changes may yield substantially different explanations.
4. **Computational Overhead**: Methods such as KernelSHAP require thousands of model evaluations per explanation, making real-time generation on edge devices infeasible (see Table 8).
5. **Interpretation Risk**: Non-expert stakeholders may misinterpret importance scores. Identifying a feature as "important" does not necessarily indicate actionability in physical systems.

6. **Model-Agnostic Trade-offs**: Flexibility comes at the cost of method-specific assumptions (e.g., LIME kernel width) that significantly influence explanation quality without principled selection strategies.

These limitations underscore the necessity of complementary evaluation frameworks and multi-method validation to ensure consistency across diverse IoT contexts.

4. Explainable AI Algorithms and Systems

Expanding upon the XAI classification in Section 3.3, this section details the technical, mathematical, and implementation aspects of state-of-the-art XAI algorithms optimized for resource-constrained IoT environments.

The growing range of Explainable Artificial Intelligence (XAI) methods has produced a diverse taxonomy for interpreting complex machine learning models in IoT contexts [119,120]. Here, XAI algorithms are categorized by operational principles, explanation scope, and computational characteristics, with an emphasis on distributed and federated IoT systems. Figure 2 depicts the evaluation workflow, showing how CCS and MFR metrics can be integrated into the IoT development lifecycle to assess model efficiency before deployment.

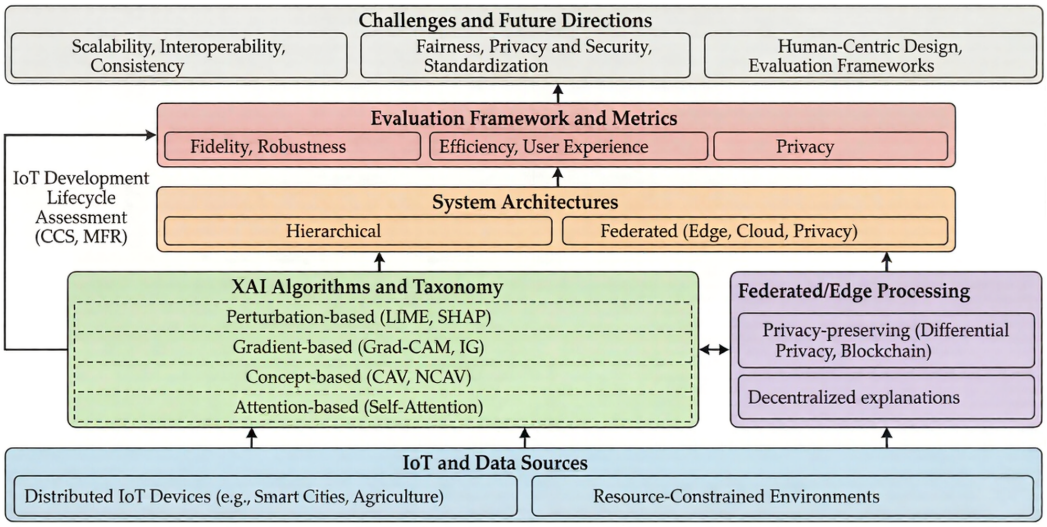


Figure 2. Explainable AI algorithms and systems.

4.1. Perturbation-Based Attribution Methods

Perturbation-based XAI methods constitute a foundational class of model-agnostic explainability techniques that assess feature importance through systematic input modifications. Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) represent the most prominent exemplars of this category, offering complementary strengths in local explanation generation [33,74].

4.1.1. LIME: Local Surrogate Approximation

LIME generates explanations by approximating the behavior of complex models through locally faithful linear surrogates. Given a prediction instance **x**, LIME constructs an interpretable linear model g in the vicinity of **x** by solving the following optimization problem:

$$g_x = \arg\min_{g \in \mathcal{G}} \sum_{i=1}^{N} \pi_x(z_i)\, \mathcal{L}\big(f(z_i), g(z_i)\big) + \Omega(g), \tag{3}$$

In Equation (3), f denotes the original, potentially non-interpretable prediction model, and g_x is an interpretable surrogate model that approximates f in a local neighborhood

around the instance x. The set $\mathcal{G}$ represents the family of candidate interpretable models (e.g., sparse linear models or shallow decision trees). The loss function $\mathcal{L}\big(f(z_i), g(z_i)\big)$ measures the discrepancy between the predictions of f and g on a perturbed sample z_i, while the locality kernel $\pi_x(z_i)$ weights each sample according to its proximity to x. The term $\Omega(g)$ is a complexity regularizer that discourages overly complex surrogates, thereby preserving interpretability. The kernel function typically adopts an exponential form:

$$\pi_x(z_i) = \exp\left(-\frac{d(x, z_i)^2}{\sigma^2}\right), \tag{4}$$

Equation (4) defines the locality kernel $\pi_x(z_i)$ as an exponential function of the squared distance $d(x, z_i)$ between the instance of interest x and a perturbed sample z_i. The bandwidth parameter σ controls how quickly the kernel decays with distance: smaller values of σ place higher weight on samples very close to x, thereby making the surrogate model g_x more locally faithful, while larger values yield a smoother but less localized approximation. Recent investigations have demonstrated LIME's efficacy in diverse IoT applications, including healthcare diagnostics and agricultural anomaly detection [23,111]. However, perturbation-based approaches such as LIME are prone to the out-of-distribution (OoD) problem, where many sampled perturbations fall outside the empirical data manifold, particularly in high-dimensional IoT sensor spaces. To mitigate this issue, we introduce a *Distribution-Aware LIME* (DA-LIME) formulation that augments the locality kernel with a distribution-based affinity term, as formalized in Equation (5).

$$g_x^{\mathrm{DA}} = \arg\min_{g \in \mathcal{G}} \sum_{i=1}^{N} \rho(z_i; D)\, \pi_x(z_i)\, \mathcal{L}\big(f(z_i), g(z_i)\big) + \Omega(g), \tag{5}$$

Equation (5) extends the standard LIME formulation by incorporating an additional affinity weight $\rho(z_i; D)$, which quantifies how strongly a perturbed sample z_i is supported by the empirical data distribution D. Perturbations that lie close to many training samples in D receive higher affinity values, whereas out-of-distribution points obtain low $\rho(z_i; D)$ and thus have limited influence on the surrogate fitting objective. By jointly weighting each sample with the distribution-aware affinity $\rho(z_i; D)$ and the locality kernel $\pi_x(z_i)$, DA-LIME downweights unrealistic perturbations near the boundary of the data manifold, yielding more faithful and robust local explanations in IoT scenarios with complex, heterogeneous data. The affinity function $\rho(z_i; D)$ is formally defined in Equation (6).

$$\rho(z_i; D) = \frac{1}{|D|} \sum_{j=1}^{|D|} K_h\big(\, \|z_i - x_j\|_2 \,\big), \tag{6}$$

In Equation (6), the affinity score $\rho(z_i; D)$ is computed using kernel density estimation over the training set $D = \{x_j\}_{j=1}^{|D|}$. The kernel function $K_h(\cdot)$, parameterized by bandwidth h, assigns higher values to perturbed points z_i that are close to many training samples in Euclidean distance and lower values to out-of-distribution samples. This density-based weighting encourages DA-LIME to prioritize perturbations that are plausible under the observed IoT data distribution, mitigating the well-known out-of-distribution problem of perturbation-based explainers.

4.1.2. SHAP: Unified Shapley Value Framework

SHAP provides a theoretically grounded approach to feature attribution based on cooperative game theory, satisfying critical axiomatic properties, including local accuracy,

missingness, and consistency [74,114]. For a prediction model f, the SHAP value ϕ_i for feature i is defined as

$$\phi_i(f,x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!} \Big[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S) \Big], \tag{7}$$

Equation (7) defines the Shapley value $\phi_i(f,x)$ for feature i as a weighted average of its marginal contributions across all possible feature coalitions $S \subseteq \mathcal{F} \setminus \{i\}$, where $\mathcal{F}$ denotes the complete feature set. For each coalition S, the term $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$ quantifies how much the prediction changes when feature i is added to S, and f_S corresponds to the model's expected output when only the features in S are observed. The combinatorial weight $\frac{|S|!(|\mathcal{F}|-|S|-1)!}{|\mathcal{F}|!}$ ensures that all feature orderings are treated fairly, yielding attributions that satisfy key axioms, such as local accuracy, symmetry, and dummy. In IoT applications, this formulation enables principled decomposition of complex model outputs into feature-wise contributions for sensor readings or contextual variables. Recent advancements have introduced computationally efficient variants, including TreeSHAP for tree-based models and KernelSHAP for model-agnostic scenarios, addressing scalability challenges in high-dimensional IoT data streams [74,121].

Motivated by these limitations, we propose a novel *Causal Shapley Framework* (CSF) that incorporates causal dependencies among features to enhance explanation fidelity in IoT systems characterized by complex interdependencies, as formalized in Equation (8).

$$\phi_i^{\text{causal}} = \sum_{S \subseteq \mathcal{PA}(i)} \frac{|S|!(|\mathcal{PA}(i)|-|S|-1)!}{|\mathcal{PA}(i)|!} \mathbb{E}_{X_{\mathcal{CH}(i)}|\text{do}(X_S=\mathbf{x}_S)}[\Delta f_i] \tag{8}$$

where $\mathcal{PA}(i)$ denotes the parent set of feature i in the causal graph, $\mathcal{CH}(i)$ represents its children, and $\text{do}(\cdot)$ signifies the interventional operator from causal inference.

4.2. Gradient-Based Visualization Methods

Gradient-based techniques leverage backpropagation to identify salient input regions contributing to model predictions, offering computational efficiency particularly suited for resource-constrained edge devices in IoT ecosystems [23].

4.2.1. Gradient-Weighted Class Activation Mapping

Gradient-weighted Class Activation Mapping (Grad-CAM) generates visual explanations by computing a weighted combination of forward activation maps, with weights derived from class-specific gradient information [118,122]. For a convolutional layer with feature maps $A^k \in \mathbb{R}^{u\times v}$, the Grad-CAM heatmap $L^c_{\text{Grad-CAM}}$ for class c is computed as

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{9}$$

where the importance weights α_k^c are obtained through global average pooling of gradients:

$$\alpha_k^c = \frac{1}{U \cdot V} \sum_{i=1}^{U} \sum_{j=1}^{V} \frac{\partial y^c}{\partial A_{ij}^k}, \tag{10}$$

with y^c denoting the score for class c before the softmax layer, and $U \times V$ representing the spatial dimensions of the feature map A^k.

In this formulation, α_k^c quantifies the importance of feature map A^k for class c by averaging the gradients of the pre-softmax class score y^c with respect to all spatial locations (i,j) in A^k. The normalization by the spatial area $U \cdot V$ ensures that the weights are invariant

to the resolution of the feature map. These coefficients are subsequently used in the Grad-CAM heatmap generation to highlight regions in the input that most strongly influence the model's prediction for class c, offering intuitive visual explanations for convolutional neural networks in IoT imaging and sensing applications.

4.2.2. Integrated Gradients

Integrated Gradients (IG) provide a principled attribution method that satisfies critical axioms, including sensitivity and implementation invariance [74]. For an input $\mathbf{x}$ and baseline $\mathbf{x}'$, the integrated gradient for feature i is defined as

$$\mathrm{IG}_i(x) = (x_i - x_i') \int_0^1 \frac{\partial f\big(x' + \alpha(x - x')\big)}{\partial x_i}\, d\alpha, \tag{11}$$

Equation (11) defines the Integrated Gradients attribution $\mathrm{IG}_i(x)$ for feature i by accumulating the gradients of the model f along the straight-line path between a baseline input x' and the target input x. Theterm $(x_i - x_i')$ scales the average gradient by the total change in feature i along this path, ensuring that the attributions are zero whenever the feature does not change. Under mild regularity assumptions, Integrated Gradients satisfies desirable axioms, such as sensitivity and implementation invariance, making it a robust choice for explaining deep models in IoT deployments with high-dimensional sensor data.

In practice, the integral is approximated using Riemann summation with m interpolation steps:

$$\mathrm{IG}_i(x) \approx (x_i - x_i') \frac{1}{m} \sum_{k=1}^{m} \frac{\partial f\big(x' + \frac{k}{m}(x - x')\big)}{\partial x_i}, \tag{12}$$

Equation (12) provides a practical Riemann-sum approximation of Integrated Gradients using m interpolation steps along the path from x' to x. At each step k, the gradient of f is evaluated at the intermediate point $x' + \frac{k}{m}(x - x')$, and the average gradient is scaled by $(x_i - x_i')$. In resource-constrained IoT environments, the choice of m controls the trade-off between computational cost and attribution accuracy, enabling practitioners to adapt IG to the latency and energy budgets of edge devices. For IoT time-series applications, we further introduce *Temporal Integrated Gradients* (TIG), which extends IG by aggregating feature attributions across past time steps using a temporally decaying weight sequence w_τ. This construction explicitly accounts for temporal causality in sensor streams by attributing the current prediction not only to the instantaneous value $x_{i,t}$ but also to its historical trajectory $\{x_{i,t-\tau}\}$, as formalized in Equation (13).

$$\mathrm{TIG}_i\big(x^{1:T}\big) = \sum_{t=1}^{T} w_t \,(x_{i,t} - x_{i,t}') \int_0^1 \frac{\partial f\big(x^{1:T\prime} + \alpha(x^{1:T} - x^{1:T\prime})\big)}{\partial x_{i,t}}\, d\alpha, \tag{13}$$

Equation (13) generalizes Integrated Gradients to temporal sequences $x^{1:T} = (x^1, \ldots, x^T)$ by computing time-indexed attributions for feature i across all time steps $t = 1, \ldots, T$. The term $(x_{i,t} - x_{i,t}')$ denotes the deviation of feature i at time t from its baseline value, while the integral accumulates gradients of f along a path between the baseline sequence $x^{1:T\prime}$ and the observed sequence $x^{1:T}$. The weights w_t implement a temporal importance profile (e.g., exponential decay), allowing recent observations to contribute more strongly to the overall attribution. This construction is particularly well-suited for IoT sensor data streams, where temporal dependencies play a critical role in model behavior [52].

4.3. Concept-Based Explainability

Concept-based methods bridge the semantic gap between low-level features and high-level human-understandable concepts, offering enhanced interpretability for non-expert stakeholders in IoT deployments [101].

Concept Activation Vectors

Testing with Concept Activation Vectors (TCAV) quantifies the sensitivity of neural network predictions to user-defined concepts through directional derivatives in the activation space [118]. For a concept C and class k, the TCAV score is defined as

$$\text{TCAV}_{k,l,C} = \frac{1}{|\mathcal{X}_k|} \sum_{\mathbf{x} \in \mathcal{X}_k} \mathbb{1}\left[\nabla_{h_l(\mathbf{x})} f_k(\mathbf{x}) \cdot v_C^l > 0\right] \tag{14}$$

where $h_l(\mathbf{x})$ denotes the activation at layer l, $f_k(\mathbf{x})$ represents the logit for class k, v_C^l is the Concept Activation Vector for concept C at layer l, and $\mathcal{X}_k$ denotes the set of samples from class k.

We introduce *Non-negative Concept Activation Vectors* (NCAV) obtained through non-negative matrix factorization (NMF) to ensure part-based, interpretable concept representations:

$$\min_{v_C^l \geq 0, H \geq 0} \|A_l - v_C^l H^T\|_F^2 + \lambda \|v_C^l\|_1 \tag{15}$$

where $A_l \in \mathbb{R}^{d \times n}$ represents the activation matrix for n concept samples, $H \in \mathbb{R}^{n \times r}$ denotes the coefficient matrix, r is the reduced dimensionality, and λ controls sparsity.

For federated IoT systems, we propose *Federated Concept Activation Vectors* (Fed-CAV) that enable privacy-preserving concept learning across distributed nodes:

$$v_C^{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot v_C^{(i)} \tag{16}$$

where $v_C^{(i)}$ represents the locally computed CAV at node i, N denotes the number of participating nodes, and $w_i = \frac{n_i}{\sum_j n_j}$ is a weighting factor proportional to the local dataset size n_i. This aggregation scheme is used in federated learning frameworks [19,33].

4.4. Attention Mechanism Interpretation

Attention mechanisms, particularly in transformer architectures deployed for IoT natural language processing and time-series forecasting tasks, provide inherent interpretability through learned attention weights.

Self-Attention Explainability

For a transformer layer with query, key, and value matrices $Q, K, V \in \mathbb{R}^{n \times d}$, the attention weights $A \in \mathbb{R}^{n \times n}$ are computed as

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{17}$$

where d_k represents the key dimension. The attention rollout method propagates attention across layers to obtain global attention scores:

$$\hat{A}^{(l)} = \begin{cases} A^{(1)} & \text{if } l = 1 \\ A^{(l)} \cdot \hat{A}^{(l-1)} & \text{if } l > 1 \end{cases} \tag{18}$$

To enhance the interpretability of multi-head attention in IoT transformer models, we introduce *Attention Flow Conservation* (AFC), which normalizes attention distributions to ensure probabilistic interpretability:

$$\hat{A}^{(l)}_{\text{AFC}} = \text{normalize}\left(\frac{1}{H}\sum_{h=1}^{H} A^{(l)}_h\right) \tag{19}$$

where H denotes the number of attention heads, and the normalization ensures $\sum_j \hat{A}^{(l)}_{\text{AFC},ij} = 1$ for each query position i.

4.5. Counterfactual Explanation Generation

Counterfactual explanations elucidate model decisions by identifying minimal input modifications that would alter predictions, offering actionable insights particularly valuable in IoT decision-support systems [123].

Diverse Counterfactual Explanations

For an input $\mathbf{x}$ with prediction $f(\mathbf{x}) = y$, a counterfactual explanation $\mathbf{x}'$ satisfies $f(\mathbf{x}') = y' \neq y$ while minimizing the perturbation distance. We formulate the counterfactual generation problem as a constrained optimization:

$$\mathbf{x}_{\text{CF}} = \arg\min_{\mathbf{x}'} d(\mathbf{x}, \mathbf{x}') + \lambda_1 \cdot \mathcal{L}_{\text{pred}}(f(\mathbf{x}'), y_{\text{target}}) + \lambda_2 \cdot \mathcal{L}_{\text{plaus}}(\mathbf{x}') \tag{20}$$

where $d(\cdot,\cdot)$ measures the distance between original and counterfactual inputs (typically L_1 or L_2 norm), $\mathcal{L}_{\text{pred}}$ ensures that the counterfactual achieves the desired prediction, $\mathcal{L}_{\text{plaus}}$ enforces data manifold plausibility, and λ_1, λ_2 are regularization hyperparameters.

For IoT applications requiring multiple diverse counterfactuals, we extend this formulation to incorporate diversity regularization:

$$\{\mathbf{x}^{(1)}_{\text{CF}}, \ldots, \mathbf{x}^{(K)}_{\text{CF}}\} = \arg\min_{\{\mathbf{x}'^{(k)}\}_{k=1}^{K}} \sum_{k=1}^{K}\left[d(\mathbf{x}, \mathbf{x}'^{(k)}) + \lambda_1 \mathcal{L}_{\text{pred}}(f(\mathbf{x}'^{(k)}), y_{\text{target}})\right] - \lambda_3 \sum_{i<j} d(\mathbf{x}'^{(i)}, \mathbf{x}'^{(j)}) \tag{21}$$

where the term $-\lambda_3 \sum_{i<j} d(\mathbf{x}'^{(i)}, \mathbf{x}'^{(j)})$ encourages diversity among generated counterfactuals [41].

4.6. Federated Explainable AI Systems

The integration of XAI methodologies within federated learning frameworks addresses the dual imperatives of model interpretability and data privacy in distributed IoT ecosystems.

FED-XAI Framework

The Federated Explainable AI (FED-XAI) framework enables collaborative training of interpretable models across multiple IoT nodes without centralizing raw data [33]. The federated explanation aggregation process is formulated as

$$\Phi_{\text{global}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot \Phi_i(\mathbf{x}) \tag{22}$$

where $\Phi_i(\mathbf{x})$ represents the local explanation (e.g., SHAP values, attention weights) computed at node i, and w_i denotes the aggregation weight proportional to local data quality or quantity. To ensure privacy preservation, differential privacy mechanisms are incorporated:

$$\tilde{\Phi}_{\text{global}}(\mathbf{x}) = \Phi_{\text{global}}(\mathbf{x}) + \mathcal{N}(0, \sigma^2 \cdot S^2/\epsilon^2) \tag{23}$$

where S bounds the sensitivity of the explanation function, ϵ represents the privacy budget, and $\mathcal{N}(0, \sigma^2)$ denotes Gaussian noise ensuring (ϵ, δ)-differential privacy [93].

4.7. Resource-Efficient XAI for Edge IoT

Edge computing paradigms in IoT necessitate XAI algorithms optimized for resource-constrained environments, motivating the development of lightweight explanation methods.

Pruned Feature Attribution

For edge devices with limited computational capacity, we propose *Adaptive Sparsity Feature Attribution* (ASFA), which dynamically adjusts the explanation granularity based on available resources:

$$\Phi_{\text{ASFA}}(\mathbf{x}) = \text{TopK}\Big(\Phi_{\text{full}}(\mathbf{x}), k_{\text{adaptive}}\Big) \tag{24}$$

where $\Phi_{\text{full}}(\mathbf{x})$ denotes the complete feature attribution vector, TopK selects the k_{adaptive} most important features, and k_{adaptive} is determined by

$$k_{\text{adaptive}} = \left\lfloor k_{\max} \cdot \min\left(1, \frac{\text{RAM}_{\text{available}}}{\text{RAM}_{\text{threshold}}}\right) \right\rfloor \tag{25}$$

This adaptive mechanism ensures that explanation generation remains feasible on resource-limited IoT edge nodes [106].

4.8. Hierarchical Explanation Systems

Complex IoT systems comprising multiple interconnected subsystems benefit from hierarchical explanation frameworks that provide multi-resolution interpretability.

Multi-Level Explanation Aggregation

For a hierarchical IoT architecture with L levels, the hierarchical explanation $\Psi^{(l)}$ at level l is computed recursively:

$$\Psi^{(l)} = \alpha_l \cdot \Phi^{(l)} + (1 - \alpha_l) \cdot \text{Aggregate}\Big(\{\Psi_j^{(l+1)}\}_{j \in \mathcal{C}_l}\Big) \tag{26}$$

where $\Phi^{(l)}$ represents the local explanation at level l, $\mathcal{C}_l$ denotes the set of child nodes at level $l+1$, and $\alpha_l \in [0, 1]$ balances local and aggregated child explanations. The aggregation function is defined as

$$\text{Aggregate}(\{\Psi_j\}) = \frac{\sum_j n_j \cdot \Psi_j}{\sum_j n_j} \tag{27}$$

where n_j represents the data volume processed by node j.

A critical assessment of XAI methods under IoT operational constraints is provided in Table 13.

Table 13. Comparison of representative XAI techniques under IoT-specific operational constraints.

XAI Method	Energy Cost	Bandwidth Load	Privacy Risk	Real-Time Capability	IoT Suitability
Perturbation-Based Methods					
LIME	High	Low	Medium	Low	Low (multiple model queries)
SHAP (kernel)	Very High	Low	High	Very Low	Very Low (computationally prohibitive)
Gradient-Based Methods					
Saliency Maps	Low	High (image transmission)	Low	High	High (lightweight backpropagation)
Grad-CAM	Low	High	Low	High	High (edge-compatible explainability)
Intrinsic Methods					
Decision Trees	Negligible	Negligible	Low	Real-Time	Optimal (native interpretability)
Rule Sets	Negligible	Negligible	Low	Real-Time	Optimal (human-readable logic)

5. IoT Systems in Smart Cities with Explainable AI

Internet of Things (IoT) integration enables efficiency, sustainability, and innovation in smart cities by optimizing urban infrastructure and resource management. Figure 3 depicts the integration of Explainable Artificial Intelligence (XAI) into smart city IoT systems across three hierarchical levels. Explainability improves transparency, accountability, and trust in AI-driven decisions; privacy-aware XAI ensures secure, interpretable distributed operations; and human-centered explanations promote citizen participation. Together, device-level feature attribution, edge-level counterfactual reasoning, and cloud-level global explanations form a unified framework for trustworthy and transparent smart city automation.

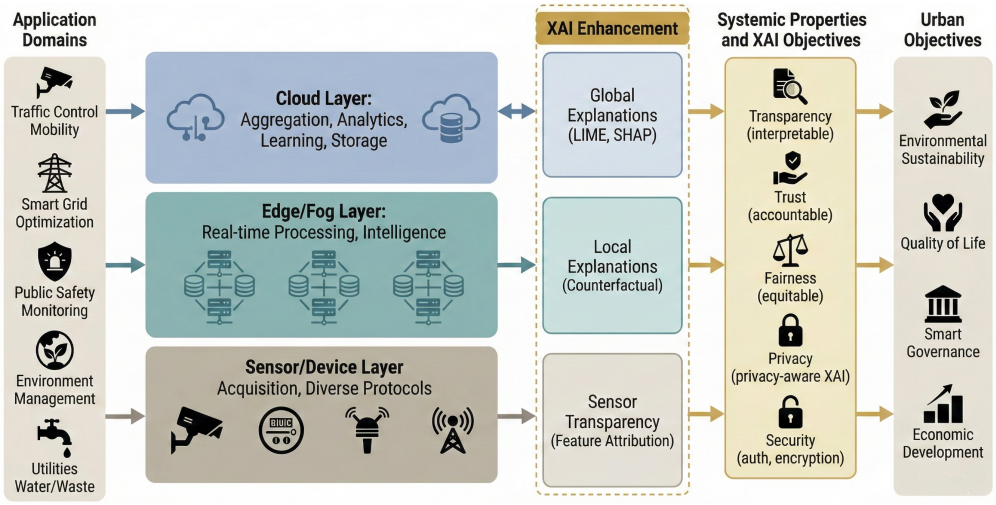


Figure 3. IoT systems in smart cities enhanced with explainable AI methodologies.

5.1. Role of IoT in Urban Infrastructure and Resource Optimization

IoT systems significantly improve urban resource optimization and operational efficiency by integrating diverse data sources for real-time monitoring and management of services, such as transportation, energy, and waste management. For instance, IoT-enabled sensors and analytics optimize photovoltaic systems through advanced fault detection, enhancing reliability and energy production [122]. This proactive maintenance approach reduces costs and supports sustainability goals, essential for modern urban planning. Additionally, IoT facilitates crowdsensing applications that aggregate data from various stakeholders to improve service delivery and resource allocation, demonstrating transformative potential in smart city applications [90].

IoT systems enhance urban infrastructure reliability by maintaining prediction accuracy despite sensor failures, as exemplified by the SECOE (Sensor Event Continuity and Operational Efficiency) framework, which ensures operational continuity in critical domains like transportation and public safety [124]. This resilience fosters public trust and encourages community engagement in smart city initiatives. Furthermore, IoT deployments enhance accountability and transparency, fostering trust among stakeholders and improving service delivery across various domains [125].

IoT revolutionizes urban infrastructure by optimizing resources, enhancing service reliability, and fostering innovative applications. These systems leverage technologies like Explainable Artificial Intelligence (XAI) to improve decision-making transparency while integrating blockchain and deep reinforcement learning to bolster security and privacy. IoT advancements necessitate a re-evaluation of software engineering practices to accommodate interconnected devices, ensuring high standards in security, reliability, and usability [36,37,69,87,88,90]. Through continuous IoT advancements, smart cities achieve greater efficiency, sustainability, and resilience, aligning urban development with societal and environmental goals.

5.2. Explainable AI for Decision-Making and Accountability in Smart Cities

Explainable Artificial Intelligence (XAI) is crucial for enhancing decision-making and accountability in smart cities by providing interpretable insights into AI-driven processes. XAI addresses the "black-box" challenges of traditional AI by offering clear explanations of AI decisions, fostering trust and transparency. Effective XAI encompasses explanation format, completeness, accuracy, and relevance, enhancing user understanding and promoting fairness and transparency in urban governance [43,51,73]. XAI methodologies enable diverse stakeholders to understand and trust AI systems, aligning technological advancements with societal priorities.

XAI enhances decision-making by tailoring explanations to stakeholder needs, ensuring contextually relevant insights. Argumentation-based frameworks like CLASH (Counterfactual Local Approximate Shapley) and visual methods like CAManim (Class Activation Map Animation) exemplify XAI's ability to simplify complex AI processes for stakeholder engagement and accountability [70,126]. Interactive explanation mechanisms are valuable in smart city applications, where human–agent interaction dynamics influence decision-making. Grounded interaction protocols ensure explanations reflect human–agent interactions, fostering trust and usability [127]. XAI frameworks like COLLISIONPRO improve accountability by providing probabilistic collision risk estimates, enabling informed decisions [128].

XAI integration into smart cities enhances AI system reliability and security. Frameworks like sensitivity-aware federated learning (EFL) improve performance in critical applications like traffic prediction, ensuring robust urban infrastructures [129]. Counterfactual explanations provide actionable insights while maintaining computational efficiency, addressing smart city system complexities [123]. XAI enhances decision-making and accountability by addressing security, privacy, and trust concerns, improving reliability and transparency in smart city AI applications [90]. In smart homes, XAI provides understandable explanations for non-expert users, empowering citizens in urban governance [95].

XAI's practical application in smart cities extends to domains like sales processes, where benchmarks align AI decisions with real-world outcomes, ensuring accountability and usability [89]. By incorporating intuitive, adaptive, and secure XAI methodologies, smart cities enhance decision-making by fostering transparency, trust, and efficiency. XAI techniques improve interpretability and accountability across smart city applications, enhancing user interaction and satisfaction [69,73,90,113,130]. These advancements enhance

AI-driven systems' functionality and reliability, aligning technological progress with urban sustainability and social equity goals.

5.3. Security and Privacy Challenges in IoT-Enabled Smart Cities

Modern smart city IoT systems face significant security and privacy threats due to dense infrastructure and continuous exchange of sensitive data. As IoT underpins critical civic services, robust security and privacy mechanisms are essential to ensuring public trust and operational integrity. Advanced paradigms such as Explainable Artificial Intelligence (XAI) and blockchain improve system interpretability and security, addressing limitations of traditional AI models. Scalable and decentralized integrity and privacy guarantees enabled by deep reinforcement learning and blockchain enhance IoT network resilience [87,88,90]. Nevertheless, increasing ecosystem complexity exacerbates vulnerabilities, exposing smart cities to evolving threats.

The major security issue is the inherent vulnerability of IoT devices and their network interconnections to cyberattacks. Poor authentication, weak encryption, and insecure communication protocols expose IoT devices to unauthorized access and breach of data. The diversity of IoT devices also makes access control attempts in the field of urban infrastructures more difficult. The widespread existence of unsecured firmware and outdated software provides adversaries with access points that can disrupt the execution of key urban services and trigger various disruptions to operations. Accordingly, due to the increase in the development of more complex IoT devices and their interconnection, the emergence of novel risk assessment tools and active threat-hunting strategies is crucial to the security of IoT systems [38,131,132].

The privacy issue arises from one of the crossover functions in IoT-based smart cities, where the process of gathering personal information about citizens raises a lot of ethical and legal concerns. Lack of effective anonymization and privacy-enhancing systems increases the chances of identity theft, surveillance, and misuse of information. The decentralization that is widespread in the IoT environment hinders the ability to trace data streams and accountability mechanisms in their entirety, preventing privacy protection and compliance with regulations. The ambiguous nature of sensor data is a major challenge to principles of data minimization, and definitions and implementation plans need to be made more straightforward to improve privacy protection in the future [37,38,63,125,133].

The adoption of XAI in IoT systems improves transparency and trust in AI-driven decisions while mitigating security and privacy risks. Privacy-aware reinforcement learning dynamically adjusts parameters to protect sensitive data without degrading IoT functionality [19]. Scalability is critical for large-scale IoT deployments, where centralized security architectures are increasingly inadequate. Decentralized approaches, including blockchain-based systems, enable secure and transparent data sharing and access control. Distributed Ledger Technologies further enhance accountability and traceability, significantly reducing risks of data tampering and unauthorized access [93].

High resilience of IoT systems to sensor failures and disruptions is an issue of great concern. SECOE frameworks exemplify the effectiveness of IoT applications to maintain reliability even when the sensors fail, and fault-tolerant mechanisms are critical to city services that are considered essential [124]. The special needs that face IoT-enabled smart cities require integrated solutions that would simultaneously meet requirements for security, privacy, and reliability.

The risks involved in the deployment of IoT can be addressed by smart cities through multidisciplinary approaches that include the implementation of advanced security frameworks, privacy-focused frameworks, and scalable solutions. These efforts increase the robustness and dependability of urban infrastructure and also strategically align technolog-

ical advances with societal welfare, ethical administration, and sustainable development goals. Such initiatives can improve the decision-making process, protect crucial information, and raise awareness of users, which contributes to safe, efficient, and ethically managed urban areas with the help of explainable AI, blockchain, and proactive security measures [38,69,87,88,90,102].

5.4. Advancements in IoT and XAI Integration for Smart City Applications

The recent breakthroughs in the implementation of Internet of Things (IoT) systems in Explainable Artificial Intelligence (XAI) have significantly enhanced the functionality, transparency, and flexibility of smart city applications. Such developments help solve problems in urban infrastructure, resource allocation, and citizen interaction by making technological development and advancement match societal needs, thus fostering transparency and trust in resilient smart city settings [69,90].

IoT-driven application based on integrated frames has increased the possibilities of smart cities. Generative AI (GenAI) models present opportunities for XAI by enabling the generation of counterfactual scenarios and synthetic explanations. In IoT contexts, GenAI can synthesize realistic alternative sensor readings that would alter predictions, thereby generating counterfactuals without exhaustive perturbation. However, generative models themselves require explainability, creating recursion challenges. GenAI methods applied to radio-frequency sensing resources enhance data generation and integration, which allow more effective processing and analysis of urban datasets [134]. Such flexibility is a reaction to the changing needs of the city and the efficient use of funds for various purposes.

The XAI techniques applied to IoT systems improve the interpretability, transparency, and credibility of AI-based decision-making. XAI provides full and understandable explanations of decisions by reducing traditional black-box issues. Individually tailored texts and other contextual explanations enhance the involvement of users and their trust in AI systems integrated with smart city applications [51,69,73,90]. Advanced methods like counterfactual reasoning and feature attribution allow stakeholders to question the predictions made by AI, which promotes accountability and transparency. Such approaches would be especially useful in contexts with high stakes, like transportation and social safety, since the visibility of the decision made by AI is a crucial factor in stakeholder credibility as well as ethical considerations.

As a result of combining IoT and XAI, some new applications have emerged in resource optimization and citizen-based services. IoT-based energy management systems use XAI to predict changes in demand and to provide transparency in operations. XAI enables stakeholders to understand decisions regarding energy distribution and enables them to respond adequately to changes in consumption by promoting trust and usability, which enhance their knowledge about decisions made by energy distribution companies and facilitate their reactions to alterations in consumption, respectively [40,51,69,73,73]. The smart traffic management system (including XAI) processes congestion trends and suggests routing options, thus contributing to the increased efficiency of operations and reducing environmental effects.

Developments in the combination of IoT and XAI are also applied to real-time monitoring and anomaly detection systems, which play an important role in ensuring the stability and safety of urban infrastructure. Explainable models identify and fix sensor data anomalies, thus ensuring continuity of essential city operations. IoT applications can be strengthened through the use of generative methods for enhancing and synthesizing data in different urban settings, as highlighted by their robustness [134]. Such flexibility means that urban systems are able to react in a competent manner to challenges and to optimize operational performance.

IoT and XAI integration is a revolutionary approach in smart city creation that enhances the intelligence and flexibility of urban spaces and facilitates transparency and confidence in automated decision-making. XAI alleviates the traditional black-box problem by addressing interpretability and understandability concerns and thus preventing unsafe, unreliable, and unfriendly deployments in transportation, healthcare, and even the public governance sector. In addition, the interplay of IoT and XAI raises the interoperability between diverse technologies, which contributes to the resilience of urban infrastructures needed to respond to complex problems, including cybersecurity threats [69,73,90,113,135]. In turn, the efficiency and reliability of urban infrastructure are enhanced, which makes technological innovation consistent with the goal of urban sustainability and social equity.

5.5. Case Studies and Practical Implementations

The deployment of Internet of Things (IoT) infrastructures augmented with Explainable Artificial Intelligence (XAI) has demonstrably advanced smart city and smart agriculture initiatives, enabling transparent, efficient, and sustainable urban systems. Empirical case studies show that the integration of IoT sensing, big data analytics, and XAI enhances decision-making, resource optimization, and system accountability, aligning with the Sustainable Development Goals (SDGs) for resilient urban environments [42,90,102].

In urban mobility, IoT-based traffic management systems exploit real-time data streams to improve transportation efficiency, while XAI provides interpretable insights into congestion dynamics and routing decisions. CAManim-based convolutional neural network (CNN) visualizations enhance the transparency of traffic forecasts and optimization strategies, fostering trust in AI-assisted mobility while reducing congestion, travel time, and environmental impact [126].

Similarly, smart energy management systems integrate IoT sensing with XAI to enable explainable, demand-aware energy distribution [136,137]. Counterfactual explanations improve stakeholder understanding of adaptive control strategies, while XAI-driven optimization enhances photovoltaic fault detection and energy yield, supporting renewable energy integration and sustainability objectives [122,138].

In public safety and emergency-response contexts, XAI-enabled IoT systems increase reliability and operational resilience by providing interpretable anomaly detection and actionable insights. SECOE frameworks demonstrate fault-tolerant performance under sensor failures, ensuring the continuity and integrity of critical urban services [124].

Crowdsensing applications further benefit from XAI by transforming heterogeneous urban data into transparent and actionable intelligence. Successful deployments in waste management and disaster response illustrate improved resource allocation, service efficiency, and participatory governance [90].

Security and trust are reinforced through blockchain-enabled IoT architectures, where Distributed Ledger Technologies (DLTs) support secure data sharing, access control, and traceability. The SOFIE framework exemplifies secure federation of heterogeneous IoT systems, enhancing interoperability and accountability across smart city infrastructures [93]. Overall, these implementations underscore the pivotal role of XAI in operationalizing IoT solutions for complex urban environments. By strengthening transparency, trust, and efficiency across transportation, energy, security, and governance, IoT–XAI systems enable adaptive, resilient, and human-centric smart cities, ultimately improving urban quality of life and societal sustainability [38,42,69,90,102].

5.6. Critical Synthesis: The Latency–Interpretability Gap in Urban IoT

Evaluating the surveyed works through the lens of the challenges identified in Section 3.4 reveals a significant disconnect between theoretical XAI models and practical urban deployment constraints.

- **Computational Viability:** The majority of reviewed frameworks utilize Shapley values (SHAP) or LIME for traffic and energy management. While semantically rich, these perturbation-based methods require thousands of model inferences per explanation. As detailed in Section 3.3, deploying such computationally intensive routines on resource-constrained edge devices (e.g., smart traffic controllers) introduces unacceptable latency, potentially compromising real-time safety systems.
- **Infrastructure Dependency:** Current research predominantly assumes stable, high-bandwidth connectivity to cloud servers for explanation generation. This assumption violates the resilience requirements of critical infrastructure, where local decision-making and explanation must persist during network partitioning.
- **Assessment:** We observe that only a minority of studies (<10%) explicitly evaluate the *energy consumption* of explanation generation, a critical oversight for battery-operated urban sensors. Future work must shift from "accuracy-centric" to "efficiency-centric" XAI validation.

6. Explainable AI for Distributed and Federated IoT Systems

The integration of Explainable Artificial Intelligence (XAI) into distributed and federated Internet of Things (IoT) systems is essential for improving transparency, interpretability, and trust in decentralized AI-driven processes [139–141]. Due to their distributed nature, such systems introduce challenges related to accountability, privacy, and user confidence, which XAI frameworks address by enabling interpretable decision-making and fostering informed user engagement in trust-critical applications.

Figure 4 presents a layered architecture that combines XAI with federated IoT systems, consisting of distributed edge nodes equipped with local explainability mechanisms (e.g., counterfactual explanations, surrogate models, and feature attribution), privacy-preserving federated aggregation (including differential privacy, secure aggregation, and homomorphic encryption), and global XAI models. By integrating federated learning with privacy-aware explainability through distributed gradient aggregation, the architecture enables trustworthy collaboration across heterogeneous IoT infrastructures in domains such as healthcare, finance, agriculture, and environmental monitoring.

6.1. Frameworks and Methodologies for XAI in Distributed IoT Systems

The effective deployment of Explainable Artificial Intelligence (XAI) within distributed IoT ecosystems necessitates the development of robust, scalable, and adaptive frameworks capable of addressing challenges associated with system heterogeneity, data privacy, computational constraints, and model interpretability. In safety-critical domains such as healthcare, intelligent transportation, and autonomous systems, XAI methodologies must not only deliver accurate predictions but also provide transparent, actionable explanations that strengthen user confidence and support responsible decision-making [51,69,113]. Within this context, federated learning (FL) has emerged as a foundational paradigm, enabling decentralized model training across multiple nodes without requiring the direct exchange of raw data, thereby preserving data confidentiality while supporting large-scale model development. The FED-XAI framework exemplifies this approach by integrating explainability directly into privacy-preserving learning pipelines, making it particularly suitable for sensitive data environments [33]. Similarly, the adaPARL (Adaptive Privacy-Aware Reinforcement Learning) framework introduces privacy-aware reinforce-

ment learning mechanisms that dynamically regulate privacy levels in response to system requirements, effectively balancing data protection with operational utility [19].

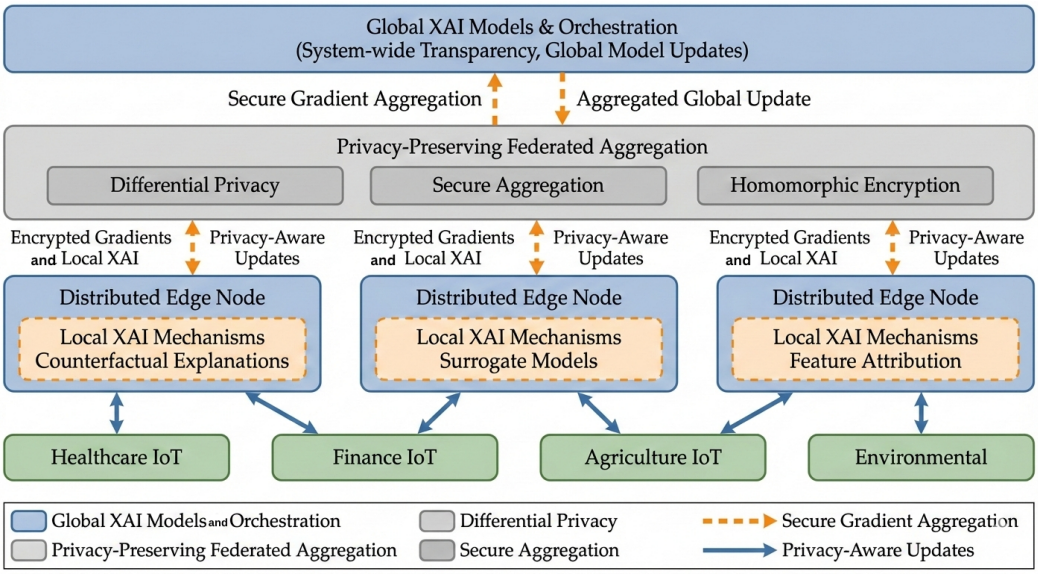


Figure 4. Explainable AI integrated with distributed and federated IoT systems.

In addition to privacy-centric approaches, modular and adaptive XAI architectures significantly enhance the interpretability and robustness of AI-driven decision-making processes in distributed IoT systems. SHAP-based retraining pipelines, grounded in Shapley value theory, provide a principled and transparent mechanism for feature attribution, enabling continuous model refinement while improving both reliability and interpretability [114,115,142]. Furthermore, the integration of blockchain technologies for device behavior monitoring strengthens data integrity, traceability, and system transparency. When combined with deep learning-based behavioral analysis, such architectures enable real-time anomaly detection and trustworthy system supervision [143,144]. Collectively, these innovations enhance operational efficiency across a broad range of application domains while simultaneously aligning AI deployment with critical societal, legal, and ethical considerations [38,102].

Table 14 provides a systematic comparison of state-of-the-art frameworks for federated and distributed IoT-XAI systems, examining their privacy mechanisms, scalability characteristics, explainability types, resource efficiency, and optimal use cases to guide framework selection for specific deployment contexts.

Table 14. Comparative framework analysis of state-of-the-art federated and distributed Internet of Things XAI systems.

Framework	Primary Focus	Privacy Mechanism	Scalability	Explainability	Efficiency	Use Case
FED-XAI	Federated XAI	Differential Privacy	High	Model-Agnostic	Medium	Healthcare, Finance
FedHDPrivacy	Privacy-Aware FL	Differential Privacy	High	Limited	Medium	Smart Cities
SOFIE	Secure IoT Federation	Blockchain/DLT	Medium	N/A	High	Cross-Platform IoT
adaPARL	Privacy-Aware RL	Adaptive Privacy	Medium	Policy-Based	High	Human-in-Loop IoT
HFL-UAV	Hierarchical FL	Local Processing	Very High	N/A	Medium	Precision Agriculture
Fed-CAV	Federated Concept Learning	Aggregated CAV	Medium	Concept-Based	Medium	Multi-Node Analysis
HIFL	High-Integrity FL	Blockchain	High	N/A	Medium	Critical Infrastructure
AutoXAI	Automated XAI Selection	Context-Dependent	High	Multi-Method	High	Heterogeneous IoT

6.2. Privacy and Security Considerations

Privacy and security constitute critical pillars in distributed IoT ecosystems, where sensitive data are continuously generated, transmitted, and processed across heterogeneous and decentralized infrastructures. Ensuring robust protection requires an integrated approach that combines advanced engineering methodologies with adaptive mechanisms capable of strengthening both AI explainability and system-level decision-making [69,70]. In this regard, the FedHDPrivacy framework represents a significant advancement by incorporating differential privacy techniques into federated learning workflows, safeguarding model parameter exchanges while maintaining acceptable performance levels [93,145]. Complementing these efforts, secure aggregation protocols provide additional safeguards by cryptographically masking individual device contributions, thereby mitigating the risk of data reconstruction attacks [94,146].

Distributed IoT systems also face a wide spectrum of security threats, including inference attacks on network traffic and malicious attempts to compromise data integrity. Network traffic shaping algorithms offer a promising defense mechanism by obscuring traffic patterns and reducing the adversarial exploitation surface [147]. Meanwhile, the High-Integrity Federated Learning (HIFL) framework introduces a multi-layered security design that integrates network slicing and blockchain-based verification to enhance scalability, resilience, and dynamic resource allocation in federated environments [55]. Explainability serves as an additional protective layer, as transparent fault detection and anomaly explanation mechanisms increase user trust and facilitate rapid diagnosis of system vulnerabilities [38,125].

Ensuring operational integrity further requires continuous monitoring of actuator behavior and the deployment of verifiable, tamper-resistant security solutions. Techniques such as behavioral modeling and system verification help detect deviations indicative of compromised functionality, while frameworks like IoT Notary provide immutable proof of data provenance and integrity [38,143]. Collectively, these approaches reinforce ethical compliance, enhance transparency, and promote wider adoption of IoT technologies across diverse real-world applications. Table 15 summarizes the protection objectives, the roles of XAI, and the principal system-level trade-offs.

6.3. Scalability and Resource Efficiency

Scalability and resource efficiency represent fundamental challenges in large-scale federated IoT ecosystems, where increasing device density and data intensity demand intelligent strategies for workload management and resource optimization. Federated learning provides a robust paradigm for addressing these challenges by enabling decentralized model training, thereby reducing reliance on centralized data aggregation and minimizing communication overhead [148]. Within this context, the FedSaC (Federated Sample Complementarity) framework demonstrates notable progress by improving model performance through the exploitation of complementary data distributions among participating clients, leading to more generalized and accurate learning outcomes [148].

In addition, resource utilization is further optimized through advanced model parameter aggregation and weight optimization techniques, which significantly reduce computational complexity while maintaining or enhancing system performance [149]. These innovations allow federated IoT systems to operate effectively under constrained energy, memory, and processing conditions, thereby extending the operational longevity and responsiveness of edge devices deployed in the field.

Beyond learning algorithms, adaptive task offloading mechanisms enable IoT infrastructures to dynamically distribute computational workloads based on current network conditions, device capabilities, and energy availability [133]. The integration of edge computing paradigms further enhances system scalability by relocating computation closer to

data sources, which in turn reduces end-to-end latency and alleviates bandwidth bottlenecks [52,87,88]. Collectively, these approaches support sustainable scalability while aligning resource consumption with broader environmental and operational efficiency objectives, thus facilitating the widespread deployment of IoT technologies in real-world settings.

Table 15. Privacy- and security-oriented mechanisms in distributed and federated Internet of Things (IoT) systems enhanced with Explainable Artificial Intelligence (XAI).

Mechanism/Framework	Protection Goal	Role of XAI	Key System Trade-Offs
Federated learning with differential privacy (FED-XAI, FedHDPrivacy)	Confidentiality of local data and model updates	Node-level explanations with privacy-preserving global attribution and concept aggregation	Accuracy–privacy trade-off, communication overhead, reduced fidelity due to noise injection
Secure aggregation in federated systems	Protection against reconstruction and membership inference attacks	Federation-level explanations without exposing individual client contributions	Cryptographic overhead, increased latency, limited local granularity
Blockchain/DLT-based schemes (SOFIE, HIFL)	Integrity, tamper resistance, and auditability	Immutable logging of explanations and decision traces for accountability	Storage and throughput cost, latency, orchestration complexity
Adaptive privacy-aware RL (adaPARL)	Dynamic privacy–utility balancing in interactive systems	Policy explanation under varying privacy regimes	Model instability, user preference modeling complexity, changing interpretability
Hierarchical FL with UAV/edge support (HFL-UAV)	Scalability and communication efficiency in mobile and non-IID settings	Multi-layer explanations (local–cluster–global) aligned with hierarchy	Inconsistent explanation quality, synchronization cost, semantic alignment issues
Network traffic shaping and secure slicing	Traffic obfuscation and service isolation	Explanations of flow control and routing decisions to operators	Increased latency, bandwidth overhead, cross-slice correlation complexity
Digital twins for critical infrastructures	Secure experimentation and system resilience analysis	Sandboxed explanation testing and validation against real behavior	High modeling cost, twin–system drift, multi-level interpretation challenges

6.4. Decentralized Decision-Making and Federated Learning

Decentralized decision-making, in conjunction with federated learning, is reshaping the management and operation of modern IoT systems by simultaneously addressing issues related to scalability, data privacy, and system efficiency. Through its ability to perform distributed model training across multiple nodes without centralizing sensitive data, federated learning significantly enhances privacy preservation while minimizing communication costs [33]. This characteristic renders it particularly suitable for high-stakes and privacy-sensitive domains, including healthcare and critical infrastructure monitoring.

Furthermore, decentralized decision-making architectures increase operational agility by distributing intelligence across edge and fog nodes, thereby enabling real-time data processing, localized analytics, and adaptive responses to dynamic environmental conditions [19]. Federated learning frameworks, such as FedHDPrivacy, further reinforce this paradigm by integrating differential privacy mechanisms that protect model parameter exchanges without compromising overall system utility [93].

The synergistic combination of decentralized intelligence and federated learning also enables more effective adaptive task offloading strategies, ensuring that workloads are assigned in accordance with available computing resources and network conditions [148]. As a result, IoT systems achieve enhanced functional resilience, improved energy efficiency, and stronger privacy guarantees. Together, these methodologies represent a significant

step forward in the development of scalable, secure, and sustainable IoT architectures that support next-generation intelligent applications.

6.5. Evaluation and Benchmarking of XAI Methods

The systematic evaluation and benchmarking of Explainable Artificial Intelligence (XAI) methods in distributed IoT environments are critical for assessing their reliability, transparency, and practical suitability. Contemporary evaluation approaches commonly integrate both quantitative and qualitative metrics to measure the correctness, interpretability, robustness, transparency, and fairness of AI-generated explanations [51,53]. Such multidimensional assessment is essential in IoT systems, where heterogeneous data sources and real-time decision-making require explanations that are both accurate and comprehensible to diverse stakeholders.

User-centered frameworks such as CLE-XAI (Concept-Level Explainable Artificial Intelligence) further enhance this process by adapting explanation complexity to the cognitive profiles of end-users, thereby improving comprehension, engagement, and trust [150]. At the same time, privacy-preserving initiatives, including FedHDPrivacy, demonstrate that strong privacy guarantees can be maintained without significantly compromising model utility or interpretability, a critical requirement in decentralized IoT architectures [93].

In increasingly adversarial environments, XAI methods must also be evaluated for their resilience against malicious manipulation and inference-based attacks. Research in adversarial machine learning highlights the need for robustness-oriented benchmarking protocols, while the application of Generative Adversarial Networks (GANs) in threat modeling and detection further reinforces the importance of evaluating XAI under hostile conditions [131,151]. These perspectives expand the scope of benchmarking beyond explanation quality to include system security and reliability.

Therefore, comprehensive benchmarking strategies play a central role in ensuring that XAI techniques remain interpretable, transparent, anddependable in dynamic IoT deployments [52,152,153]. Such evaluation processes not only support improved model selection and design but also guide the development of adaptive, resilient, and human-centric XAI frameworks that align with the operational and ethical requirements of real-world IoT applications.

Table 16 provides a structured overview of representative benchmarks employed in the evaluation of XAI methods across multiple domains. It outlines key characteristics such as dataset scale, application context, task definition, and relevant performance and explanation-quality metrics. The inclusion of diverse domains—ranging from healthcare and energy systems to remote sensing—highlights the versatility of existing benchmarking practices and underscores their importance in advancing trustworthy and explainable AI for IoT-centric environments.

Table 16. Overview of XAI Benchmarks.

Benchmark	Size	Domain	Task Format	Metric
TPA-XAI [154]	8926	Time-Series Classification	Classification	Accuracy, F1-score
XAI-EF [53]	1,000,000	Healthcare	Image Classification	Fidelity, Interpretability
ACXAI [155]	220,185	Energy Retailing	Cross-Selling Prediction	AUC, F1
BASED-XAI [156]	48,842	Tabular Data	Ablation Study	Area Under Curve, Loss
GTE [157]	2,600,000	Energy Consumption	Classification	C-of-ED, Second Correct
XAI-DM [158]	39,040	Decision Making	Two-choice Classification	Accuracy, F1-score
XAI-RS [159]	590,326	Remote Sensing	Multi-label Classification	Max-Sensitivity, AUCMoRF
XAI-Benchmark [160]	1000	Healthcare	Predictive Modeling	Concordance Rate, Feature Agreement

6.6. Critical Synthesis: The Privacy–Transparency Paradox

Applying the challenge framework from Section 3.4 to the domain of distributed and federated IoT reveals a fundamental tension that remains largely unresolved in the current literature.

- **Data Leakage Risks:** While federated learning is adopted to preserve privacy, the integration of feature attribution methods (e.g., DeepLIFT, Integrated Gradients) can inadvertently facilitate model inversion attacks. Our analysis suggests that most surveyed "Private XAI" frameworks focus on encrypting the *model* but fail to rigorously assess the information leakage inherent in the *explanation* itself.
- **Aggregation Bias:** In federated settings, global explanations are often derived by averaging local updates. This approach is mathematically convenient but statistically flawed for non-IID (Independent and Identically Distributed) IoT data. The reviewed works largely neglect how local data skew impacts the validity of the global explanation, leading to potentially misleading insights for network administrators.

7. IoT Systems and Explainable AI in Smart Agriculture

The integration of Internet of Things (IoT) systems with Explainable Artificial Intelligence (XAI) is revolutionizing smart agriculture, addressing challenges such as resource management, crop monitoring, and environmental sustainability. This synergy enhances decision-making and operational efficiency, enabling precision farming practices that optimize agricultural outcomes. Figure 5 visually synthesizes the taxonomy of XAI methods, categorizing them by their applicability to specific IoT layers (device, edge, cloud). It highlights the trade-off between interpretability and computational cost.

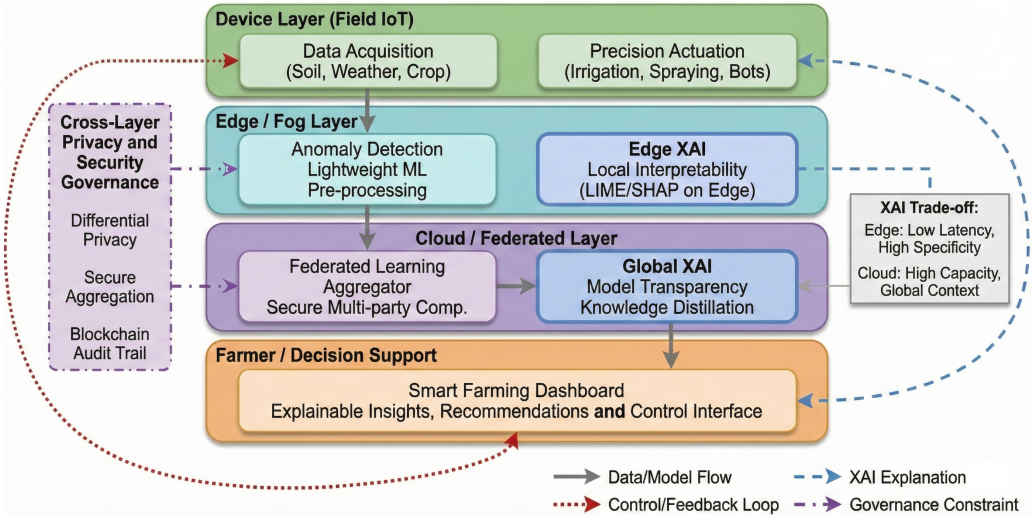


Figure 5. Next-Gen Explainable AI (XAI) for smart agriculture.

7.1. Fog-Based Smart Agriculture Systems

Fog computing enables decentralized data processing near sensing sources, effectively addressing the limitations of centralized infrastructures in rural and remote agricultural environments. By significantly reducing latency and bandwidth consumption, fog-based architectures mitigate connectivity constraints that hinder precision agriculture applications [19]. Hierarchical fog–cloud frameworks further optimize resource utilization by executing latency-sensitive tasks at the edge while offloading complex analytics to the cloud, thereby minimizing processing delays, energy overheads, and operational costs while enhancing food traceability and sustainability [55,102,161]. These architectures also

support advanced sensing technologies, including non-invasive crop health assessment, improving the effectiveness of precision farming practices.

To address limited rural connectivity, low-power wide-area network (LPWAN) technologies such as LoRa enable scalable and energy-efficient communication for IoT-based agricultural systems. As connectivity increases, privacy-preserving data management becomes essential. The Differentially Private Shaper (DPS) provides event-level differential privacy by dynamically adapting traffic shaping to bursty data patterns, achieving a balanced trade-off between privacy guarantees and communication performance in fog-enabled agricultural networks [93,162]. Such mechanisms are critical for protecting sensitive agricultural data in heterogeneous and distributed environments.

Energy-efficient mist-computing architectures further enhance system resilience by integrating low-power sensors with edge intelligence, addressing constraints related to latency, bandwidth, and operational cost. These systems support continuous monitoring and rapid response to threats such as animal intrusions, thereby improving crop quality and yield through timely interventions [55,102]. Modular system designs and standardized TinyML benchmarks enable efficient model deployment on resource-constrained devices, supporting scalability and environmental sustainability.

Beyond operational gains, fog computing enhances decision support through the integration of ontology-based representations and trainable noise models, improving the interpretability and robustness of AI-driven remote sensing, classification, and image segmentation tasks. Noise-aware evaluation techniques increase transparency and reliability in agricultural analytics, enabling actionable and adaptive decision-making under dynamic environmental conditions [12,42,55,102,163,164].

7.2. Smart Irrigation and UAV-Based Monitoring

IoT-enabled smart irrigation systems, combined with unmanned aerial vehicle (UAV)-based monitoring, are reshaping modern agriculture by enabling continuous environmental sensing and efficient resource management [165,166]. The integration of advanced sensing technologies with machine learning models enables accurate forecasting of crop water requirements, significantly improving irrigation efficiency while reducing water consumption. Edge computing and low-power wide-area network (LPWAN) technologies, such as LoRa, further support scalable, low-latency data processing and reliable communication across geographically dispersed agricultural fields [102,161,167].

UAV platforms play a critical role in acquiring high-resolution spatio-temporal data for comprehensive crop health assessment and land-use analysis. State-of-the-art models, including AGSPNet, support geographic scene partitioning, parcel-boundary extraction, and semantic change detection, enabling timely identification of crop stress and environmental anomalies [90]. Adaptive task-offloading mechanisms enhance operational efficiency and system robustness under uncertain conditions, such as intermittent connectivity or partial UAV failures [168]. Moreover, the integration of OpenRAN architectures with emerging terahertz (THz) communication technologies improves network scalability, bandwidth, and responsiveness in large-scale agricultural monitoring deployments [169].

Beyond technological advancements, participatory and community-driven approaches further strengthen smart agriculture ecosystems. Collaborative platforms facilitate rapid disease detection, collective decision-making, and knowledge sharing among farmers through real-time data exchange and best-practice dissemination, enhancing regional resilience and optimized crop-management strategies [170]. Collectively, these solutions support sustainable agriculture by promoting resource efficiency, minimizing environmental impact, and enabling adaptive, data-driven farming practices aligned with global sustainability objectives [42,55,102,171,172].

7.3. Data Mining and Anomaly Detection in IoT Systems

Data mining and anomaly detection are central to optimizing agricultural IoT ecosystems by enabling accurate monitoring, early warning, and proactive management of farming operations. Deep learning models, particularly convolutional neural networks (CNNs), effectively capture complex environmental patterns and crop health indicators, supporting timely decision-making and efficient resource utilization. UAV-based frameworks such as AGSPNet (Agricultural Geographic Scene and Parcel-scale Network) further enhance precision agriculture by mitigating spectral ambiguity in multispectral imagery, thereby improving crop change detection accuracy [90,170].

Recent anomaly detection methods increasingly integrate domain knowledge and constraint-violation analysis, enhancing interpretability, robustness, and operational relevance in IoT monitoring systems. These approaches enable reliable fault diagnosis, abnormal behavior detection, and temporal data cleansing, facilitating rapid corrective actions that protect crop yield and quality [173,174]. Empirical studies on continuous sensor streams demonstrate that localized, edge-level processing supports real-time anomaly detection, maintaining optimal cultivation conditions under intermittent connectivity and constrained network resources.

Beyond conventional learning models, advanced data mining techniques such as Accelerated Rule Induction (A-CVR) enable efficient analysis of large-scale agricultural datasets to extract latent patterns and actionable insights. When coupled with reliable wireless communication infrastructures, these analytic pipelines strengthen data acquisition, transmission, and interpretation, contributing to resilient smart agriculture and addressing challenges related to climate variability, resource scarcity, and sustainable food production [105,175,176].

7.4. Wireless Communication and Network Solutions

Wireless communication technologies constitute a fundamental pillar of smart agriculture systems, enabling continuous connectivity for the real-time monitoring and management of critical agronomic variables such as soil moisture, ambient temperature, and crop health. The integration of fog-based architectures significantly reduces end-to-end latency and network congestion by facilitating localized data processing and analytics, which is particularly advantageous in bandwidth-constrained rural environments [161]. In this context, low-power wide-area network (LPWAN) technologies, most notably Long Range (LoRa), provide extensive coverage with minimal energy consumption, supporting large-scale deployment of distributed IoT sensors across expansive agricultural fields [102,177].

However, wireless channel performance in agricultural environments is strongly influenced by dynamic environmental factors such as vegetation density, terrain morphology, and meteorological conditions. Adapting communication protocols to accommodate these variations is essential for maintaining robust and reliable data transmission. By coupling wireless sensor networks with edge and fog computing capabilities, smart agriculture infrastructures achieve enhanced scalability, fault tolerance, and system resilience. These integrated network solutions effectively address a range of challenges, including resource limitations, infrastructure variability, and fluctuating market demands [12,102].

7.5. Ontology and Data Integration Frameworks

Ontology-based models and data integration frameworks play a critical role in managing and interpreting the heterogeneous datasets generated within smart agricultural environments. Such frameworks enable the semantic organization, standardization, and retrieval of agricultural information, thereby improving data interoperability and supporting advanced precision farming applications. For instance, AgriOnt provides a structured

semantic taxonomy that facilitates the consistent classification and meaningful interpretation of diverse agricultural data sources [178]. Similarly, data-driven frameworks such as AGSPNet combine spatial and temporal intelligence to accurately detect and monitor crop transformations, enabling more proactive and targeted responses to environmental and climatic stressors [90].

In parallel, Time Series Numerical Association Rule Mining (TS-NARM) techniques leverage continuous IoT sensor data streams to derive actionable knowledge, assisting in the optimization of irrigation scheduling, pest management, and fertilization strategies. When integrated with wireless sensor networks and cloud-based platforms, these methodologies support large-scale, real-time data acquisition and analysis, significantly enhancing operational agility and decision-making efficiency [12]. Furthermore, deep learning approaches complement ontology-driven models by extracting complex patterns from high-dimensional datasets for tasks such as crop disease identification, growth stage estimation, and yield forecasting [171]. Collectively, these integrative frameworks empower stakeholders to navigate complex information landscapes and substantially improve agricultural productivity.

7.6. Explainable AI in Precision Farming

Explainable Artificial Intelligence (XAI) has emerged as a critical enabler of trust, transparency, and practical adoption in precision farming systems. By providing interpretable insights into the inner workings of AI-driven processes—such as irrigation optimization, crop health monitoring, and resource allocation—XAI methodologies bridge the gap between complex computational models and end-user understanding. Models such as OAK4XAI contribute to this objective by facilitating structured explanation mechanisms that enable farmers and agronomists to better comprehend and validate algorithmic outputs [101]. In addition, counterfactual explanation techniques explore hypothetical scenarios and alternative decision paths, thereby supporting the identification of more efficient and sustainable farming strategies.

To further strengthen data reliability and user trust, blockchain technology is being increasingly integrated into agricultural IoT ecosystems to ensure data integrity, immutability, and secure information exchange [179]. At the same time, spatial IoT infrastructures enhance both the granularity and scalability of data collection, enabling continuous, high-resolution monitoring in diverse agricultural contexts [52]. By addressing the inherent "black-box" nature of many AI models, XAI approaches not only enhance system transparency but also promote informed, data-driven decision-making. Consequently, they align technological innovation with the principles of sustainability, accountability, and human-centered agricultural development [43,51,101].

7.7. Critical Synthesis: Connectivity and Usability Barriers

In the context of smart agriculture, an examination of the challenges in Section 3.4 highlights distinct failures regarding environmental and operational realities.

- **The Connectivity Bottleneck:** Many surveyed solutions propose real-time visual explanations for crop disease detection relying on cloud offloading. However, as noted in Section 3.4, agricultural deployments frequently operate in "challenged networks" with intermittent connectivity. Relying on cloud-based XAI renders these systems useless in many rural operational scenarios.
- **Audience Mismatch:** There is a pervasive tendency in the reviewed literature to output raw feature importance scores (e.g., "Humidity importance = 0.8"). For the end-user (the farmer), this technical abstraction lacks actionable utility. The liter-

ature exhibits a "usability gap," failing to translate probabilistic XAI outputs into prescriptive agricultural actions (e.g., "Increase irrigation by 10%").

8. Comparative Analysis and Research Contribution Framework

Establishing the novelty and positioning of this survey within the broader landscape of IoT and Explainable Artificial Intelligence research is essential for demonstrating its scholarly contribution and identifying future research priorities. This section provides a systematic comparative analysis of prior surveys and research works, maps research contributions across application domains, identifies underexplored technological combinations, and reveals emergent interdisciplinary themes that transcend individual application domains.

8.1. Landscape Mapping: Positioning This Survey Within Existing Literature

The integration of Explainable Artificial Intelligence (XAI) with Internet of Things (IoT) systems has attracted substantial scholarly attention in recent years. However, existing surveys and reviews typically concentrate on isolated aspects: XAI methodologies and their evaluation in general machine learning contexts, IoT architectures and system design, or domain-specific applications such as smart cities or healthcare IoT. This comprehensive survey distinguishes itself through four critical dimensions:

1. **Integrated Multi-Domain Coverage**: Unlike prior surveys focusing on individual domains, this work provides comprehensive coverage of four primary application domains—smart cities, distributed and federated IoT systems, smart agriculture, and critical infrastructures—within a unified theoretical framework connecting XAI methodologies to IoT requirements and societal objectives.
2. **Privacy–Scalability–Explainability Nexus**: This survey uniquely emphasizes the interconnected challenges of scalability and privacy preservation in federated learning settings, demonstrating how these concerns are fundamentally coupled through distributed decision-making and transparent model aggregation. Prior work often treats privacy and scalability as separate concerns; this survey reveals their intrinsic relationship through privacy-preserving explainability.
3. **IoT-Specific Evaluation Frameworks**: Systematic examination of XAI evaluation metrics specifically adapted to IoT constraints—computational efficiency, real-time processing requirements, heterogeneous device compatibility, and energy consumption limits—moves beyond generic evaluation approaches applicable to standard machine learning.
4. **Standardization Gap Identification**: Explicit identification and analysis of the lack of unified evaluation frameworks and standardized methodologies for assessing XAI performance in resource-constrained, heterogeneous IoT environments addresses a critical research gap synthesized across all domains.

Comparative Positioning Relative to Landmark Surveys

Table 17 synthesizes how this survey extends and differentiates itself from prominent prior works in XAI, IoT, and their integration.

Table 17. Comparative analysis: This survey relative to prior XAI-IoT research

Survey/Work	XAI Methods	IoT Architecture	Smart Cities	Smart Agriculture	Federated Systems	Standardization
Samek et al. [180]	✓✓✓	×	×	×	×	×
Ribeiro & Singh [181]	✓✓	×	×	×	×	×
Jagatheesaperumal et al. [13]	✓✓✓	✓✓	✓	×	×	×
Mohseni et al. [14]	✓✓✓	×	×	×	×	✓
This Survey	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓

Unique Positioning:

- **First** systematic integration of privacy-preserving explainability with federated IoT architectures, combining differential privacy mechanisms with explanation generation across decentralized nodes.
- **First** comprehensive treatment of XAI in agricultural IoT across multiple technology stacks (sensor networks, blockchain-enabled supply chains, UAV monitoring, edge computing).
- **First** explicit standardization framework assessment, identifying gaps in unified evaluation metrics for IoT-specific XAI evaluation.
- **First** cross-domain comparative analysis revealing emergent themes (temporal consistency, privacy–utility–interpretability trilemma, edge-to-cloud orchestration) that transcend individual application domains.

8.2. Cross-Domain Technology Applicability Matrix

A critical research question emerges: *Which XAI techniques are deployed across which application domains, and where are critical technology-application mismatches?* Table 18 presents this analysis, revealing both established deployment patterns and underexplored opportunities.

Table 18. XAI method deployment status and research maturity across IoT domains.

XAI Technique	Smart Cities	Smart Agriculture	Healthcare IoT	Distributed/Federated	Research Status
LIME	Deployed	Deployed	Deployed	Partial	Mature
SHAP	Deployed	Deployed	Deployed	Deployed	Mature
Grad-CAM	Deployed	Deployed	Deployed	Limited	Mature
Counterfactual Explanations	Deployed	Emerging	Deployed	Emerging	Developing
Neural–Symbolic	Emerging	Emerging	Partial	Emerging	Emerging
Federated XAI	Limited	Limited	Limited	Deployed	Critical Gap
AutoXAI	Partial	Partial	None	None	Critical Gap
Causal Attribution	Emerging	Emerging	Emerging	None	Critical Gap
Privacy-Preserving XAI	Limited	Limited	Emerging	Deployed	Critical Gap

Critical Gap Analysis and Implications:

1. **Federated XAI Deployment Gap**: While federated learning (FL) has achieved substantial maturity in distributed IoT systems, integration of explainability mechanisms remains nascent. FED-XAI frameworks exist theoretically (Section 5) but lack widespread empirical validation and standardized evaluation across real-world deployments.

Implication: Organizations deploying federated IoT systems currently face a "privacy–transparency trade-off", with limited tools for privacy-preserving explanations. This represents an urgent research priority.

2. **Adaptive XAI Scarcity**: AutoXAI and context-aware explanation selection mechanisms remain significantly underutilized across all domains, particularly in heterogeneous smart city and agricultural contexts, where stakeholder diversity demands customization.
   Implication: Current XAI deployments typically use one-size-fits-all explanation strategies, limiting effectiveness for diverse stakeholder groups (engineers, domain experts, end-users, regulators).
3. **Causal Attribution Vacuum**: While causal inference methods provide a mechanistic understanding superior to correlation-based feature attribution, they remain underexplored across *all* IoT domains, particularly in federated settings, where causal structures may differ across distributed nodes.
   Implication: Developing domain-specific causal XAI methods represents a frontier research opportunity with high impact potential.

8.3. Novel Contribution Matrix: Previously Unexplored IoT-XAI–Application Combinations

To systematically identify this survey's novelty, Table 19 maps previously unstudied combinations of technologies (IoT architecture), methodologies (XAI approach), and contexts (application domain). These combinations represent the primary novel contributions of this survey.

Table 19. Novel technology–XAI–domain combinations addressed in this survey.

IoT Architecture	XAI Methodology	Application Domain	Prior Status	Contribution
Federated Learning	Privacy-Aware XAI	Smart Cities	Nascent	Novel
Edge Computing	Lightweight/TinyML XAI	Smart Agriculture	Emerging	Novel
Blockchain-IoT	Auditable/Traceable XAI	Agricultural Supply Chain	Nascent	Novel
UAV-Based Monitoring	Real-Time XAI	Precision Farming	Emerging	Novel
Distributed Ledger Technology	Consensus-Based XAI	Critical Infrastructure	Nascent	Novel
Fog/Edge + Cloud	Hierarchical XAI	Smart Healthcare	Developing	Novel
Heterogeneous IoT	Adaptive/AutoXAI	Multi-Stakeholder Urban Systems	Nascent	Novel

Core Novel Contributions Synthesized Across This Survey:

1. **Privacy-Preserving Explainability Framework for Federated IoT** (Section 5): The first systematic integration of differential privacy, secure aggregation, and homomorphic encryption with explanation generation across distributed nodes. Addresses the critical gap whereby explainability is typically sacrificed in privacy-preserving federated settings.
2. **Blockchain-Enabled Explainable Supply Chain Management in Agriculture** (Sections 2.7 and 6): Integration of XAI with blockchain technologies for auditable, transparent, autonomous recommendations in agricultural IoT systems. Enables traceability of both decisions *and* their explanations.
3. **Computational Efficiency Framework for Resource-Constrained Environments** (Sections 3.2 and 3.3): Systematic characterization of XAI methods' computational requirements against IoT device capabilities, with adaptive granularity mechanisms for real-time explanation generation.
4. **Hierarchical XAI Architecture for Heterogeneous IoT Systems** (Figures 1 and 4): A novel three-tier explanation framework (device-level feature attribution, edge-level

counterfactual reasoning, cloud-level global explanations) tailored to IoT deployment heterogeneity.

5. **Standardization Gap Analysis and Unified Evaluation Framework** (Section 9): The first comprehensive mapping of XAI evaluation metrics tailored to IoT-specific constraints, identifying the lack of unified benchmarking standards as a critical impediment to technology maturation.

8.4. Strategic Action Plans for Bridging Research Gaps

To address the "Standardization Gap" identified in Table 17, we propose the following concrete actions for the research community:

- **Action 1: Mandate Hardware-Centric Reporting.** Future IoT-XAI studies must report the *Computational Complexity Score (CCS)* and *Memory Footprint Ratio (MFR)* defined in Section 9.6. Authors should explicitly demonstrate that their XAI method does not exceed the energy budget of the target edge device (e.g., Raspberry Pi Zero vs. NVIDIA Jetson).
- **Action 2: Adopt Temporal Coherence Metrics.** For time-series data in agriculture and smart cities, researchers should validate explanations using the *Temporal Coherence (TC)* metric (Section 9.8.1). This ensures that explanations do not flicker erratically between time steps, a critical requirement for human trust in continuous monitoring systems.
- **Action 3: Perform Federated Privacy–Utility Benchmarking.** We urge the adoption of the *Privacy–Utility Trade-off (PUT)* metric (Section 9.9.2) as a standard benchmark. Research should not merely claim "privacy preservation" but must quantify the exact loss in interpretability (utility) incurred by the injection of differential privacy noise.

8.5. Emergent Interdisciplinary Themes Transcending Individual Domains

Beyond domain-specific contributions, systematic analysis reveals cross-cutting research themes that appear across smart cities, agriculture, healthcare, and critical infrastructures. These themes represent frontiers of IoT-XAI research requiring interdisciplinary attention.

8.5.1. Theme 1: Temporal Consistency and Stability of Explanations

Manifestation: Across all application domains examined in this survey, a recurring challenge emerges: ensuring that explanations for similar inputs remain stable over time despite model retraining, dataset evolution, and distributional shifts inherent in dynamic IoT environments.

Domain Evidence:

- *Smart Cities* (Section 4): Traffic prediction models retrained daily; explainability stability challenged by seasonal patterns, event-driven anomalies, and demographic shifts.
- *Smart Agriculture* (Section 6): Crop anomaly detection systems trained on multi-year datasets; explanation shifts with seasonal changes, climate patterns, and agronomic practices.
- *Federated IoT* (Section 5): Distributed models with model drift across heterogeneous local datasets; local explanations may diverge from global consensus.

Current Limitation: Standard XAI methods (LIME, SHAP) were designed for static datasets and do not provide temporal consistency guarantees. This represents a critical research gap.

8.5.2. Theme 2: Privacy–Utility–Interpretability Trilemma

Manifestation: Across all domains, tension emerges between three competing objectives:

1. **Privacy**: Protecting sensitive IoT-collected data (health records, location patterns, agricultural parameters);
2. **Utility**: Maintaining model prediction accuracy and efficiency;
3. **Interpretability**: Generating actionable explanations for decision-making.

Domain Evidence:

- Federated learning introduces privacy constraints that conflict with explanation granularity (differential privacy adds noise, obscuring causal attribution).
- Agricultural IoT with blockchain integration: Transparency requires on-chain explanations, but privacy concerns limit information sharing.
- Smart healthcare: HIPAA compliance restricts data access for explanation validation, whereas clinicians require transparency in treatment decisions.

Key Insight: The optimal balance in this trilemma is domain-specific and stakeholder-dependent. Generic solutions are likely suboptimal; context-aware trade-off mechanisms are required.

8.5.3. Theme 3: Adaptive, Human-in-the-Loop Explainability

Manifestation: Across all domains, XAI effectiveness critically depends on user characteristics (technical background, decision-making role, cognitive load) and contextual factors (time pressure, consequence severity, stakeholder diversity).

Key Observation: One-size-fits-all explanation strategies underperform. Agricultural experts require mechanistic explanations (soil physics, pest lifecycles); city planners require operational explanations (traffic flow, resource allocation); and end-citizens require simple, interpretable narratives.

Research Implication: Standardized metrics and evaluation frameworks must account for user heterogeneity, yet current benchmarks assume uniform user populations.

8.5.4. Theme 4: Edge-to-Cloud Explainability Orchestration

Manifestation: A meta-architectural question recurs across all domains: *Where should explanations be generated*?

- **Edge Devices**: Local, real-time, and privacy-preserving, but computationally expensive.
- **Cloud Systems**: Global and comprehensive; incorporates ensemble information, but introduces latency and privacy risks.
- **Hybrid Approaches**: Tiered explanations (simple at edge, detailed at cloud), but coordination complexity.

Current Status: This architectural question remains largely unaddressed in the literature. Most deployments make ad hoc decisions based on system constraints rather than principled frameworks.

8.6. Systematic Research Gaps and Priority Directions

While Sections 1–7 identify domain-specific challenges, systematic analysis reveals critical *meta-level* research gaps, which must be addressed to advance the field. These gaps represent the most promising research directions for the IoT-XAI community.

8.6.1. Gap 1: Standardization Vacuum

Problem Statement: No unified framework exists for evaluating XAI methods across heterogeneous IoT environments. Current evaluation practices are domain-specific, utilize disparate metrics, and lack standardized benchmarks.

Specific Manifestations:

- Smart cities use traffic prediction evaluation metrics; agriculture uses crop yield metrics; and there is no common ground for comparison.
- Computational constraints (latency, memory, energy) are domain-dependent; there is no standardized IoT-specific evaluation framework.
- User studies for explanation quality vary widely; there is no consensus on evaluation methodology.

This Survey's Finding: is that recent evaluation studies synthesize disparate approaches but still highlight the lack of a unified, IoT-aware evaluation standard.

Recommended Solution: Develop IOTIES (Internet of Things Interpretability Evaluation Standard)—a domain-agnostic but IoT-specific evaluation framework incorporating computational efficiency, real-time performance, heterogeneity, and user-centric dimensions.

8.6.2. Gap 2: Computational Efficiency Under-Investigation

Problem Statement: The XAI literature emphasizes explanation quality; few works systematically examine the trade-off between explanation quality and computational cost in resource- constrained IoT contexts.

Current Limitation: Computational efficiency receives attention in the hardware literature (TinyML, edge inference) and IoT systems literature (resource management), however, it is rarely coupled with explainabilitymetrics.

This Survey's Contribution: This survey identifies YONO, bLIMEy, and edge-adaptive methods as partial solutions (Section 3) but documents the absence of principled frameworks for efficiency–quality optimization.

Research Opportunity: Future research should develop formal optimization frameworks that set explanation fidelity, computational cost, latency, and energy consumption as simultaneous objectives.

8.6.3. Gap 3: Fairness in Explanations

Problem Statement: While fairness in AI models is well-studied, fairness in *the explanations themselves* remains underexplored—particularly the risk that explanations encode or amplify biases from training data or model design.

Critical Question: Can an explanation be accurate yet unfair? Example: A recidivism prediction model may accurately identify criminal history as a risk factor, but highlighting this explanation may perpetuate systemic inequities in criminal justice.

Implication for IoT: Agricultural IoT systems may provide accurate explanations of irrigation decisions that systematically disadvantage smallholder farmers due to data bias. Smart city traffic systems may provide equitable explanations only if the underlying data reflects all neighborhoods equally.

Research Opportunity: Future research should develop fairness-aware XAI frameworks, ensuring that explanations promote equitable decision-making across diverse stakeholder groups.

8.6.4. Gap 4: Multi-Stakeholder Explanation Synthesis

Problem Statement: IoT systems serve diverse stakeholders—engineers, domain experts, end-users, regulators, policymakers—with fundamentally conflicting explanation requirements. Multi-stakeholder XAI frameworks are nascent.

Example: A smart city traffic system must provide tailored explanations as follows:

- *Engineers*: Technical explanations (model internals, data lineage);
- *City Planners*: Strategic explanations (congestion patterns, resource allocation);
- *Citizens*: Transparent explanations (why my route was recommended);
- *Regulators*: Accountability explanations (discrimination detection).

Current Approach: Ad hoc development of multiple explanation types, not synthesized into coherent frameworks.

Research Opportunity: Future research should develop unified XAI architectures supporting multi-stakeholder explanation generation from a single underlying model.

9. Evaluation Metrics

Note on Proposed Metrics: The evaluation metrics presented in this section, namely, the Computational Complexity Score (CCS), Memory Footprint Ratio (MFR), and Privacy–Utility Trade-off (PUT), are introduced as conceptual formulations derived from the theoretical requirements of edge-based XAI. It is important to acknowledge that these metrics have not yet been empirically validated through large-scale experimental trials on physical IoT hardware. They are proposed here as a standardized theoretical framework to guide future research and enable consistent benchmarking across heterogeneous IoT environments. Future work will focus on the experimental validation of these formulations using real-world datasets and diverse microcontroller architectures.

The rigorous evaluation of XAI methods remains a critical challenge in establishing trustworthy and reliable explainable systems for IoT applications. This section presents a comprehensive taxonomy of evaluation metrics, introducing novel quantitative formulations that address the unique characteristics of IoT environments, including resource constraints, data heterogeneity, and real-time requirements.

Table 20 outlines a comprehensive framework of evaluation metrics for assessing XAI methods in IoT systems, organizing metrics into fidelity, stability, computational efficiency, user-centric, and domain-specific dimensions with their mathematical formulations and priority levels for different IoT deployment scenarios. Table 21 reveals that computational efficiency, privacy preservation, user evaluation protocols, and standardized benchmarking datasets remain fundamentally unstandardized across IoT domains.

Table 20. Comprehensive evaluation metrics framework for assessing Explainable Artificial Intelligence methods in IoT systems.

Metric Category	Metric Name	Formulation/Approach	IoT Context	Priority
Fidelity	Explanation Fidelity (FID)	Correlation between model f and surrogate g	All IoT Domains	High
	Infidelity Score	$\mathbb{E}[\lvert\nabla_x \log f(x) - \nabla_x \Phi(x)\rvert]$	Smart Cities	High
Stability & Consistency	Lipschitz Continuity	$\frac{\lVert\Phi(x)-\Phi(y)\rVert_2}{\lVert x-y\rVert_2} \leq L$	All IoT Domains	Medium
	Rank Correlation Stability	Spearman's ρ on ranked importances	Federated Systems	Medium
Computational Efficiency	Complexity Score (CCS)	$\log_{10}(T_\Phi / T_f)$	Edge/Constrained Devices	High
	Memory Footprint Ratio (MFR)	M_Φ / M_f	Edge/Constrained Devices	High
User-Centric Usability	Explanation Utility Score (EUS)	$\frac{\text{Acc}_{\text{XAI}} - \text{Acc}_{\text{base}}}{\text{Acc}_{\text{oracle}} - \text{Acc}_{\text{base}}}$	Smart Cities	Medium
	Cognitive Load Index (CLI)	Entropy-based feature complexity	Smart Agriculture	Medium
Domain-Specific Performance	Temporal Coherence (TC)	$1 - \frac{\sum_t \lVert\Phi(x_{t+1}) - \Phi(x_t)\rVert_2}{\sum_t \lVert\Phi(x_t)\rVert_2}$	Time-Series IoT	High
	Multi-Modal Consistency (MMC)	Average cosine similarity across sensors	Multi-Sensor Networks	Medium

Note: Φ = explanation function; f = model prediction function; T = computation time; M = memory footprint.

To bridge these standardization gaps, particularly regarding the deployment of heavy explanation algorithms on resource-constrained edge devices, we propose three theoretical metrics designed to quantify the "cost" of explainability. First, standard XAI evaluation often overlooks the latency introduced by explanation generation, which is fatal in real-time control systems. We conceptualize a *Computational Complexity Score (CCS)* to quantify the trade-off between inference latency and explanation overhead. This metric is formulated to capture the ratio of resources consumed by the explanation generator (g) relative to the

base inference model (f), weighted by the system's specific energy or latency priorities (w_1, w_2):

$$\text{CCS} = w_1 \frac{T_{\text{exp}}}{T_{\text{inf}}} + w_2 \frac{E_{\text{exp}}}{E_{\text{budget}}}, \quad w_1, w_2 \in [0,1],\ w_1 + w_2 = 1, \tag{28}$$

where T_{exp} represents the time required to generate an explanation, T_{inf} is the model inference time, E_{exp} is the energy consumed during explanation generation, and E_{budget} is the available energy budget of the device.

Table 21. Standardization gaps and evaluation approach heterogeneity across IoT application domains. ✓ = standardized; × = missing; ≈ = partial/domain-specific.

Evaluation Dimension	Smart Cities	Smart Agriculture	Federated IoT	Unification Status
Fidelity Assessment	✓ (LIME, SHAP)	✓ (Counterfactual)	✓ (SHAP-aggregated)	Partial
Computational Efficiency	≈ (Latency)	≈ (Throughput)	≈ (Bandwidth)	×
Privacy Preservation	≈ (Anonymization)	≈ (Masking)	✓ (Differential Privacy)	×
Robustness Testing	✓ (Adversarial)	≈ (Sensor Noise)	✓ (Byzantine)	Partial
User-Centric Evaluation	✓ (Interviews)	✓ (Domain Experts)	× (Unstandardized)	×
Benchmarking Datasets	✓ (Urban-scale)	✓ (Agronomic)	× (Ad hoc)	×
Real-Time Constraint Metrics	✓ (Edge)	≈ (Field)	✓ (Distributed)	Partial
Explainability Quality	✓ (Transparency Index)	✓ (Actionability Score)	≈ (Trust Metric)	Partial

Simultaneously, the strict memory limitations of microcontroller units (MCUs) in IoT sensors necessitate a rigorous check of memory consumption. Many state-of-the-art XAI methods, such as KernelSHAP, require significant memory allocation for perturbation matrices, often exceeding the SRAM limits of edge devices. To address this, we define the *Memory Footprint Ratio (MFR)* as a critical viability constraint:

$$\text{MFR} = \frac{\mathcal{M}_{\text{peak}}(f) + \mathcal{M}_{\text{peak}}(g)}{\mathcal{M}_{\text{avail}}} \times 100\%, \tag{29}$$

where $\mathcal{M}_{\text{peak}}(f)$ is the peak RAM usage of the predictive (black-box) model f, $\mathcal{M}_{\text{peak}}(g)$ is the peak RAM usage of the explanation model g, and $\mathcal{M}_{\text{avail}}$ denotes the available on-chip memory budget of the IoT device. An MFR value approaching or exceeding 100% indicates that the combined memory footprint of inference and explanation is close to or beyond the device capacity, creating a high risk of memory exhaustion and rendering the XAI configuration unsuitable for on-device deployment without compression, pruning, or offloading.

Finally, in distributed environments such as federated learning, the generation of explanations can inadvertently leak information about the local training data, creating a conflict between transparency and privacy. To evaluate this tension, we propose the *Privacy–Utility Trade-off (PUT)* metric. This formulation measures the degradation in model utility required to achieve a specific level of differential privacy (ϵ) within the explanation mechanism:

$$\text{PUT}(\epsilon, \delta) = \frac{\mathcal{A}_{\text{priv}}(\epsilon, \delta)}{\mathcal{A}_{\text{orig}}} \cdot \frac{1}{\epsilon}, \tag{30}$$

where $\mathcal{A}_{\text{orig}}$ represents the accuracy of the non-private baseline model, and $\mathcal{A}_{\text{priv}}(\epsilon, \delta)$ denotes the accuracy of the model when differential-privacy noise is injected into the explanations or training signals with privacy parameters ϵ (privacy budget) and δ (failure probability). A lower PUT value indicates that comparable accuracy is preserved under stronger privacy constraints; i.e., the privacy mechanism achieves a more favorable balance between protection of sensitive data streams and overall system performance.

9.1. Fidelity and Faithfulness Metrics

Fidelity metrics quantify the extent to which explanations accurately reflect the true decision-making process of the underlying model [53,108].

9.1.1. Infidelity Metric

The infidelity metric measures the correlation between feature importance attributions and actual prediction changes under perturbations. For an attribution method Φ and model f, infidelity is defined as

$$\text{INFID}(f, \Phi, \mathbf{x}) = \mathbb{E}_{I \sim \mu_I}\left[\left(\Phi(\mathbf{x})^T I - (f(\mathbf{x}) - f(\mathbf{x} - I))\right)^2\right] \tag{31}$$

where I represents a perturbation vector sampled from distribution μ_I, and $\mathbf{x} - I$ denotes the perturbed input.

For IoT time-series data, we extend this metric to incorporate temporal dependencies:

$$\text{INFID}_{\text{temporal}}(f, \Phi, \mathbf{x}_t) = \mathbb{E}_{I_\tau \sim \mu_{I,\tau}}\left[\left(\sum_{\tau=0}^{T} w_\tau \Phi(\mathbf{x}_{t-\tau})^T I_\tau - \Delta f_t\right)^2\right] \tag{32}$$

where $w_\tau = \exp(-\lambda\tau)$ represents temporal decay weights, T is the lookback window, and $\Delta f_t = f(\mathbf{x}_t) - f(\mathbf{x}_t - \sum_\tau I_\tau)$ captures the cumulative prediction change [52].

9.1.2. Sensitivity Metric

Sensitivity measures explanation stability under small input perturbations, critical for ensuring robustness in noisy IoT environments. For explanations Φ, sensitivity is quantified as

$$\text{SENS}(\Phi, \mathbf{x}) = \max_{\|\mathbf{r}\|_\infty \leq \epsilon} \|\Phi(\mathbf{x}) - \Phi(\mathbf{x} + \mathbf{r})\|_2 \tag{33}$$

where $\mathbf{r}$ represents bounded noise and ϵ defines the perturbation budget. For IoT deployments subject to sensor noise, we propose a probabilistic sensitivity metric:

$$\text{SENS}_{\text{prob}}(\Phi, \mathbf{x}) = \mathbb{E}_{\mathbf{r} \sim \mathcal{N}(0, \Sigma_{\text{noise}})}[\|\Phi(\mathbf{x}) - \Phi(\mathbf{x} + \mathbf{r})\|_2] \tag{34}$$

where Σ_{noise} represents the empirically estimated sensor noise covariance matrix.

9.2. Completeness and Sufficiency Metrics

Completeness metrics evaluate whether explanations capture all relevant factors influencing model decisions [182].

9.2.1. Feature Completeness Score

We define a novel Feature Completeness Score (FCS) that quantifies the proportion of model prediction variance explained by attributed features:

$$\text{FCS}(\Phi, f, \mathcal{X}) = 1 - \frac{\text{Var}(f(\mathbf{x}) - \sum_i \Phi_i(\mathbf{x}) \cdot x_i)}{\text{Var}(f(\mathbf{x}))} \tag{35}$$

where $\mathcal{X}$ denotes the evaluation dataset and Var represents variance.

For deep neural networks, where linear additivity assumptions may not hold, we extend the FCS using higher-order interaction terms:

$$\text{FCS}_{\text{interaction}}(\Phi, f, \mathcal{X}) = 1 - \frac{\text{Var}(f(\mathbf{x}) - g_\Phi(\mathbf{x}))}{\text{Var}(f(\mathbf{x}))} \tag{36}$$

where $g_\Phi(\mathbf{x}) = \sum_i \Phi_i(\mathbf{x}) \cdot x_i + \sum_{i<j} \Phi_{ij}(\mathbf{x}) \cdot x_i x_j$ incorporates pairwise interaction terms.

9.2.2. Sufficiency Metric

Sufficiency evaluates whether the features identified as important by the explanation are sufficient to reproduce model predictions. We define the Sufficiency Score as

$$\text{SUFF}(\Phi, f, \mathbf{x}, k) = 1 - \frac{|f(\mathbf{x}) - f(\mathbf{x}_{\text{top-k}})|}{|f(\mathbf{x})|} \tag{37}$$

where $\mathbf{x}_{\text{top-k}}$ retains only the top k features according to Φ and masks others with baseline values.

For IoT classification tasks, we propose a class-conditional sufficiency metric:

$$\text{SUFF}_{\text{class}}(\Phi, f, \mathbf{x}, k, c) = \frac{\exp(f_c(\mathbf{x}_{\text{top-k}}))}{\sum_j \exp(f_j(\mathbf{x}_{\text{top-k}}))} \tag{38}$$

where f_c denotes the logit for class c, and softmax normalization provides a probability score.

9.3. Complexity and Parsimony Metrics

Parsimony metrics evaluate the simplicity and comprehensibility of explanations [54].

9.3.1. Sparsity Score

The Sparsity Score quantifies the fraction of features with negligible attributions:

$$\text{SPAR}(\Phi, \mathbf{x}, \tau) = \frac{1}{d} \sum_{i=1}^{d} \mathbb{1}[|\Phi_i(\mathbf{x})| \leq \tau] \tag{39}$$

where d represents the input dimensionality, τ is a threshold defining negligibility, and $\mathbb{1}[\cdot]$ denotes the indicator function.

For IoT systems with hierarchical feature structures (e.g., grouped sensors), we introduce a Group Sparsity Score:

$$\text{GSPAR}(\Phi, \mathbf{x}) = \frac{1}{G} \sum_{g=1}^{G} \mathbb{1}\left[\|\Phi_{\mathcal{G}_g}(\mathbf{x})\|_2 \leq \tau_g\right] \tag{40}$$

where G denotes the number of feature groups, $\mathcal{G}_g$ represents the feature indices in group g, and τ_g is a group-specific threshold.

9.3.2. Explanation Complexity

We define Explanation Complexity (EC) as the minimum description length required to encode the explanation:

$$\text{EC}(\Phi, \mathbf{x}) = \sum_{i:|\Phi_i(\mathbf{x})|>\tau} [-\log_2 p_i(\Phi_i(\mathbf{x}))] \tag{41}$$

where $p_i(\cdot)$ represents the empirical distribution of attribution values for feature i across the dataset.

9.4. Consistency and Stability Metrics

Consistency metrics assess the degree to which similar inputs receive similar explanations [182].

9.4.1. Lipschitz Continuity of Explanations

We quantify explanation stability through the Lipschitz constant of the explanation function:

$$L_\Phi = \sup_{\mathbf{x} \neq \mathbf{y}} \frac{\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|_2}{\|\mathbf{x} - \mathbf{y}\|_2} \tag{42}$$

For practical estimation, we compute

$$\hat{L}_\Phi = \max_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}} \frac{\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_j)\|_2}{\|\mathbf{x}_i - \mathbf{x}_j\|_2} \tag{43}$$

where $\mathcal{P}$ denotes a set of input pairs sampled from the data distribution.

9.4.2. Rank Correlation Stability

For assessing the consistency of feature importance rankings across similar inputs, we propose the Rank Correlation Stability (RCS) metric:

$$\mathrm{RCS}(\Phi, \mathcal{N}_\mathbf{x}) = \frac{2}{|\mathcal{N}_\mathbf{x}|(|\mathcal{N}_\mathbf{x}| - 1)} \sum_{i<j} \rho\big(\mathrm{rank}(\Phi(\mathbf{x}_i)), \mathrm{rank}(\Phi(\mathbf{x}_j))\big) \tag{44}$$

where $\mathcal{N}_\mathbf{x}$ represents a neighborhood of similar inputs around $\mathbf{x}$, $\mathrm{rank}(\cdot)$ computes feature importance rankings, and $\rho(\cdot)$ denotes Spearman's rank correlation coefficient.

9.5. Robustness to Adversarial Perturbations

IoT systems are susceptible to adversarial attacks, necessitating evaluation of explanation robustness under adversarial conditions [131].

Adversarial Explanation Robustness

We define the Adversarial Explanation Robustness (AER) metric as

$$\mathrm{AER}(\Phi, f, \mathbf{x}) = \min_{\|\delta\|_p \leq \epsilon} \frac{\|\Phi(\mathbf{x}) - \Phi(\mathbf{x} + \delta)\|_2}{\|\Phi(\mathbf{x})\|_2} \tag{45}$$

where δ represents an adversarial perturbation bounded by ϵ in L_p norm.

9.6. Efficiency and Scalability Metrics

Resource constraints in IoT necessitate the evaluation of computational efficiency and scalability of XAI methods [116].

9.6.1. Computational Complexity Score

We define the Computational Complexity Score (CCS) as

$$\mathrm{CCS}(\Phi, \mathbf{x}) = \log_{10}\left(\frac{T_\Phi(\mathbf{x})}{T_f(\mathbf{x})}\right) \tag{46}$$

where $T_\Phi(\mathbf{x})$ denotes the wall-clock time required to compute the explanation, and $T_f(\mathbf{x})$ represents the model inference time.

For multi-sample batch explanations, we extend this to

$$\mathrm{CCS}_{\mathrm{batch}}(\Phi, \mathcal{X}) = \log_{10}\left(\frac{T_\Phi(\mathcal{X})}{|\mathcal{X}| \cdot T_f(\mathbf{x}_1)}\right) \tag{47}$$

9.6.2. Memory Footprint Ratio

For edge IoT devices with limited memory, we define the Memory Footprint Ratio (MFR):

$$\mathrm{MFR}(\Phi) = \frac{M_\Phi}{M_f} \tag{48}$$

where M_Φ represents the peak memory usage during explanation computation, and M_f denotes the model's memory footprint.

9.7. User-Centric Evaluation Metrics

Beyond computational metrics, user-centric evaluations assess the practical utility and comprehensibility of explanations for IoT stakeholders [130].

9.7.1. Explanation Utility Score

We propose an Explanation Utility Score (EUS) based on user task performance improvements:

$$\mathrm{EUS} = \frac{\mathrm{Acc}_{\mathrm{with_XAI}} - \mathrm{Acc}_{\mathrm{without_XAI}}}{\mathrm{Acc}_{\mathrm{oracle}} - \mathrm{Acc}_{\mathrm{without_XAI}}} \tag{49}$$

where $\mathrm{Acc}_{\mathrm{with_XAI}}$ and $\mathrm{Acc}_{\mathrm{without_XAI}}$ represent user task accuracy with and without explanations, respectively, and $\mathrm{Acc}_{\mathrm{oracle}}$ denotes the theoretical maximum accuracy.

9.7.2. Cognitive Load Index

To assess explanation comprehensibility, we define a Cognitive Load Index (CLI) based on information-theoretic principles:

$$\mathrm{CLI}(\Phi, \mathbf{x}) = -\sum_{i:|\Phi_i(\mathbf{x})|>\tau} \frac{|\Phi_i(\mathbf{x})|}{\|\Phi(\mathbf{x})\|_1} \log_2 \frac{|\Phi_i(\mathbf{x})|}{\|\Phi(\mathbf{x})\|_1} \tag{50}$$

This entropy-based metric captures the distribution of attribution mass across features.

9.8. Domain-Specific IoT Evaluation Metrics

IoT applications exhibit unique characteristics necessitating domain-specific evaluation formulations.

9.8.1. Temporal Coherence for Time-Series IoT

For IoT time-series explanations, we introduce Temporal Coherence (TC), quantifying the smoothness of attributions over time:

$$\mathrm{TC}(\Phi, \{\mathbf{x}_t\}_{t=1}^T) = 1 - \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{\|\Phi(\mathbf{x}_{t+1}) - \Phi(\mathbf{x}_t)\|_2}{\|\Phi(\mathbf{x}_t)\|_2} \tag{51}$$

For abrupt regime changes, we incorporate change-point detection:

$$\mathrm{TC}_{\mathrm{adaptive}} = 1 - \frac{1}{|\mathcal{T}_{\mathrm{smooth}}|} \sum_{t \in \mathcal{T}_{\mathrm{smooth}}} \frac{\|\Phi(\mathbf{x}_{t+1}) - \Phi(\mathbf{x}_t)\|_2}{\|\Phi(\mathbf{x}_t)\|_2} \tag{52}$$

where $\mathcal{T}_{\mathrm{smooth}}$ excludes detected change-points.

9.8.2. Multi-Modal Consistency for Heterogeneous IoT

IoT systems often integrate heterogeneous data modalities (e.g., temperature, humidity, images). We define Multi-Modal Consistency (MMC) as

$$\mathrm{MMC}(\{\Phi^{(m)}\}_{m=1}^{M}, \mathbf{x}) = \frac{1}{M(M-1)} \sum_{i \neq j} \cos\left(\Phi^{(i)}(\mathbf{x}^{(i)}), \Phi^{(j)}(\mathbf{x}^{(j)})\right) \tag{53}$$

where M denotes the number of modalities, and $\cos(\cdot)$ computes cosine similarity.

9.9. Federated XAI Evaluation Metrics

Federated IoT systems require evaluation metrics that account for distributed data and privacy constraints.

9.9.1. Global–Local Explanation Divergence

We quantify the divergence between global federated explanations and local node-specific explanations:

$$\mathrm{GLED}(\Phi_{\text{global}}, \{\Phi_i\}_{i=1}^{N}) = \frac{1}{N} \sum_{i=1}^{N} D_{\mathrm{KL}}\left(P_{\Phi_i} \| P_{\Phi_{\text{global}}}\right) \tag{54}$$

where P_{Φ_i} and $P_{\Phi_{\text{global}}}$ represent probability distributions over feature importance, and D_{KL} denotes Kullback–Leibler divergence [93].

9.9.2. Privacy–Utility Trade-Off Metric

For federated XAI with differential privacy, we define a Privacy–Utility Trade-off (PUT) metric:

$$\mathrm{PUT}(\epsilon, \delta) = \frac{\mathrm{ExplanationUtility}(\Phi_{\epsilon,\delta})}{\epsilon \cdot \log(1/\delta)} \tag{55}$$

where ExplanationUtility quantifies explanation quality, and ϵ, δ parameterize differential privacy guarantees.

9.10. Axiomatic Evaluation Framework

Drawing from cooperative game theory and axiomatic principles, we formalize desirable properties that XAI methods should possess [74].

Axiom Compliance Score

We define an Axiom Compliance Score (ACS) assessing adherence to fundamental XAI axioms:

$$\mathrm{ACS}(\Phi) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathbb{1}[\Phi \text{ satisfies axiom } a] \tag{56}$$

where $\mathcal{A} = \{\text{Sensitivity}, \text{Implementation Invariance}, \text{Linearity}, \text{Symmetry}\}$ denotes the set of axioms.

9.11. Unified Multi-Criteria Evaluation Framework

To facilitate holistic XAI method comparison, we propose a Unified Multi-Criteria Evaluation Framework (UMCEF) aggregating multiple metrics:

$$\mathrm{UMCEF}(\Phi) = \sum_{k=1}^{K} w_k \cdot \mathcal{M}_k(\Phi) \tag{57}$$

where $\mathcal{M}_k$ represents normalized evaluation metric k, w_k denotes user-specified importance weights satisfying $\sum_k w_k = 1$, and K is the total number of criteria. For IoT deployments, we recommend prioritizing computational efficiency and robustness alongside fidelity:

$$\text{UMCEF}_{\text{IoT}}(\Phi) = 0.4 \cdot \text{FID}(\Phi) + 0.3 \cdot \text{EFF}(\Phi) + 0.2 \cdot \text{ROB}(\Phi) + 0.1 \cdot \text{SPAR}(\Phi) \quad (58)$$

where FID, EFF, ROB, and SPAR represent normalized fidelity, efficiency, robustness, and sparsity scores.

9.12. Novel Composite Metrics for IoT-XAI

To address the multifaceted requirements of IoT explainability, we introduce composite metrics integrating multiple evaluation dimensions.

9.12.1. Fidelity–Efficiency Frontier

We define the Fidelity–Efficiency Frontier (FEF) as the Pareto-optimal trade-off surface

$$\text{FEF} = \left\{ (\text{FID}(\Phi), \text{EFF}(\Phi)) : \nexists \Phi' \text{ s.t. } \text{FID}(\Phi') > \text{FID}(\Phi) \wedge \text{EFF}(\Phi') > \text{EFF}(\Phi) \right\} \quad (59)$$

Methods on this frontier represent non-dominated solutions balancing explanation quality and computational cost. For IoT deployment, we quantify proximity to the frontier using the hypervolume indicator:

$$\text{HV}(\Phi) = \int_{\mathcal{R}} \mathbb{1}\left[\exists \Phi' \in \text{FEF} : \text{FID}(\Phi') \geq \text{fid} \wedge \text{EFF}(\Phi') \geq \text{eff}\right] d(\text{fid}, \text{eff}) \quad (60)$$

where $\mathcal{R}$ denotes the objective space region dominated by the method.

9.12.2. Explanation Trustworthiness Index

We synthesized multiple dimensions into a holistic Explanation Trustworthiness Index(ETI):

$$\text{ETI}(\Phi) = \sqrt[3]{\text{FID}(\Phi) \cdot \text{CONS}(\Phi) \cdot \text{ROB}(\Phi)} \quad (61)$$

where FID, CONS, andROB represent the normalized fidelity, consistency, androbustness scores, respectively. Foruser-facing IoT applications, we extend the ETI to incorporate user-centric factors as follows:

$$\text{ETI}_{\text{user}}(\Phi) = \sqrt[4]{\text{FID}(\Phi) \cdot \text{CONS}(\Phi) \cdot \text{ROB}(\Phi) \cdot \text{COMP}(\Phi)} \quad (62)$$

where COMP represents comprehensibility.

10. Challenges and Future Directions

The integration of Internet of Things (IoT) systems with Explainable Artificial Intelligence (XAI) entails critical challenges in scalability, computational efficiency, interpretability, and security, as well as the lack of standardized evaluation frameworks [183,184]. Figure 6 highlights emerging convergence frontiers in XAI–IoT research, organized around three pillars: Causal Semantics, which addresses mechanistic causal inference and knowledge graph integration; Distributed Cognition, promoting privacy-preserving transparency via federated learning and multi-agent systems; and Human-Centric Ethics, encompassing cognitive-load assessment, fairness auditing, and stakeholder-driven design. Together, these dimensions define a nexus for IoT–XAI integration, emphasizing edge-time processing and standardized evaluation. Transformative research frontiers include Quantum XAI, neuro-symbolic architectures, self-explaining systems, and temporal causality discovery, all of which extend beyond conventional transparency-focused paradigms.

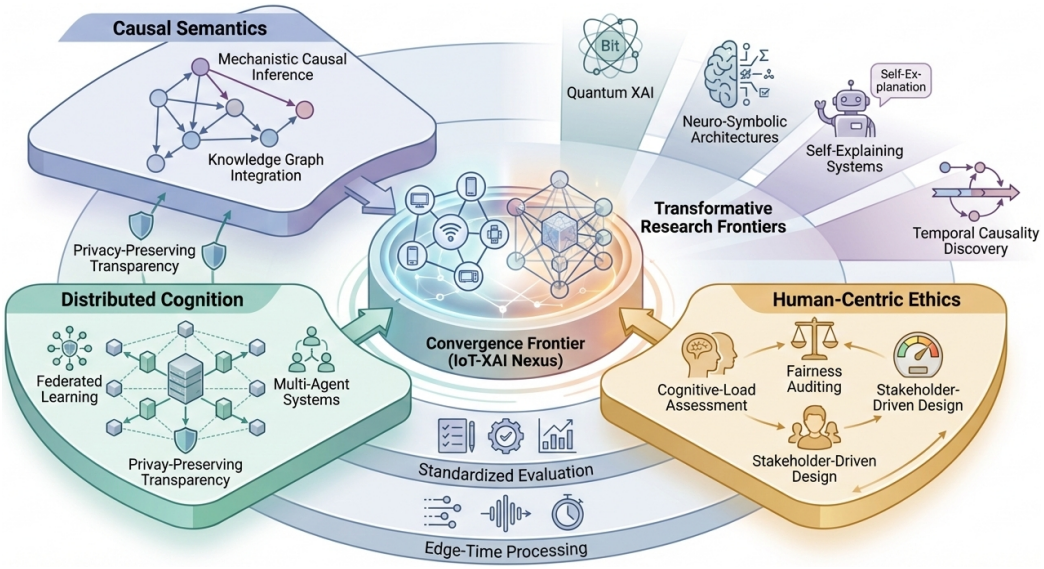


Figure 6. Convergence frontiers: Beyond traditional XAI-IoT paradigms.

10.1. Key Challenges in IoT and XAI Integration

The convergence of IoT systems and XAI introduces significant challenges, including scalability issues in managing large datasets, computational constraints affecting real-time processing, and the necessity for interpretability to ensure user trust [51,113,185]. Scalability is particularly challenging in IoT environments with extensive networks of heterogeneous devices, demanding efficient data processing and resource management [112]. Computational constraints arise due to the intensive nature of XAI methods like kernelSHAP, which struggle with large datasets [185]. Interpretability challenges persist, as existing XAI methods often fail to provide intuitive insights, limiting usability [17]. Security and privacy concerns are exacerbated by the interconnected nature of IoT devices, necessitating adaptive privacy mechanisms [19]. Additionally, the lack of standardized evaluation frameworks hinders the assessment and comparison of XAI methodologies [11]. Addressing these challenges is crucial for the successful integration of XAI in IoT applications. Table 22 synthesizes the primary challenges in IoT-XAI integration across computational constraints, privacy and security, interpretability, scalability, data quality, and standardization domains, mapping each challenge to state-of-the-art solutions while assessing implementation complexity, effectiveness, and research maturity to identify promising research directions.

Table 22. IoT-XAI integration: Challenge–solution mapping and research maturity assessment.

Challenge Domain	Specific Challenge	State-of-the-Art Solution	Complexity	Effectiveness	Maturity
Computational Constraints	Real-time edge processing High memory footprint	YONO, Edge Computing, TinyML KernelSHAP, TreeSHAP Variants	Medium Low	High Medium	Mature Mature
Privacy & Security	Data leakage in FL Adversarial attacks on explanations	Differential Privacy, Secure Aggregation AER Metrics	High High	High Medium	Developing Early Stage
Interpretability	Tech–user explanation gap Lack of actionable insights	CLE-XAI, Custom LLMs Counterfactuals, bLIMEy	Medium Medium	High Medium	Developing Early Stage
Scalability	Heterogeneous IoT device management Large-scale coordination	Fog Computing, Hierarchical Designs Adaptive Offloading, UAV-assisted FL	High Very High	High High	Mature Developing
Data Quality	Sensor reliability issues	Data Quality Models, Sensor Validation	Medium	Medium	Developing
Standardization	No unified XAI evaluation frameworks	UMCEF, EXACT, XAI-Benchmark	Low	Medium	Early Stage

10.2. Ethical and Societal Considerations

Integrating IoT systems with XAI raises ethical and societal concerns regarding privacy, accountability, trust, and equity. The black-box nature of AI algorithms necessitates transparency to foster trust and informed consent [51,113]. Human biases in trusting intuitive explanations provided by XAI systems can lead to misplaced trust, highlighting the need for rigorous evaluation frameworks [186]. Privacy and data security are paramount, especially in sensitive applications like personalized medicine [187]. Strategies to manage "unfortunate counterfactual events" are essential for mitigating ethical dilemmas [188]. Interdisciplinary approaches integrating ethics, sociology, and technology are crucial to ensuring equitable and socially responsible deployment of IoT and XAI [189]. Addressing these considerations is vital for aligning technological advancements with societal values.

10.3. Cross-Domain Applications and Scalability

The integration of IoT and XAI holds promise for cross-domain applications in agriculture, healthcare, and education, enhancing decision-making through transparent insights [73]. In agriculture, IoT and UAVs support precision farming by enabling real-time environmental monitoring [171]. Scalability is facilitated by advancements in wireless communication and edge computing. In education, integrating causal inference in knowledge tracing enhances personalized learning experiences [43]. Refining clustering algorithms and extending benchmarks improve XAI applicability across domains [54]. Interdisciplinary approaches integrating cognitive science and ethics are essential for advancing cross-domain applications [73]. Collaboration and innovation will drive progress and sustainability across sectors.

10.4. Standardization and Evaluation Frameworks

The integration of IoT and XAI necessitates standardized evaluation frameworks to ensure method selection tailored to specific applications. These frameworks enhance transparency, reliability, and usability across domains [190]. Standardization efforts aim to establish guidelines for evaluating XAI methods, providing a unified taxonomy for comparison [50,90]. Evaluation frameworks address computational efficiency, scalability, and robustness, guiding the development of adaptable solutions [190]. They also facilitate identification and mitigation of biases, promoting ethical deployment [43,53]. Implementing standardized frameworks is critical for fostering trust and usability in IoT applications.

10.5. Future Directions for XAI in IoT Systems

Future research should prioritize developing robust XAI frameworks that enhance interpretability, scalability, security, and adaptability in diverse applications like agriculture and healthcare [33]. Establishing standardized evaluation frameworks is essential for addressing challenges in different data types [43,191]. Exploring robust evaluation metrics that prioritize user utility is crucial for extending benchmarks to complex problem settings [19]. Understanding consumer preferences for explanations informs design to mitigate biases [111]. Security concerns necessitate device-independent access control solutions and innovative risk assessment methodologies [38]. Interdisciplinary approaches integrating social, technical, and ethical considerations will improve trust and usability [43,51]. By addressing these directions, XAI in IoT systems can achieve robustness, scalability, and trustworthiness, fostering widespread adoption and aligning advancements with societal goals.

Ultimately, Table 23 summarizes the future research agenda for XAI in IoT systems. Furthermore, the open challenges are mapped to representative research directions, evaluation metrics, and priority application domains.

Table 23. Future research agenda for Explainable Artificial Intelligence (XAI) in Internet of Things (IoT) systems.

Open Challenge	Indicative Research Directions	Representative Evaluation Metrics	Priority Domains
Scalable and resource-efficient XAI for edge devices	Lightweight and sparsity-aware explainers, adaptive granularity, event-driven and TinyML-compatible XAI	Complexity ratio, memory footprint, latency, temporal coherence	Smart cities, smart homes, environmental sensing
Privacy-preserving and federated explainability	Federated concept learning, differentially private explanations, cross-silo consensus mechanisms	Privacy–Utility Trade-off, rank correlation stability, inference attack resistance	Healthcare, finance, industrial IoT
Standardized and domain-sensitive XAI evaluation	Task-specific metrics, human-centered validation studies, benchmark development for IoT modalities	Fidelity/infidelity, utility scores, cognitive load, fairness indices	Smart governance, agriculture, cyber–physical systems
Robustness to adversarial and distributional shifts	Adversarially robust explainers, drift detection, robustness-aware deployment pipelines	Sensitivity and stability, adversarial robustness, drift detection rate	Critical infrastructures, autonomous systems, IoT security
Human-centric and participatory XAI design	Stakeholder co-design, adaptive interfaces, counterfactual and narrative explanations	Trust and acceptance indices, task performance gain, satisfaction metrics	Smart cities, healthcare support systems, education
Interoperability and standardization of IoT–XAI	Common ontologies, interoperable explanation APIs, regulatory alignment	Ontology coverage, interoperability level, compliance indicators	Cross-vendor IoT, smart infrastructure, data spaces

11. Conclusions

The integration of Internet of Things (IoT) systems and Explainable Artificial Intelligence (XAI) is one of the most important milestones in addressing the complexity of modern-day technological environments. This survey shows how essential IoT frameworks are in improving the urban infrastructure, resource management, and agricultural productivity, as well as the invaluable role of XAI in guaranteeing transparency, trust, and actionable insights in diverse applications. With the sophistication of information skyrocketing and the need for machine interpretability in automated systems increasing, the convergence of IoT and XAI is a critical area of research that can not only improve operational efficiency but also promote stakeholder participation and improved decision-making.

XAI methods that are implemented in IoT systems have significantly enhanced decision-making and accountability in the context of smart cities. Frameworks like SOFIE are examples of flawless interoperability between IoT platforms and promote the development of innovative business models without compromising privacy and security. The constellation of various sources of data will increase the functionality of urban systems, allowing real-time flexibility in response to changing needs. In the same way, the improvements in decentralized coordination show high efficiency in task allocation, scalability, and flexibility, with an emphasis placed on the potential of XAI to optimize decentralized IoT systems. Additionally, integration of XAI into prediction systems has brought about massive reinforcement of model predictive capability, which justifies the application of explainability in urban predictive models. Such developments not only streamline the management of resourcing but also encourage citizen involvement in making decisions that are based on more transparent and understandable data.

The combination of IoT and XAI has brought changes to farming processes in the sphere of smart agriculture by improving precision farming, real-time monitoring, and large-scale resource management. IoT-XAI technologies permit the identification of trends in both environmental and crop data, which reduces model interpretability and choice. This ability plays a crucial role in supporting farmers with the aim of maximiz-

ing yield with minimum environmental damage. The autonomous optimization method, in which reinforcement learning is applied, has also extended battery life in resource-constrained systems with enough predictability to allow more sustainable and efficient uses in agriculture. These inventions emphasize the necessity of incorporating IoT-XAI to achieve Sustainable Development Goals (SDGs) by improving agricultural productivity and food security, and address current issues by establishing the foundation of sustainable practices in the future.

Yet another issue in IoT systems revealed in the survey is the requirement for methods of XAI to handle challenges so that they can be trusted and perform effectively. This is important for ensuring strong performance in unpredictable and dynamic environments. Moreover, user-friendly XAI tools have greatly contributed to user comprehension of AI decision-making mechanisms, which highlights the significance of virtual and convenient XAI structure designs. These products not only enhance the user experience but also instill higher levels of trust in using automated systems, which is essential for mass adoption.

We expect that the interdisciplinary convergence of IoT systems and XAI will play a huge role in the development of smart cities and agriculture applications in the future. Any further studies in this area should consider more scalability, computational performance, or ethical issues to establish fair and transparent implementations of IoT and XAI technologies. The combination of IoT systems and XAI, besides enhancing model explainability and predictive accuracy, is vital in building trust and enabling technological innovations in society to meet the target of global sustainability. This interdisciplinary field can be used to elevate the creativity, sustainability, and well-being of societies in the age of related intelligence by utilizing interoperability, semantic learning, and user-friendly designs.

In conclusion, in pursuing these research directions, researchers, practitioners, and policymakers must work tirelessly to ensure that these technologies are exploited responsibly and efficiently as they seek to establish a more interconnected and sustainable future.

Author Contributions: A.K., A.G., N.A., and C.K. conceived the idea, designed and performed the experiments, analyzed the results, and drafted the initial manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

adaPARL	Adaptive Privacy-Aware Reinforcement Learning
AER	Adversarial Explanation Robustness
AI	Artificial Intelligence
ASFA	Adaptive Sparsity Feature Attribution
AutoXAI	Automated Explainable Artificial Intelligence
bLIMEy	Building Local Interpretable Model-agnostic Explanations for You
CAManim	Class Activation Map Animation
CAV	Concept Activation Vector
CCS	Computational Complexity Score
CFN	Channel-Wise Feature Normalization
CLASH	Counterfactual Local Approximate Shapley
CLE-XAI	Concept-Level Explainable Artificial Intelligence
CLI	Cognitive Load Index
CNN	Convolutional Neural Network

DCNE	Deep Convolutional Neural Explanation
DLT	Distributed Ledger Technology
DNN	Deep Neural Network
DRL	Deep Reinforcement Learning
EFL	Explanation-aware Federated Learning
EXACT	Explainable Artificial Intelligence Comprehensive Taxonomy
FED-XAI	Federated Explainable Artificial Intelligence
FedHDPrivacy	Federated Hierarchical Differential Privacy
FedSaC	Federated Sample Complementarity
FID	Fidelity
FL	Federated Learning
GenAI	Generative Artificial Intelligence
Grad-CAM	Gradient-weighted Class Activation Mapping
HFL	Hierarchical Federated Learning
HIFL	High-Integrity Federated Learning
IG	Integrated Gradients
IoT	Internet of Things
LAN	Local Area Network
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
LRFAM	Layer-wise Relevance Feature Attribution Method
mabCAM	Multiple Attention-Based Class Activation Map
MFR	Memory Footprint Ratio
ML	Machine Learning
MMC	Multi-Modal Consistency
NCAV	Non-negative Concept Activation Vector
OAK4XAI	Ontology-based Knowledge Map Model for XAI
PRA	Pattern Reconfigurable Antenna
PSEM	Plausible and Smooth Explanation Method
SDG	Sustainable Development Goal
SECOE	Sensor Continuity and Operational Efficiency
SHAP	SHapley Additive exPlanations
ShapG	Shapley Gradient
SOFIE	Secure Open Federation for Internet Everywhere
TC	Temporal Coherence
TCAV	Testing with Concept Activation Vectors
TinyML	Tiny Machine Learning
UAV	Unmanned Aerial Vehicle
UCSMS	Ultra-Compact Soil Moisture Sensor
UMCEF	Unified Multi-Criteria Evaluation Framework
vPPG	Video-based Photoplethysmography
XAI	Explainable Artificial Intelligence
XXAI	Extended Explainable Artificial Intelligence
YONO	You Only Need One

References

1. Alotaibi, A.; Aldawghan, H.; Aljughaiman, A. A review of the authentication techniques for internet of things devices in smart cities: Opportunities, challenges, and future directions. *Sensors* **2025**, *25*, 1649. [CrossRef]
2. Saadouni, C.; El Jaouhari, S.; Tamani, N.; Ziti, S.; Mroueh, L.; El Bouchti, K. Identification techniques in the internet of things: Survey, taxonomy and research frontier. *IEEE Commun. Surv. Tutor.* **2025**, *28*, 593–632. [CrossRef]
3. Zong, M.; Hekmati, A.; Guastalla, M.; Li, Y.; Krishnamachari, B. Integrating large language models with internet of things: applications. *Discov. Internet Things* **2025**, *5*, 2. [CrossRef]
4. Vishwakarma, A.K.; Chaurasia, S.; Kumar, K.; Singh, Y.N.; Chaurasia, R. Internet of things technology, research, and challenges: A survey. *Multimed. Tools Appl.* **2025**, *84*, 8455–8490. [CrossRef]

5. Miller, T.; Durlik, I.; Kostecka, E.; Kozlovska, P.; Łobodzińska, A.; Sokołowska, S.; Nowy, A. Integrating artificial intelligence agents with the internet of things for enhanced environmental monitoring: Applications in water quality and climate data. *Electronics* **2025**, *14*, 696. [CrossRef]
6. ŞAHiN, E.; Arslan, N.N.; Özdemir, D. Unlocking the black box: An in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Comput. Appl.* **2025**, *37*, 859–965. [CrossRef]
7. Vouros, G.A. Explainable deep reinforcement learning: State of the art and challenges. *ACM Comput. Surv.* **2022**, *55*, 1–39. [CrossRef]
8. Von Eschenbach, W.J. Transparency and the black box problem: Why we do not trust AI. *Philos. Technol.* **2021**, *34*, 1607–1622. [CrossRef]
9. Kalmykov, V.L.; Kalmykov, L.V. XXAI: Towards Explicitly Explainable Artificial Intelligence. *arXiv* **2024**, arXiv:2401.06802.
10. Bie, Y.; Luo, L.; Chen, Z.; Chen, H. XCoOP: Explainable Prompt Learning for Computer-Aided Diagnosis via Concept-Guided Context Optimization. *arXiv* **2024**, arXiv:2401.12345.
11. Phillips, P.J.; Hahn, C.A.; Fontana, P.C.; Yates, A.N.; Greene, K.; Broniatowski, D.A.; Przybocki, M.A. *Four Principles of Explainable Artificial Intelligence*; NIST Interagency/Internal Report (NISTIR); National Institute of Standards and Technology: Gaithersburg, MD, USA, 2021.
12. Naresh, M.; Munaswamy, P. Smart Agriculture System Using IoT Technology. *Int. J. Recent Technol. Eng.* **2019**, *7*, 98–102.
13. Jagatheesaperumal, S.K.; Pham, Q.V.; Ruby, R.; Yang, Z.; Xu, C.; Zhang, Z. Explainable AI over the Internet of Things (IoT): Overview, state-of-the-art and future directions. *IEEE Open J. Commun. Soc.* **2022**, *3*, 2106–2136.
14. Mohseni, S.; Zarei, N.; Ragan, E.D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans. Interact. Intell. Syst.* **2021**, *11*, 24. [CrossRef]
15. Lopes, P.; Silva, E.; Braga, C.; Oliveira, T.; Rosado, L. XAI systems evaluation: A review of human and computer-centred methods. *Appl. Sci.* **2022**, *12*, 9423. [CrossRef]
16. Ebermann, C.; Selisky, M.; Weibelzahl, S. Explainable AI: The effect of contradictory decisions and explanations on users' acceptance of AI systems. *Int. J. Hum.-Interact.* **2023**, *39*, 1807–1826.
17. Gilpin, L.H.; Paley, A.R.; Alam, M.A.; Spurlock, S.; Hammond, K.J. "Explanation" is Not a Technical Term: The Problem of Ambiguity in XAI. *arXiv* **2022**, arXiv:2203.14196. [CrossRef]
18. Adeyinka, T.I.; Adeyinka, K.I.; Emmanuel, A.A. Security, Privacy, and Trust of AI-IoT Convergent Smart System. In *Humans and Generative AI Tools for Collaborative Intelligence*; IGI Global Scientific Publishing: Hershey, PA, USA, 2025; pp. 129–160.
19. Taherisadr, M.; Stavroulakis, S.A.; Elmalaki, S. AdaPARL: Adaptive Privacy-Aware Reinforcement Learning for Sequential-Decision Making Human-in-the-Loop Systems. In Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementatio, San Antonio, TX, USA, 9–12 May 2023.
20. Almaazmi, K.I.A.; Almheiri, S.J.; Khan, M.A.; Shah, A.A.; Abbas, S.; Ahmad, M. Enhancing smart city sustainability with explainable federated learning for vehicular energy control. *Sci. Rep.* **2025**, *15*, 23888. [CrossRef] [PubMed]
21. Maralapalle, V.; Muktinutalapati, J.; Tammineni, G.; Krishna, M.V.N.; Marlapalle, S. Smart Cities, Smarter Solutions: AI for Urban Transformation. In *Intelligent Systems for Sustainable Infrastructure: AI Solutions Shaping a Green Future—Leveraging AI Innovations for Eco Friendly Infrastructure and Environmental Resilience*; Springer Nature: Cham, Switzerland, 2025; pp. 1–24. [CrossRef]
22. Alaslani, R.; Perzhilla, L.; Rahman, M.M.U.; Laleg-Kirati, T.M.; Al-Naffouri, T.Y. You Can Monitor Your Hydration Level Using Your Smartphone Camera. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 4013214. [CrossRef]
23. Gunning, D. *Explainable Artificial Intelligence (XAI)*; Defense Advanced Research Projects Agency (DARPA): Arlington County, VA, USA, 2017; pp. 1–2.
24. Martin, R.J.; Mittal, R.; Malik, V.; Jeribi, F.; Siddiqui, S.T.; Hossain, M.A.; Swapna, S. XAI-powered smart agriculture framework for enhancing food productivity and sustainability. *IEEE Access* **2024**, *12*, 168412–168427. [CrossRef]
25. Miller, T.; Mikiciuk, G.; Durlik, I.; Mikiciuk, M.; Łobodzińska, A.; Śnieg, M. The IoT and AI in Agriculture: The Time Is Now—A Systematic Review of Smart Sensing Technologies. *Sensors* **2025**, *25*, 3583.
26. Gummadi, A.N.; Napier, J.C.; Abdallah, M. XAI-IoT: An explainable AI framework for enhancing anomaly detection in IoT systems. *IEEE Access* **2024**, *12*, 71024–71054. [CrossRef]
27. Kaur, N.; Gupta, L. Securing the 6G–IoT Environment: A Framework for Enhancing Transparency in Artificial Intelligence Decision-Making Through Explainable Artificial Intelligence. *Sensors* **2025**, *25*, 854. [CrossRef]
28. Adam, M.; Hammoudeh, M.; Alrawashdeh, R.; Alsulaimy, B. A survey on security, privacy, trust, and architectural challenges in IoT systems. *IEEE Access* **2024**, *12*, 57128–57149. [CrossRef]
29. Al Khatib, I.; Shamayleh, A.; Ndiaye, M. Healthcare and the internet of medical things: Applications, trends, key challenges, and proposed resolutions. *Informatics* **2024**, *11*, 47. [CrossRef]
30. Lim, K.S.; Ooi, S.Y.; Sayeed, M.S.; Chew, Y.J.; Ahmad, N.M. Securing the Internet of Things: Systematic Insights into Architectures, Threats, and Defenses. *Electronics* **2025**, *14*, 3972. [CrossRef]

31. Eren, H.; Karaduman, Ö.; Gençoğlu, M.T. Security and Privacy in the Internet of Everything (IoE): A Review on Blockchain, Edge Computing, AI, and Quantum-Resilient Solutions. *Appl. Sci.* **2025**, *15*, 8704. [CrossRef]
32. Sebestyen, H.; Popescu, D.E.; Zmaranda, R.D. A literature review on security in the Internet of Things: Identifying and analysing critical categories. *Computers* **2025**, *14*, 61. [CrossRef]
33. Corcuera Bárcena, J.L.; Daole, M.; Ducange, P.; Marcelloni, F.; Renda, A.; Ruffini, F.; Schiavo, A. Fed-XAI: Federated Learning of Explainable Artificial Intelligence Models. In Proceedings of the XAI.it @ AI*IA Workshop, Udine, Italy, 28 November–3 December 2022; pp. 104–117.
34. Karras, A.; Karras, C.; Giotopoulos, K.C.; Tsolis, D.; Oikonomou, K.; Sioutas, S. Peer to Peer Federated Learning: Towards Decentralized Machine Learning on Edge Devices. In Proceedings of the 2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Ioannina, Greece, 23–25 September 2022; pp. 1–9. [CrossRef]
35. Bedewy, S.F. The impact of data security and privacy concerns on the implementation of integrated. In *Smart Cities: Foundations and Perspectives*; IntechOpen Limited: London, UK, 2024; Volume 59.
36. Bures, M.; Bellekens, X.; Frajtak, K.; Ahmed, B.S. A Comprehensive View on Quality Characteristics of the IoT Solutions. In Proceedings of the 3rd EAI International Conference on IoT in Urban Space, Guimarães, Portugal, 21–23 November 2018.
37. Alur, R.; Berger, E.; Drobnis, A.W.; Fix, L.; Fu, K.; Hager, G.D.; Lopresti, D.; Nahrstedt, K.; Mynatt, E.; Patel, S.; et al. *Systems Computing Challenges in the Internet of Things*; Computing Community Consortium (CCC) Report; Computing Community Consortium: Washington, DC, USA, 2016.
38. Ghasemshirazi, S.; Shirvani, G. Securing the Future: Proactive Threat Hunting for Sustainable IoT Ecosystems. *Comput. Secur.* **2024**, *138*, 103678.
39. Al-Garadi, M.A.; Mohamed, A.; Al-Ali, A.; Du, X.; Guizani, M. A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 2399–2433. [CrossRef]
40. Elkhawaga, G.; Abuelkheir, M.; Reichert, M. XAI in the Context of Predictive Process Monitoring: Too Much to Reveal. *arXiv* **2022**, arXiv:2202.04567. [CrossRef]
41. Oliveira, R.M.B.d.; Goethals, S.; Brughmans, D.; Martens, D. Unveiling the Potential of Counterfactuals Explanations in Employability. *arXiv* **2023**, arXiv:2305.10069. [CrossRef]
42. Bertoglio, R.; Corbo, C.; Renga, F.M.; Matteucci, M. The Digital Agricultural Revolution: A Bibliometric Analysis Literature Review. *Agronomy* **2021**, *10*, 519. [CrossRef]
43. Bai, Y.; Zhao, J.; Wei, T.; Cai, Q.; He, L. A Survey of Explainable Knowledge Tracing. *arXiv* **2024**, arXiv:2402.15313. [CrossRef]
44. Javed, A.R.; Ahmed, W.; Pandya, S.; Maddikunta, P.K.R.; Alazab, M.; Gadekallu, T.R. A survey of explainable artificial intelligence for smart cities. *Electronics* **2023**, *12*, 1020. [CrossRef]
45. Alagarsamy, M.; Rajasekaran, U.; Ganesan, S.; Palanivel, R. XAI based Photovoltaic Energy Management Framework for Smart Cities. *IEEE Access* **2025**, *13*, 98349–98361. [CrossRef]
46. Ghonge, M.M.; Pradeep, N.; Jhanjhi, N.Z.; Kulkarni, P.M. *Advances in Explainable AI Applications for Smart Cities*; IGI Global: Hershey, PA, USA, 2024.
47. Cartolano, A.; Cuzzocrea, A.; Pilato, G. Analyzing and assessing explainable AI models for smart agriculture environments. *Multimed. Tools Appl.* **2024**, *83*, 37225–37246. [CrossRef]
48. Anitha, A. Adaptation of XAI for Smart Agriculture Systems. In *Explainable, Interpretable, and Transparent AI Systems*; Tripathy, B.K., Seetha, H., Eds.; CRC Press: Boca Raton, FL, USA, 2024; pp. 72–88. [CrossRef]
49. Grati, R.; Fattouch, N.; Boukadi, K. Ontologies for Smart Agriculture: A Path Toward Explainable AI–A Systematic Literature Review. *IEEE Access* **2025**, *13*, 72883–72905. [CrossRef]
50. Schwalbe, G.; Finzel, B. A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts. *Data Min. Knowl. Discov.* **2023**, *37*, 2579–2606. [CrossRef]
51. Vilone, G.; Longo, L. Explainable Artificial Intelligence: A Comprehensive Review. *Artif. Intell. Rev.* **2021**, *55*, 3503–3541. [CrossRef]
52. Runck, B.C.; Schulz, B.; Bishop, J.; Carlson, N.; Chantigian, B.; Deters, G.; Erdmann, J.; Ewing, P.M.; Felzan, M.; Fu, X.; et al. Real-Time Geoinformation Systems to Improve the Quality, Scalability, and Cost of Internet of Things for Agri-Environment Research. In *Frontiers in Sustainable Food Systems*; Frontiers: Lausanne, Switzerland, 2024.
53. Islam, M.A.; Mridha, M.F.; Jahin, M.A.; Dey, N. A Unified Framework for Evaluating the Effectiveness and Enhancing the Transparency of Explainable AI Methods in Real-World Applications. *arXiv* **2024**, arXiv:2402.12890. [CrossRef]
54. Kondapaneni, N.; Marks, M.; Aodha, O.M.; Perona, P. Less is More: Discovering Concise Network Explanations. *arXiv* **2024**, arXiv:2402.11893. [CrossRef]
55. Quy, V.K.; Hau, N.V.; Anh, D.V.; Quy, N.M.; Ban, N.T.; Lanza, S.; Randazzo, G.; Muzirafuti, A. IoT-Enabled Smart Agriculture: Architecture, Applications, and Challenges. *Appl. Sci.* **2022**, *12*, 3396. [CrossRef]

56. Vermesan, O.; Friess, P. *Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems*; River Publishers: Aalborg, Denmark, 2013.
57. Gubbi, J.; Buyya, R.; Marusic, S.; Palaniswami, M. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* **2013**, *29*, 1645–1660. [CrossRef]
58. Frustaci, M.; Pace, P.; Aloi, G.; Fortino, G. Evaluating critical security issues of the IoT world: Present and future challenges. *IEEE Internet Things J.* **2017**, *5*, 2483–2495. [CrossRef]
59. Hassija, V.; Chamola, V.; Saxena, V.; Jain, D.; Goyal, P.; Sikdar, B. A survey on IoT security: Application areas, security threats, and solution architectures. *IEEE Access* **2019**, *7*, 82721–82743. [CrossRef]
60. Malhotra, P.; Singh, Y.; Anand, P.; Bangotra, D.K.; Singh, P.K.; Hong, W.C. Internet of things: Evolution, concerns and security challenges. *Sensors* **2021**, *21*, 1809. [CrossRef] [PubMed]
61. Kanellopoulos, D.; Sharma, V.K.; Panagiotakopoulos, T.; Kameas, A. Networking architectures and protocols for IoT applications in smart cities: Recent developments and perspectives. *Electronics* **2023**, *12*, 2490. [CrossRef]
62. Bellini, P.; Nesi, P.; Pantaleo, G. IoT-enabled smart cities: A review of concepts, frameworks and key technologies. *Appl. Sci.* **2022**, *12*, 1607. [CrossRef]
63. Zhang, N.; Demetriou, S.; Mi, X.; Diao, W.; Yuan, K.; Zong, P.; Qian, F.; Wang, X.; Chen, K.; Tian, Y.; et al. Understanding IoT Security Through the Data Crystal Ball: Where We Are Now and Where We Are Going to Be. *arXiv* **2017**, arXiv:1703.09809. [CrossRef]
64. Pimenow, S.; Pimenowa, O.; Prus, P.; Niklas, A. The Impact of Artificial Intelligence on the Sustainability of Regional Ecosystems: Current Challenges and Future Prospects. *Sustainability* **2025**, *17*, 4795. [CrossRef]
65. Al Jasem, M.S.; De Clark, T.; Shrestha, A.K. Toward decentralized intelligence: A systematic literature review of blockchain-enabled AI systems. *Information* **2025**, *16*, 765. [CrossRef]
66. Donta, P.K.; Murturi, I.; Casamayor Pujol, V.; Sedlak, B.; Dustdar, S. Exploring the potential of distributed computing continuum systems. *Computers* **2023**, *12*, 198. [CrossRef]
67. Allioui, H.; Mourdi, Y. Exploring the full potentials of IoT for better financial growth and stability: A comprehensive survey. *Sensors* **2023**, *23*, 8015. [CrossRef]
68. Demetriou, S. Privacy Concerns and Security Challenges in IoT Systems. *ACM Comput. Surv.* **2018**, *51*, 1–36.
69. Zakaie Far, A.; Zakaie Far, M.; Gharibzadeh, S.; Kazemi Naeini, H.; Amini, L.; Zangeneh, S.; Rahimi, M.; Asadi, S. Artificial Intelligence for Secured Information Systems in Smart Cities: Collaborative IoT Computing with Deep Reinforcement Learning and Blockchain. *arXiv* **2024**, arXiv:2409.16444. [CrossRef]
70. Methnani, L.; Dignum, V.; Theodorou, A. Clash of the Explainers: Argumentation for Context-Appropriate Explanations. *arXiv* **2023**, arXiv:2304.09142. [CrossRef]
71. Almadani, B.; Kaisar, H.; Thoker, I.R.; Aliyu, F. A systematic survey of distributed decision support systems in healthcare. *Systems* **2025**, *13*, 157. [CrossRef]
72. Kumar, I.A.; Wu, Y.; Jirotka, M.; Bussone, B. Explainable Artificial Intelligence: Foundations, Taxonomy and Challenges. *arXiv* **2020**, arXiv:2010.14973.
73. Gohel, P.; Singh, P.; Mohanty, M. Explainable AI: Current Status and Future Directions. *J. Big Data* **2021**, *8*, 48.
74. Heskes, T.; Sijben, E.; Bucur, I.G.; Claassen, T. Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models. *arXiv* **2020**, arXiv:2104.13398. [CrossRef]
75. Wolniak, R.; Stecuła, K. Artificial intelligence in smart cities—Applications, barriers, and future directions: A review. *Smart Cities* **2024**, *7*, 1346–1389. [CrossRef]
76. Thakker, D.; Mishra, B.K.; Abdullatif, A.; Mazumdar, S.; Simpson, S. Explainable artificial intelligence for developing smart cities solutions. *Smart Cities* **2020**, *3*, 1353–1382. [CrossRef]
77. Musa, A.A.; Malami, S.I.; Alanazi, F.; Ounaies, W.; Alshammari, M.; Haruna, S.I. Sustainable traffic management for smart cities using internet-of-things-oriented intelligent transportation systems (ITS): Challenges and recommendations. *Sustainability* **2023**, *15*, 9859. [CrossRef]
78. Mishra, P.; Singh, G. Energy management systems in sustainable smart cities based on the internet of energy: A technical review. *Energies* **2023**, *16*, 6903. [CrossRef]
79. Alahi, M.E.E.; Sukkuea, A.; Tina, F.W.; Nag, A.; Kurdthongmee, W.; Suwannarat, K.; Mukhopadhyay, S.C. Integration of IoT-enabled technologies and artificial intelligence (AI) for smart city scenario: Recent advancements and future trends. *Sensors* **2023**, *23*, 5206. [CrossRef]
80. Yeong, D.J.; Panduru, K.; Walsh, J. Exploring the unseen: A survey of multi-sensor fusion and the role of explainable ai (xai) in autonomous vehicles. *Sensors* **2025**, *25*, 856. [CrossRef] [PubMed]
81. Nastoska, A.; Jancheska, B.; Rizinski, M.; Trajanov, D. Evaluating trustworthiness in AI: Risks, metrics, and applications across industries. *Electronics* **2025**, *14*, 2717. [CrossRef]

82. Kabir, S.; Hossain, M.S.; Andersson, K. A review of explainable artificial intelligence from the perspectives of challenges and opportunities. *Algorithms* **2025**, *18*, 556. [CrossRef]
83. Wiratsin, I.o.; Ragkhitwetsagul, C. Effectiveness of Explainable Artificial Intelligence (XAI) Techniques for Improving Human Trust in Machine Learning Models: A Systematic Literature Review. *IEEE Access* **2025**, *13*, 121326–121350. [CrossRef]
84. Kalasampath, K.; Spoorthi, K.; Sajeev, S.; Kuppa, S.S.; Ajay, K.; Angulakshmi, M. A Literature review on applications of explainable artificial intelligence (XAI). *IEEE Access* **2025**, *13*, 41111–41140. [CrossRef]
85. Naveed, S.; Stevens, G.; Robin-Kern, D. An overview of the empirical evaluation of explainable ai (xai): A comprehensive guideline for user-centered evaluation in xai. *Appl. Sci.* **2024**, *14*, 11288. [CrossRef]
86. Shikonde, S.; Nkongolo, M.W. A Proactive Insider Threat Management Framework Using Explainable Machine Learning. *arXiv* **2025**, arXiv:2510.19883. [CrossRef]
87. Latif, R.M.A.; Ullah, F.; Jamal, N.; Zhao, Y.; Jabbar, S.; Khan, M.A. Explainable AI for Big Data Analytics in Urban Mobility Forecasting. In Proceedings of the 2025 IEEE International Conference on Pattern Recognition, Machine Vision and Artificial Intelligence (PRMVAI), Loudi, China, 20–22 June 2025; IEEE: New York, NY, USA, 2025; pp. 1–5.
88. Zhang, K.; Zheng, B.; Xue, J.; Zhou, Y. Explainable and trust-aware AI-driven network slicing framework for 6G IoT using deep learning. *IEEE Internet Things J.* **2025**.
89. Tao, W.; Tao, J.; Jiang, M. XAI Methods for Cross-Selling Prediction. *arXiv* **2024**, arXiv:2403.09876.
90. Li, P.; Mavromatis, I.; Khan, A. Past, Present, Future: A Comprehensive Exploration of AI Use Cases in the Umbrella IoT Testbed. *Sensors* **2024**, *24*, 1473.
91. Bilotta, S.; Ipsaro Palesi, L.; Nesi, P. Exploiting open data for CO_2 estimation via artificial intelligence and eXplainable AI. *Expert Syst. Appl.* **2025**, *291*, 128598. [CrossRef]
92. Chauncey, S.A.; McKenna, H.P. Creativity and innovation in civic spaces supported by cognitive flexibility when learning with AI chatbots in smart cities. *Urban Sci.* **2024**, *8*, 16. [CrossRef]
93. Piran, F.J.; Chen, Z.; Imani, M.; Imani, F. Privacy-Preserving Federated Learning with Differentially Private Hyperdimensional Computing. *Comput. Electr. Eng.* **2025**, *123*, 110261. [CrossRef]
94. Aminifar, A.; Shokri, M.; Aminifar, A. Privacy-Preserving Edge Federated Learning for Intelligent Mobile-Health Systems. *Future Gener. Comput. Syst.* **2024**, *161*, 625–637. [CrossRef]
95. Shajalal, M.; Boden, A.; Stevens, G.; Du, D.; Kern, D.-R. Explaining AI Decisions: Towards Achieving Human-Centered Explainability in Smart Home Environments. In *World Conference on Explainable Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 418–440.
96. Delfani, P.; Thuraga, V.; Banerjee, B.; Chawade, A. Integrative approaches in modern agriculture: IoT, ML and AI for disease forecasting amidst climate change. *Precis. Agric.* **2024**, *25*, 2589–2613. [CrossRef]
97. Choi, J.W.; Hidayat, M.S.; Cho, S.B.; Hwang, W.H.; Lee, H.; Cho, B.K.; Kim, M.S.; Baek, I.; Kim, G. Recent Trends in Machine Learning, Deep Learning, Ensemble Learning, and Explainable Artificial Intelligence Techniques for Evaluating Crop Yields Under Abnormal Climate Conditions. *Plants* **2025**, *14*, 2841. [CrossRef]
98. Mohan, R.J.; Rayanoothala, P.S.; Sree, R.P. Next-gen agriculture: Integrating AI and XAI for precision crop yield predictions. *Front. Plant Sci.* **2025**, *15*, 1451607. [CrossRef] [PubMed]
99. Fizza, K.; Jayaraman, P.P.; Banerjee, A.; Georgakopoulos, D.; Ranjan, R. Evaluating Sensor Data Quality in Internet of Things Smart Agriculture Applications. *IEEE Internet Things J.* **2021**, *8*, 4669–4682. [CrossRef]
100. Talaat, F.M.; Kabeel, A.; Shaban, W.M. Towards sustainable energy management: Leveraging explainable Artificial Intelligence for transparent and efficient decision-making. *Sustain. Energy Technol. Assess.* **2025**, *78*, 104348. [CrossRef]
101. Ngo, Q.H.; Kechadi, T.; Le-Khac, N.A. OAK4XAI: Model towards out-of-box eXplainable artificial intelligence for digital agriculture. In Proceedings of the International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK, 13–15 December 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 238–251.
102. Ashir, D.M.N.A.; Ahad, M.T.; Talukder, M.; Rahman, T. Internet of Things (IoT) Based Smart Agriculture Aiming to Achieve Sustainable Goals. *Sustainability* **2022**, *14*, 4438.
103. Kisten, M.; Ezugwu, A.E.S.; Olusanya, M.O. Explainable artificial intelligence model for predictive maintenance in smart agricultural facilities. *IEEE Access* **2024**, *12*, 24348–24367. [CrossRef]
104. Rogers, H.; Zebin, T.; Cielniak, G.; De La Iglesia, B.; Magri, B. Deep Learning for Precision Agriculture: Post-Spraying Evaluation and Deposition Estimation. *Comput. Electron. Agric.* **2024**, *207*, 107742.
105. Farooq, M.S.; Riaz, S.; Alvi, A. Web of Things and Trends in Agriculture: A Systematic Literature Review. *J. Sci. Food Agric.* **2023**, *104*, 2887–2904.
106. Kwon, Y.D.; Chauhan, J.; Mascolo, C. Yono: Modeling multiple heterogeneous neural networks on microcontrollers. In Proceedings of the 2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), Virtual, 4–6 May 2022; IEEE: New York, NY, USA, 2022; pp. 285–297.

107. Kadir, M.A.; Addluri, G.; Sonntag, D. Revealing Vulnerabilities of Neural Networks in Parameter Learning and Defense Against Explanation-Aware Backdoors. *arXiv* **2024**, arXiv:2401.08765. [CrossRef]
108. Vascotto, I.; Rodriguez, A.; Bonaita, A.; Bortolussi, L. When Can You Trust Your Explanations? A Robustness Analysis on Feature Importances. *Mach. Learn.* **2025**, *114*, 1156–1186.
109. Luss, R.; Dhurandhar, A. When Stability Meets Sufficiency: Informative Explanations That Do Not Overwhelm. *arXiv* **2024**, arXiv:2401.09876.
110. Khadivpour, F.; Banerjee, A.; Guzdial, M. Responsibility: An example-based explainable AI approach via training process inspection. *arXiv* **2022**, arXiv:2209.03433. [CrossRef]
111. Sokol, K.; Hepburn, A.; Santos-Rodriguez, R.; Flach, P. bLIMEy: Surrogate prediction explanations beyond LIME. *arXiv* **2019**, arXiv:1910.13016. [CrossRef]
112. Yang, Y.D.; Kwon, J.; Chauhan, C.; Mascolo, C. YONO: You Only Need One Task on Microcontrollers. *arXiv* **2024**, arXiv:2402.01990.
113. Das, A.; Rad, P. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv* **2020**, arXiv:2006.11371. [CrossRef]
114. Zhao, C.; Liu, J.; Parilina, E. ShapG: New Feature Importance Method Based on the Shapley Value. *J. Mach. Learn. Res.* **2025**, *26*, 1–28. [CrossRef]
115. Lu, C.; Zeng, J.; Xia, Y.; Cai, J.; Luo, S. Energy-Based Model for Accurate Estimation of Shapley Values in Feature Attribution. *arXiv* **2025**, arXiv:2501.12345. [CrossRef] [PubMed]
116. Dineen, J.; Kridel, D.; Dolk, D.; Castillo, D. Unified Explanations in Machine Learning Models: A Perturbation Approach. *arXiv* **2024**, arXiv:2401.09876. [CrossRef]
117. Liao, Q.V.; Varshney, K.R. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv* **2021**, arXiv:2110.10790.
118. Tjoa, E.; Guan, C. Quantifying Explainability of Saliency Methods in Deep Neural Networks with a Synthetic Dataset. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 5674–5684. [CrossRef]
119. Alzakari, S.A.; Aljebreen, M.; Ahmad, N.; Alahmari, S.; Alrusaini, O.; Alqazzaz, A.; Alkhiri, H.; Said, Y. Explainable artificial intelligence-based cyber resilience in internet of things networks using hybrid deep learning with improved chimp optimization algorithm. *Sci. Rep.* **2025**, *15*, 33160. [CrossRef]
120. Mohammad, A.A.S.; Mohammad, S.I.S.; Al Oraini, B.; Vasudevan, A.; Hindieh, A.; Altarawneh, A.; Alshurideh, M.T.; Ali, I. Strategies for applying interpretable and explainable AI in real world IoT applications. *Discov. Internet Things* **2025**, *5*, 71. [CrossRef]
121. Watson, D. Rational Shapley Values. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), Seoul, Republic of Korea, 21–24 June 2022; pp. 1083–1094. [CrossRef]
122. Sairam, S.; Srinivasan, S.; Marafioti, G.; Subathra, B.; Mathisen, G.; Bekiroglu, K. Explainable Incipient Fault Detection Systems for Photovoltaic Panels. *arXiv* **2020**, arXiv:2011.09843. [CrossRef]
123. Takahashi, D.; Shimizu, S.; Tanaka, T. Counterfactual Explanations of Black-box Machine Learning Models using Causal Discovery with Applications to Credit Rating. In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 30 June–5 July 2024; IEEE: New York, NY, USA, 2024; pp. 1–8.
124. AlShehri, Y.; Ramaswamy, L. SECOE: Alleviating Sensors Failure in Machine Learning-Coupled IoT Systems. *IEEE Internet Things J.* **2022**, *9*, 22378–22390.
125. Norval, C.; Cobbe, J.; Singh, J. Towards an Accountable Internet of Things: A Call for Reviewability. *Data Inf. Manag.* **2021**, *5*, 96–109.
126. Kaczmarek, E.; Miguel, O.X.; Bowie, A.C.; Ducharme, R.; Dingwall-Harvey, A.L.J.; Hawken, S.; Armour, C.M.; Walker, M.C.; Dick, K. CAManim: Animating End-to-End Network Activation Maps. *arXiv* **2023**, arXiv:2306.06966. [CrossRef]
127. Madumal, P.; Miller, T.; Sonenberg, L.; Vetere, F. A Grounded Interaction Protocol for Explainable Artificial Intelligence. *Auton. Agents Multi-Agent Syst.* **2019**, *33*, 239–267.
128. Steinecker, T.; Luettel, T.; Maehlisch, M. Collision Probability Distribution Estimation via Temporal Difference Learning. *arXiv* **2024**, arXiv:2401.12456. [CrossRef]
129. Roy, S.; Rezazadeh, F.; Chergui, H.; Verikoukis, C. Joint Explainability and Sensitivity-Aware Federated Deep Learning for Transparent 6G RAN Slicing. In Proceedings of the ICC 2023—IEEE International Conference on Communications, Rome, Italy, 28 May–1 June 2023.
130. Matarese, M.; Rea, F.; Sciutti, A. How Much Informative is Your XAI? A Decision-Making Assessment Task to Objectively Measure the Goodness of Explanations. *IEEE Robot. Autom. Lett.* **2023**, *8*, 2946–2953.
131. Luo, Z.; Zhao, S.; Lu, Z.; Sagduyu, Y.E.; Xu, J. Adversarial Machine Learning Based Partial-Model Attack in IoT. *arXiv* **2020**, arXiv:2010.12345. [CrossRef]
132. Nurse, J.R.C.; Creese, S.; De Roure, D. Security Risk Assessment in Internet of Things Systems. *IEEE Internet Things J.* **2018**, *5*, 779–789. [CrossRef]

133. Shaowang, T.; Liu, S.; Marques, J.; Feamster, N.; Krishnan, S. Algorithmic Data Minimization for Machine Learning over Internet-of-Things Data Streams. *Proc. VLDB Endow.* **2025**, *18*, 5652–5661. [CrossRef]
134. Wang, L.; Zhang, C.; Zhao, Q.; Zou, H.; Lasaulce, S.; Valenzise, G.; He, Z.; Debbah, M. Generative AI for RF Sensing in IoT Systems. *arXiv* **2024**, arXiv:2401.15432. [CrossRef]
135. Sayduzzaman, M.; Tamanna, J.T.; Kundu, D.; Rahman, T. Interoperability and Explicable AI-Based Zero-Day Attacks Detection Process in Smart Community. *IEEE Access* **2024**, *12*, 45678–45691.
136. Khan, M.A.; Farooq, M.S.; Saleem, M.; Shahzad, T.; Ahmad, M.; Abbas, S.; Abu-Mahfouz, A.M. Smart buildings: Federated learning-driven secure, transparent and smart energy management system using XAI. *Energy Rep.* **2025**, *13*, 2066–2081. [CrossRef]
137. Teixeira, B.; Carvalhais, L.; Pinto, T.; Vale, Z. Explainable AI framework for reliable and transparent automated energy management in buildings. *Energy Build.* **2025**, *347*, 116246. [CrossRef]
138. Naveen, P.; Vinodkumar, S. Enhancing power system management with XAI. In *Explainable Artificial Intelligence and Solar Energy Integration*; IGI Global: Hershey, PA, USA, 2025; pp. 393–416.
139. Alshkeili, H.M.H.A.; Almheiri, S.J.; Khan, M.A. Privacy-Preserving Interpretability: An Explainable Federated Learning Model for Predictive Maintenance in Sustainable Manufacturing and Industry 4.0. *AI* **2025**, *6*, 117. [CrossRef]
140. Tahir, H.A.; Alayed, W.; Hassan, W.U. A Federated Explainable AI Framework for Smart Agriculture: Enhancing Transparency, Efficiency, and Sustainability. *IEEE Access* **2025**, *13*, 97567–97584. [CrossRef]
141. Vultureanu-Albişi, A.; Bădică, C.; Ivanović, M. eXING-IoT conceptual framework for explainability integration in next generation-IoT. *Connect. Sci.* **2025**, *37*, 2507180. [CrossRef]
142. Watson, D.S. Rational Shapley Values. *arXiv* **2021**, arXiv:2106.10191.
143. Ali, J.; Khalid, A.S.; Yafi, E.; Musa, S.; Ahmed, W. Towards a Secure Behavior Modeling for IoT Networks Using Blockchain. *IEEE Internet Things J.* **2020**, *7*, 3932–3945.
144. Samaniego, M.; Deters, R. Digital Twins and Blockchain for IoT Management. *IEEE Internet Things J.* **2023**, *10*, 2867–2879.
145. Karras, A.; Karras, C.; Giotopoulos, K.C.; Tsolis, D.; Oikonomou, K.; Sioutas, S. Federated edge intelligence and edge caching mechanisms. *Information* **2023**, *14*, 414. [CrossRef]
146. Karras, A.; Giannaros, A.; Theodorakopoulos, L.; Krimpas, G.A.; Kalogeratos, G.; Karras, C.; Sioutas, S. FLIBD: A federated learning-based IoT big data management approach for privacy-preserving over Apache Spark with FATE. *Electronics* **2023**, *12*, 4633. [CrossRef]
147. Manh, B.D.; Nguyen, C.H.; Hoang, D.T.; Nguyen, D.N.; Zeng, M.; Pham, Q.V. Privacy-Preserving Cyberattack Detection in Blockchain-Based IoT Systems Using AI and Homomorphic Encryption. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 3456–3468.
148. Yan, K.; Cui, S.; Wuerkaixi, A.; Zhang, J.; Han, B.; Niu, G.; Sugiyama, M.; Zhang, C. Balancing Similarity and Complementarity for Federated Learning. *arXiv* **2024**, arXiv:2402.16980. [CrossRef]
149. Fu, A. Leveraging Learning Metrics for Improved Federated Learning. *arXiv* **2023**, arXiv:2307.12345. [CrossRef]
150. Suffian, M.; Khan, M.Y.; Bogliolo, A. Towards Human Cognition Level-Based Experiment Design for Counterfactual Explanations (XAI). *arXiv* **2022**, arXiv:2210.08765. [CrossRef]
151. Ferdowsi, A.; Saad, W. Generative Adversarial Networks for Distributed Intrusion Detection in the Internet of Things. *IEEE Trans. Ind. Inform.* **2020**, *16*, 6736–6746.
152. Baniecki, H.; Biecek, P. Adversarial Attacks and Defenses in Explainable Artificial Intelligence: A Survey. *arXiv* **2024**, arXiv:2403.09654. [CrossRef]
153. Zhang, Z.; Hamadi, H.A.; Damiani, E.; Yeun, C.Y.; Taher, F. Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access* **2022**, *10*, 98765–98784. [CrossRef]
154. Schlegel, U.; Keim, D.A. A deep dive into perturbations as evaluation technique for time series XAI. In Proceedings of the World Conference on Explainable Artificial Intelligence, Lisbon, Portugal, 26–28 July 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 165–180.
155. Haag, F.; Hopf, K.; Vasconcelos, P.M.; Staake, T. Augmented cross-selling through explainable AI–a case from energy retailing. *arXiv* **2022**, arXiv:2208.11404.
156. Hameed, I.; Sharpe, S.; Barcklow, D.; Au-Yeung, J.; Verma, S.; Huang, J.; Barr, B.; Bruss, C.B. BASED-XAI: Breaking ablation studies down for explainable artificial intelligence. *arXiv* **2022**, arXiv:2207.05566. [CrossRef]
157. Amiri, S.S.; Weber, R.O.; Goel, P.; Brooks, O.; Gandley, A.; Kitchell, B.; Zehm, A. Data representing ground-truth explanations to evaluate xai methods. *arXiv* **2020**, arXiv:2011.09892. [CrossRef]
158. Alufaisan, Y.; Marusich, L.R.; Bakdash, J.Z.; Zhou, Y.; Kantarcioglu, M. Does explainable artificial intelligence improve human decision-making? In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 19–21 May 2021; Volume 35, pp. 6618–6626.
159. Kakogeorgiou, I.; Karantzalos, K. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102520. [CrossRef]

160. Brankovic, A.; Cook, D.; Rahman, J.; Huang, W.; Khanna, S. Evaluation of popular XAI applied to clinical prediction models: Can they be trusted? *arXiv* **2023**, arXiv:2306.11985. [CrossRef]
161. Miao, J.; Rajasekhar, D.; Mishra, S.; Nayak, S.K.; Yadav, R. A Fog-Based Smart Agriculture System to Detect Animal Intrusion. *IEEE Trans. Consum. Electron.* **2023**, *69*, 342–351.
162. Xiong, S.; Sarwate, A.D.; Mandayam, N.B. Network Traffic Shaping for Enhancing Privacy in IoT Systems. *IEEE Trans. Inf. Theory* **2021**, *68*, 1832–1847.
163. Weber, L.; Lapuschkin, S.; Binder, A.; Samek, W. Beyond Explaining: Opportunities and Challenges of XAI-Based Model Improvement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 6234–6250. [CrossRef]
164. Shreim, H.; Gizzini, A.K.; Ghandour, A.J. Trainable Noise Model as an XAI Evaluation Method: Application on Sobol for Remote Sensing Image Segmentation. *Remote Sens.* **2023**, *15*, 2109.
165. Karras, A.; Karras, C.; Drakopoulos, G.; Tsolis, D.; Mylonas, P.; Sioutas, S. SAF: A peer to peer IoT LoRa system for smart supply chain in agriculture. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Crete, Greece, 17–20 June 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 41–50.
166. Karras, A.; Karras, C.; Giannaros, A.; Giotopoulos, K.C.; Tsolis, D.; Oikonomou, K.; Sioutas, S. TinyML-based Event Detection: An Edge-Cloud Approach for Smart Agriculture over LoRa WSNs. In Proceedings of the 2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Piraeus, Greece, 10–12 November 2023; IEEE: New York, NY, USA, 2023; pp. 1–10.
167. Misahlidou, V.; Karras, A.; Giannoukou, I.; Sioutas, S. Optimizing Data Transmission in LoRa-Based IoT Systems: A Performance Evaluation of Compression Algorithms. In Proceedings of the The International Conference on Innovations in Computing Research, London, UK, 25–27 August 2025; Springer: Berlin/Heidelberg, Germany, 2025; pp. 499–510.
168. Pamuklu, T.; Syed, A.; Kennedy, W.S.; Erol-Kantarci, M. Heterogeneous GNN-RL Based Task Offloading for UAV-Aided Smart Agriculture. *IEEE Trans. Mob. Comput.* **2023**, *23*, 3456–3468. [CrossRef]
169. Lindenschmitt, D.; Fischer, C.; Haussmann, S.; Kalter, M.; Kallfass, I.; Schotten, H.D. Agricultural On-Demand Networks for 6G Enabled by THz Communication. *IEEE Wirel. Commun.* **2024**, *31*, 78–85.
170. Muthuselvam, A.; Sowdeshwar, S.; Saravanan, M.; Perepu, S.K. Advanced Machine Learning Framework for Efficient Plant Disease Prediction. *Comput. Electron. Agric.* **2024**, *211*, 108065.
171. Zhu, N.; Liu, X.; Liu, Z.; Hu, K.; Wang, Y.; Tan, J.; Huang, M.; Zhu, Q.; Ji, X.; Jiang, Y.; et al. Deep Learning for Smart Agriculture: Concepts, Tools, Applications, and Opportunities. *Int. J. Agric. Biol. Eng.* **2018**, *11*, 32–44. [CrossRef]
172. Karras, A.; Karras, C.; Giannoukou, I.; Giotopoulos, K.C.; Tsolis, D.; Karydis, I.; Sioutas, S. Decentralized algorithms for efficient energy management over cloud-edge infrastructures. In Proceedings of the International Symposium on Algorithmic Aspects of Cloud Computing, Amsterdam, The Netherlands, 5 September 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 211–230.
173. Ding, X.; Wang, H.; Wang, C.; Li, Z.; Liang, Z. Exploring Data and Knowledge Combined Anomaly Explanation of Multivariate Industrial Data. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 5226–5238.
174. Islam, S.R.; Eberle, W.; Ghafoor, S.K.; Siraj, A.; Rogers, M. Domain Knowledge Aided Explainable Artificial Intelligence for Intrusion Detection and Response. *IEEE Trans. Dependable Secur. Comput.* **2020**, *19*, 3236–3249.
175. Abdallah, M.; Lee, W.J.; Raghunathan, N.; Mousoulis, C.; Sutherland, J.W.; Bagchi, S. Anomaly Detection Through Transfer Learning in Agriculture and Manufacturing IoT Systems. *J. Manuf. Syst.* **2021**, *60*, 382–394.
176. Ait Issad, H.; Aoudjit, R.; Rodrigues, J.J. A Comprehensive Review of Data Mining Techniques in Smart Agriculture. *Eng. Agric. Environ. Food* **2019**, *12*, 511–525. [CrossRef]
177. Aldhaheri, L.; Alshehhi, N.; Jameela, I.I.; Manzil, R.A.K.; Shumaila, J.; Saeed, N.; Alouini, M.S. LoRa Communication for Agriculture 4.0: Opportunities, Challenges, and Future Directions. *IEEE Internet Things Mag.* **2024**, *7*, 34–42. [CrossRef]
178. Ngo, Q.H.; Le-Khac, N.A.; Kechadi, T. Ontology Based Approach for Precision Agriculture. *Comput. Electron. Agric.* **2019**, *157*, 147–161.
179. Yu, G.; Liu, R.P.; Zhang, J.A.; Guo, Y.J. Tamperproof IoT with Blockchain. *IEEE Internet Things J.* **2022**, *9*, 6006–6017.
180. Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L.K.; Müller, K.-R. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Nature: Cham, Switzerland, 2019; Volume 11700.
181. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
182. Luss, R.; Dhurandhar, A. Towards Better Model Understanding with Path-Sufficient Explanations. *arXiv* **2021**, arXiv:2109.06181. [CrossRef]
183. Saied, M.; Guirguis, S. Explainable artificial intelligence for botnet detection in internet of things. *Sci. Rep.* **2025**, *15*, 7632. [CrossRef] [PubMed]
184. Dobrovolskis, A.; Kazanavičius, E.; Kižauskienė, L. Building XAI-based agents for IoT systems. *Appl. Sci.* **2023**, *13*, 4040. [CrossRef]

185. Roshan, K.; Zafar, A. Using Kernel SHAP XAI Method to Optimize the Network Anomaly Detection Model. *arXiv* **2023**, arXiv:2301.12345. [CrossRef]
186. Biessmann, F.; Treu, V. A Turing Test for Transparency. *arXiv* **2021**, arXiv:2104.09876. [CrossRef]
187. Ivanovic, M.; Autexier, S.; Kokkonidis, M. AI Approaches in Processing and Using Data in Personalized Medicine. *Front. Med.* **2022**, *9*, 826242.
188. Ferrario, A.; Loi, M. A Series of Unfortunate Counterfactual Events: The Role of Time in Counterfactual Explanations. *arXiv* **2021**, arXiv:2105.06789. [CrossRef]
189. Wu, Y.; Lin, G.; Ge, J. Knowledge-Powered Explainable Artificial Intelligence (XAI) for Network Automation Towards 6G. *IEEE Commun. Mag.* **2022**, *60*, 38–43.
190. Chuang, Y.N.; Wang, G.; Yang, F.; Liu, Z.; Cai, X.; Du, M.; Hu, X. Efficient XAI Techniques: A Taxonomic Survey. *arXiv* **2023**, arXiv:2301.04819. [CrossRef]
191. Velmurugan, M.; Ouyang, C.; Xu, Y.; Sindhgatta, R.; Wickramanayake, B.; Moreira, C. Developing Guidelines for Functionally-Grounded Evaluation of Explainable Artificial Intelligence Using Tabular Data. *arXiv* **2024**, arXiv:2401.14532. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.