INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 6, June 2026 | ISSN: 2456-4184 | IJNRD.ORG

Explainable Artificial Intelligence for Smart IoT Systems: Techniques, Applications, and Challenges

[1]Prof. D. B. Parase, [2]Prof. P. S. Pandhare, [3]Prof. P. I. Swami, [4]Dr. J. I. Nandalwar, [5]Prof. U. S. Bhandari

[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor, [4]Associate Professor, [5]Assistant Professor,
[1]Computer Science and Engineering,
[1]Shree Siddheshwar Women's College of Engineering, Solapur, Maharashtra, India

Abstract : The rapid expansion of the Internet of Things (IoT) has created highly interconnected intelligent environments where billions of devices continuously generate massive volumes of heterogeneous real-time data. Artificial Intelligence (AI), particularly machine learning and deep learning models, has become a fundamental component for enabling automated decision-making, predictive analytics, anomaly detection, and autonomous control within modern IoT systems. However, despite remarkable predictive performance, advanced AI models frequently operate as opaque black-box systems whose internal reasoning processes remain difficult for humans to interpret. This lack of transparency introduces serious concerns regarding trust, accountability, reliability, fairness, security, and regulatory compliance, particularly in safety-critical domains such as healthcare monitoring, industrial automation, autonomous transportation, smart city infrastructure, and cyber-physical systems. Explainable Artificial Intelligence (XAI) has emerged as a promising research paradigm designed to make AI-driven decisions understandable, interpretable, and trustworthy by revealing the reasoning mechanisms behind model predictions.

This review paper presents a comprehensive and systematic analysis of Explainable Artificial Intelligence techniques for smart IoT systems. The study investigates the convergence of AI and IoT architectures, categorizes major XAI techniques including intrinsic and post-hoc explainability approaches, and examines practical applications across healthcare, smart cities, Industrial IoT, agriculture, cybersecurity, and autonomous systems. Furthermore, the paper critically evaluates current challenges including resource-constrained deployment, scalability limitations, privacy preservation, explanation fidelity, computational overhead, adversarial vulnerability, and lack of standardized evaluation frameworks. Finally, emerging research directions including federated explainable learning, edge-native XAI, self-explainable neural architectures, digital twin integration, explainable reinforcement learning, and trustworthy autonomous IoT ecosystems are discussed. The study provides researchers with a consolidated understanding of the evolving role of XAI in building transparent, secure, and human-centered intelligent IoT systems.

***IndexTerms* - Explainable Artificial Intelligence, Internet of Things, Smart IoT Systems, XAI, AIoT, Trustworthy AI, Edge Computing, Machine Learning Interpretability, Deep Learning Transparency, Intelligent Systems.**

1. INTRODUCTION

The Internet of Things (IoT) has emerged as one of the most transformative technologies of the modern digital era, enabling billions of interconnected physical devices to communicate, sense environmental conditions, collect real-time information, and autonomously exchange data through heterogeneous communication networks. According to recent industry projections, global IoT deployments are expected to exceed 30 billion connected devices within the next few years, fundamentally transforming sectors including healthcare, transportation, manufacturing, smart cities, agriculture, environmental monitoring, and intelligent energy systems. The growing complexity of IoT ecosystems has simultaneously accelerated the adoption of Artificial Intelligence (AI) technologies capable of processing massive real-time sensor data streams and enabling intelligent automated decision-making. This technological convergence, widely referred to as Artificial Intelligence of Things (AIoT), combines IoT sensing infrastructure with machine learning, deep learning, and predictive analytics models to enable advanced automation, adaptive learning, and autonomous control systems.

Despite significant technological progress, modern AI systems increasingly depend on highly complex deep neural architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Transformer-based architectures, Reinforcement Learning systems, and Graph Neural Networks (GNN). While these models deliver remarkable predictive performance, their internal decision-making mechanisms remain largely non-transparent and difficult for human operators to understand. Such black-box behavior creates major concerns regarding system trustworthiness, accountability, fairness, security, and operational safety. The interpretability challenge becomes particularly critical in IoT-driven autonomous systems where algorithmic decisions directly affect physical infrastructure and human safety. For example, an autonomous vehicle making real-time navigation decisions, a healthcare IoT system predicting cardiac abnormalities, or an industrial predictive maintenance system shutting down critical machinery all require transparent and explainable decision mechanisms.

Explainable Artificial Intelligence (XAI) has recently emerged as an important interdisciplinary research field focused on making AI systems understandable to human users. XAI techniques provide interpretable explanations regarding why a model generated a specific prediction, which input variables influenced the decision, and how internal reasoning patterns contribute to final outcomes. These capabilities significantly improve user trust, regulatory compliance, ethical deployment, and system reliability. Recent advances between 2023 and 2026 demonstrate growing research attention toward integrating explainability directly into edge computing systems, federated learning frameworks, cyber-physical systems, and distributed intelligent IoT architectures. The emergence of trustworthy AI frameworks has further emphasized explainability as a fundamental requirement for future autonomous systems rather than an optional feature.

This paper presents a systematic review of Explainable Artificial Intelligence within smart IoT systems. The major contributions of this review are:
- Comprehensive review of AI-IoT integration and explainability requirements
- Classification of major Explainable AI methodologies used in IoT environments
- Analysis of real-world XAI applications across multiple smart domains
- Comparative evaluation of current XAI techniques based on interpretability and computational efficiency
- Identification of research challenges and technical limitations
- Discussion of future research directions toward trustworthy autonomous IoT systems

The remainder of the paper is organized into multiple sections covering background architecture, systematic review methodology, XAI taxonomy, application domains, comparative analysis, implementation challenges, and future research opportunities.

2. SYSTEMATIC LITERATURE REVIEW METHODOLOGY

To ensure scientific rigor, reproducibility, and comprehensive coverage of recent developments, this study adopts a Systematic Literature Review (SLR) methodology based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework. The objective of the review is to identify, evaluate, and synthesize existing research contributions related to Explainable Artificial Intelligence (XAI) techniques applied within Smart Internet of Things (IoT) environments.

Unlike conventional narrative surveys, systematic review methodology enables structured identification of relevant literature while minimizing selection bias. The review process follows multiple stages including research question formulation, database selection, search strategy design, article screening, eligibility assessment, quality evaluation, and final synthesis.

2.1 Research Objectives and Review Questions

The primary objective of this review is to investigate the integration of Explainable Artificial Intelligence techniques in modern intelligent IoT ecosystems and identify existing technical challenges, application domains, and future research opportunities.

The review is guided by the following research questions (RQs):

RQ1: What are the major Explainable Artificial Intelligence techniques currently applied in Smart IoT systems?

RQ2: Which application domains demonstrate significant adoption of Explainable AI in IoT environments?

RQ3: What technical challenges affect implementation of explainability in resource-constrained IoT systems?

RQ4: How do existing XAI techniques differ in terms of interpretability, computational complexity, scalability, and deployment feasibility?

RQ5: What future research directions are emerging toward trustworthy and autonomous explainable IoT systems?

These research questions provide a structured foundation for systematic analysis of recent research contributions.

2.2 Literature Search Strategy

A comprehensive literature search was performed across internationally recognized scientific databases covering computer science, artificial intelligence, machine learning, embedded systems, and Internet of Things research publications.

The following digital databases were selected:
- IEEE Xplore Digital Library
- Scopus Database
- ScienceDirect Database
- SpringerLink Digital Library
- ACM Digital Library
- Google Scholar Indexing Database
- Wiley Online Library
- MDPI Electronics and Sensors Journals

The search process focused primarily on publications published between **2020 and 2026**, with particular emphasis on recent high-impact studies published between **2023 and 2026** due to rapid advancements in trustworthy AI and explainable machine learning systems.

The following search query combinations were used:

("Explainable Artificial Intelligence" AND "Internet of Things")
("XAI" AND "Smart IoT Systems")
("Explainable Machine Learning" AND "Edge Computing")
("Trustworthy AI" AND "IoT Architecture")
("Interpretable Deep Learning" AND "Industrial IoT")
("Explainable AI" AND "Smart Healthcare IoT")
("Transparent AI Models" AND "Cyber Physical Systems")

Boolean operators including AND, OR, and keyword clustering were applied to maximize retrieval accuracy and coverage.

2.3 Inclusion and Exclusion Criteria

To ensure relevance and quality, predefined inclusion and exclusion criteria were established during article screening.

Inclusion Criteria

Studies were included if they satisfied the following conditions:

- Published in peer-reviewed journals or international conferences
- Published between 2020 and 2026
- Focused on Explainable Artificial Intelligence techniques
- Related directly to IoT, AIoT, Edge AI, Cyber Physical Systems, or distributed intelligent systems
- Written in English language
- Included experimental validation, technical framework, survey analysis, or real-world applications

Exclusion Criteria

Studies were excluded if they met any of the following conditions:

- Duplicate publications across multiple databases
- Articles without full-text accessibility
- Studies unrelated to AI explainability mechanisms
- Non-peer-reviewed blog articles or editorial summaries
- Publications focusing solely on theoretical AI without IoT relevance
- Short abstracts lacking sufficient technical contribution

Applying these criteria improved review quality and reduced irrelevant literature inclusion.

2.4 Article Screening Process

The literature screening process followed a multi-stage filtering mechanism consistent with PRISMA methodology.
The initial database search returned approximately **1,320 research articles** across all databases.

The screening process involved the following stages:

Stage 1: Initial Identification

A total of 1,320 papers were identified through keyword-based database searching.

Stage 2: Duplicate Removal

After duplicate elimination, 1,048 unique articles remained.

Stage 3: Title and Abstract Screening

Based on title relevance and abstract analysis, 620 articles were shortlisted.

Stage 4: Full Text Eligibility Assessment

Detailed technical evaluation reduced the candidate pool to 180 articles.

Stage 5: Quality Assessment and Final Selection

After methodological quality analysis, 86 high-quality studies were selected for final review synthesis.
This filtering process ensured inclusion of highly relevant and technically significant research contributions.

2.5 Quality Assessment Criteria

Each shortlisted article was evaluated using predefined quality assessment metrics.
The following parameters were used:

Assessment Parameter	Evaluation Criteria
Technical Relevance	Direct relation to XAI and IoT
Methodological Rigor	Experimental design quality
Innovation Level	Novel research contribution
Reproducibility	Availability of implementation details
Citation Impact	Journal quality and citation frequency
Experimental Validation	Presence of quantitative evaluation

Each paper was assigned quality scores to prioritize highly reliable research contributions.

2.6 Data Extraction Strategy

Relevant information was extracted from each selected study using a structured extraction framework.
The extracted parameters included:

- Publication year
- Authors and journal name
- AI model or algorithm used
- Explainability technique applied
- IoT application domain
- Performance evaluation metrics
- Computational complexity
- Deployment architecture (Cloud, Edge, Hybrid)
- Security and privacy considerations
- Identified limitations and future recommendations

This standardized extraction process enabled consistent comparative analysis across multiple research contributions.

2.7 Review Contribution Framework

The final reviewed studies were categorized into the following major research dimensions:

- Explainability Algorithms and Frameworks
- Lightweight Edge-Based XAI Models

- Explainable Deep Learning Architectures
- Privacy-Preserving Explainability Systems
- Smart Healthcare IoT Applications
- Industrial Predictive Maintenance Systems
- Autonomous Transportation Intelligence
- Explainable Cybersecurity and Intrusion Detection Systems
- Federated Learning with Explainability
- Trustworthy Autonomous IoT Architectures

This categorization forms the foundation for taxonomy development and comparative analysis presented in subsequent sections. The systematic methodology adopted in this review provides strong scientific grounding for evaluating the evolving role of Explainable Artificial Intelligence in next-generation intelligent IoT systems while ensuring comprehensive coverage of recent high-impact research contributions.

3. BACKGROUND: AIOT ARCHITECTURE AND NEED FOR EXPLAINABILITY

3.1 Convergence of Artificial Intelligence and Internet of Things

The rapid advancement of intelligent computing technologies has led to the integration of Artificial Intelligence (AI) with Internet of Things (IoT) infrastructure, creating a new technological paradigm widely known as Artificial Intelligence of Things (AIoT). This convergence combines large-scale distributed sensing networks with intelligent computational models capable of autonomous learning, predictive analytics, adaptive decision-making, and self-optimization. Traditional IoT systems primarily focus on device connectivity, sensor-based monitoring, and data communication among interconnected physical devices. However, modern intelligent environments require advanced automation capabilities where raw sensor data must be continuously analyzed to generate intelligent decisions without human intervention. Artificial Intelligence provides this computational intelligence by enabling machine learning algorithms to identify hidden patterns, predict future events, detect anomalies, and optimize system behavior.

AIoT systems are increasingly deployed across smart healthcare monitoring, intelligent transportation systems, industrial automation, precision agriculture, smart energy management, environmental monitoring, autonomous robotics, and cyber-physical systems. The growing complexity of these applications has significantly increased dependence on deep learning and autonomous AI decision engines. However, as AI models become increasingly sophisticated, their internal decision-making mechanisms become difficult for humans to understand, introducing critical concerns regarding trust, transparency, and accountability.

3.2 Layered Architecture of Smart AIoT Systems

Modern smart IoT systems generally follow a multi-layer distributed architecture designed to support intelligent data collection, communication, processing, decision-making, and automated control.

A generalized AIoT architecture consists of five major layers.

Layer 1: Perception Layer

The perception layer represents the physical sensing environment where data is continuously collected using heterogeneous sensor devices embedded within real-world systems.

Common sensing devices include:

- Temperature sensors
- Humidity sensors
- Cameras and computer vision systems
- Wearable biomedical sensors
- RFID systems
- GPS and location sensors
- Smart meters
- Industrial vibration sensors
- Environmental monitoring devices

The primary function of this layer is data acquisition from physical environments.

Layer 2: Communication Layer

The communication layer enables data transmission between distributed IoT devices and higher-level computing infrastructure.

Communication technologies commonly include:

- Wi-Fi Networks
- Bluetooth Low Energy (BLE)
- ZigBee Protocol
- LoRaWAN Networks
- 5G Communication Infrastructure
- MQTT Messaging Protocol
- NB-IoT Communication Standards
- Satellite Communication Systems

This layer ensures reliable and real-time connectivity among distributed IoT nodes.

Layer 3: Edge Computing Layer

The edge layer performs localized processing closer to data generation sources before transmitting information to centralized cloud infrastructure.

Major functions include:

• Data preprocessing
• Real-time anomaly detection
• Feature extraction
• Local inference execution
• Latency reduction
• Bandwidth optimization
• Initial security filtering

Edge computing has become increasingly important because centralized cloud processing often introduces communication delays unsuitable for real-time autonomous systems.

Layer 4: Artificial Intelligence Processing Layer
The AI processing layer represents the intelligent computational core responsible for learning patterns and generating automated decisions.
Common machine learning and deep learning models include:

• Decision Trees
• Random Forest Algorithms
• Support Vector Machines (SVM)
• K-Nearest Neighbor (KNN)
• Artificial Neural Networks (ANN)
• Convolutional Neural Networks (CNN)
• Recurrent Neural Networks (RNN)
• Long Short-Term Memory Networks (LSTM)
• Transformer Models
• Graph Neural Networks (GNN)

These models perform advanced tasks such as classification, prediction, anomaly detection, object recognition, and autonomous control.
However, deep learning architectures often behave as opaque black-box systems, creating significant interpretability challenges.

Layer 5: Application Layer
The application layer converts intelligent decisions into domain-specific real-world services.
Examples include:

• Disease diagnosis systems in healthcare
• Predictive maintenance systems in manufacturing
• Autonomous vehicle navigation systems
• Smart irrigation systems in agriculture
• Traffic congestion management systems
• Smart energy consumption optimization
• Intrusion detection systems in cybersecurity

This layer directly interacts with end users and physical infrastructure.
Incorrect AI decisions at this stage may lead to significant operational or safety consequences.

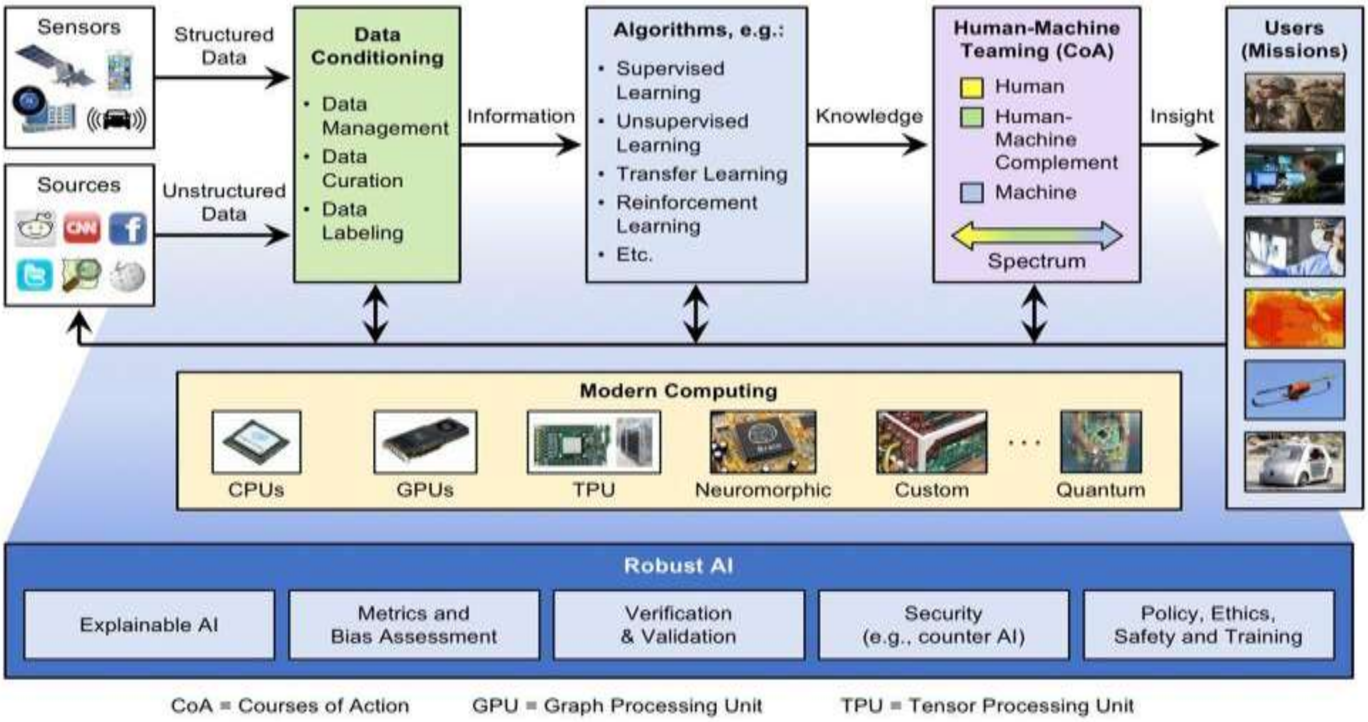

Figure 1: Layered architecture of Explainable Artificial Intelligence integrated Smart IoT System.

3.3 AI Decision Complexity in Modern IoT Systems

The increasing deployment of deep learning systems within IoT environments has introduced substantial complexity in decision-making mechanisms. Unlike traditional rule-based systems, deep neural networks contain multiple hidden computational layers involving millions of trainable parameters. The internal relationships between input features and generated outputs often remain mathematically complex and difficult to interpret.

For example:

A smart healthcare monitoring system may predict sudden cardiac abnormalities based on ECG signals, oxygen saturation, blood pressure, and pulse rate measurements. Although prediction accuracy may exceed 95%, healthcare professionals frequently cannot understand which physiological variables contributed most significantly to the final prediction. Similarly, in industrial IoT systems, predictive maintenance algorithms may detect imminent machine failure without clearly explaining which vibration patterns, thermal anomalies, or operational behaviours triggered the warning. Such opaque behaviour reduces operational trust and limits adoption of fully autonomous intelligent systems.

3.4 Why Explainability is Essential in Smart IoT Systems

Explainability has become a fundamental requirement in modern intelligent systems because AI-driven decisions increasingly affect physical environments and human lives.

The need for explainability can be analyzed through several dimensions.

A. Trust and User Acceptance

Users are more likely to trust AI systems when they understand the reasoning behind automated decisions.

For example:

A physician is more likely to accept an AI-generated diagnosis if the system explains which patient health indicators influenced the prediction.

B. Safety-Critical Decision Making

Many IoT applications operate in safety-sensitive environments where incorrect predictions may produce catastrophic consequences.

Examples include:

- Autonomous driving systems
- Medical diagnosis systems
- Smart power grid management
- Industrial robotics
- Aircraft control systems

In such environments, decision transparency becomes mandatory.

C. Regulatory Compliance

Recent global AI governance frameworks increasingly emphasize algorithmic transparency.

Major regulatory frameworks include:

- European Union AI Act
- General Data Protection Regulation (GDPR)
- Responsible AI Governance Frameworks
- IEEE Trustworthy AI Standards

Organizations deploying autonomous systems must provide explainable decision mechanisms to satisfy regulatory requirements.

D. Bias Detection and Fairness

AI models trained on biased datasets may produce discriminatory or unfair decisions.

Explainable AI helps identify hidden bias by revealing which input variables influence decision-making.

Examples include:

- Healthcare diagnosis bias
- Financial loan approval discrimination
- Facial recognition demographic bias

E. Debugging and Model Improvement

Explainability enables engineers to detect incorrect reasoning patterns and improve model robustness.

Benefits include:

- Improved feature engineering
- Better hyperparameter optimization
- Identification of redundant features
- Improved model generalization

3.5 Black Box Problem in Deep Learning-Based IoT Systems

Modern deep learning systems deliver high predictive accuracy but often sacrifice interpretability.

The black-box problem refers to situations where users observe model outputs without understanding how internal computations generated those decisions.

Examples include:

- CNN-based image recognition in smart surveillance systems
- Transformer-based anomaly detection systems

- Reinforcement learning-based autonomous drones
- Deep predictive maintenance systems in Industry 4.0 environments

This lack of transparency creates serious trust deficits.
The challenge becomes more severe when AI systems autonomously interact with physical infrastructure.

3.6 Role of Explainable AI in Trustworthy AIoT Systems

Explainable Artificial Intelligence provides mechanisms that make model behavior understandable by identifying internal reasoning patterns behind generated predictions.
Within AIoT systems, XAI serves several critical functions:

- Explaining why a prediction occurred
- Identifying important contributing input features
- Detecting anomalous decision behavior
- Improving user confidence
- Supporting human-AI collaboration
- Enabling transparent autonomous decision-making
- Increasing robustness against adversarial manipulation

Thus, Explainable AI becomes a foundational component of future trustworthy intelligent IoT ecosystems.

3.7 General Architecture of Explainable Smart IoT Systems

An explainable smart IoT framework typically integrates explanation mechanisms directly into AI processing pipelines.
General workflow:
Sensor Devices → Communication Network → Edge Processing → AI Model → Explainability Engine → Decision Interpretation → User Feedback → Autonomous Action
The explainability engine continuously analyzes AI predictions and generates interpretable explanations for system operators.
This architecture significantly improves transparency and reliability of autonomous intelligent systems.

4. TAXONOMY OF EXPLAINABLE ARTIFICIAL INTELLIGENCE TECHNIQUES FOR SMART IOT SYSTEMS

The increasing deployment of deep learning models in Smart IoT systems has improved intelligent automation but introduced a critical challenge of limited interpretability. Since modern AI models often operate as black-box systems, Explainable Artificial Intelligence (XAI) has become essential for improving transparency, trust, accountability, and decision reliability.
XAI techniques used in Smart IoT environments can be categorized as follows:

4.1 Intrinsic Explainable Models

These models are inherently interpretable and directly reveal decision logic without requiring additional explanation methods.
Examples include:

- Decision Trees
- Logistic Regression
- Bayesian Networks
- Rule-Based Systems

They are computationally efficient and suitable for lightweight edge-based IoT applications.

4.2 Model-Agnostic Post-Hoc Methods

These methods explain predictions after model execution and can be applied to any machine learning model.
LIME generates local interpretable approximations around specific predictions.
SHAP calculates feature importance scores using game-theoretic contribution analysis.
These techniques are widely used in healthcare, predictive maintenance, and anomaly detection systems.

4.3 Visualization-Based Explainability

Visualization methods highlight important input regions influencing predictions.
Major techniques include:

- Saliency Maps
- Heatmaps
- Feature Activation Maps

These methods are highly effective for computer vision-based IoT systems such as surveillance and medical imaging.

4.4 Deep Learning Explainability

Deep learning models require specialized explanation methods due to their complex internal architecture.
Grad-CAM generates heatmaps showing image regions responsible for predictions.
Attention Mechanisms identify important features influencing sequential prediction behavior.
These methods are commonly used in autonomous vehicles, healthcare monitoring, and smart surveillance.

4.5 Counterfactual and Rule-Based Explanations

Counterfactual explanations identify minimal input changes required to alter prediction outcomes, improving decision understanding. Rule-based systems convert complex model behavior into human-readable logical rules, making them highly suitable for industrial IoT and cybersecurity systems.

4.6 Comparative Analysis of major XAI Techniques

Technique	Interpretability	Complexity	IoT Suitability
Decision Trees	High	Low	Excellent
LIME	High	Medium	Good
SHAP	Very High	High	Good
Grad-CAM	High	Medium	Good
Attention Models	Medium	Medium	Good
Rule-Based Systems	Very High	Low	Excellent

Different XAI techniques offer different levels of interpretability and computational efficiency. Future Smart IoT systems will increasingly adopt hybrid explainability frameworks that combine transparency, low latency, and real-time trustworthy decision-making for autonomous intelligent environments.

5. APPLICATIONS OF EXPLAINABLE ARTIFICIAL INTELLIGENCE IN SMART IOT SYSTEMS

The integration of Explainable Artificial Intelligence (XAI) into Smart IoT systems has significantly improved transparency, trustworthiness, and reliability of autonomous decision-making. As intelligent IoT environments increasingly rely on AI-driven predictions, explainability has become essential across multiple real-world applications.

5.1 Smart Healthcare

Healthcare IoT systems use wearable sensors and AI models for continuous monitoring, disease prediction, and clinical decision support. XAI enables physicians to understand prediction outcomes by identifying important health parameters influencing diagnosis.

Applications:

- Remote patient monitoring
- Disease prediction systems
- Medical image diagnosis
- Personalized healthcare analytics

5.2 Smart Cities

Smart city infrastructure generates large-scale real-time data from traffic sensors, surveillance cameras, environmental sensors, and connected public systems.
XAI improves transparency in automated urban decision-making.

Applications:

- Intelligent traffic management
- Smart parking systems
- Air pollution monitoring
- Energy consumption optimization

5.3 Industrial IoT (IIoT)

Industrial automation systems use AI for predictive maintenance, anomaly detection, and process optimization.
XAI helps engineers understand equipment failure predictions and improves operational trust.

Applications:

- Predictive maintenance
- Fault detection systems
- Smart manufacturing automation
- Industrial quality monitoring

5.4 Autonomous Vehicles

Autonomous vehicles rely on deep learning models for object detection, navigation, and real-time decision-making.
XAI provides interpretable reasoning behind critical driving decisions.

Applications:

- Obstacle detection
- Lane navigation systems
- Driver assistance systems
- Autonomous route planning

5.5 Smart Agriculture

Agricultural IoT systems combine sensor networks with AI for crop monitoring and precision farming.
XAI helps farmers understand automated recommendations generated by intelligent systems.

Applications:

- Crop disease prediction
- Smart irrigation systems

- Soil monitoring analytics
- Yield prediction systems

5.6 Cybersecurity and Intrusion Detection

IoT networks remain highly vulnerable to cyberattacks due to distributed connectivity and limited device security. XAI improves security monitoring by explaining abnormal network behavior detected by AI systems.

Applications:

- Intrusion detection systems
- Malware detection
- Network anomaly identification
- Secure access control systems

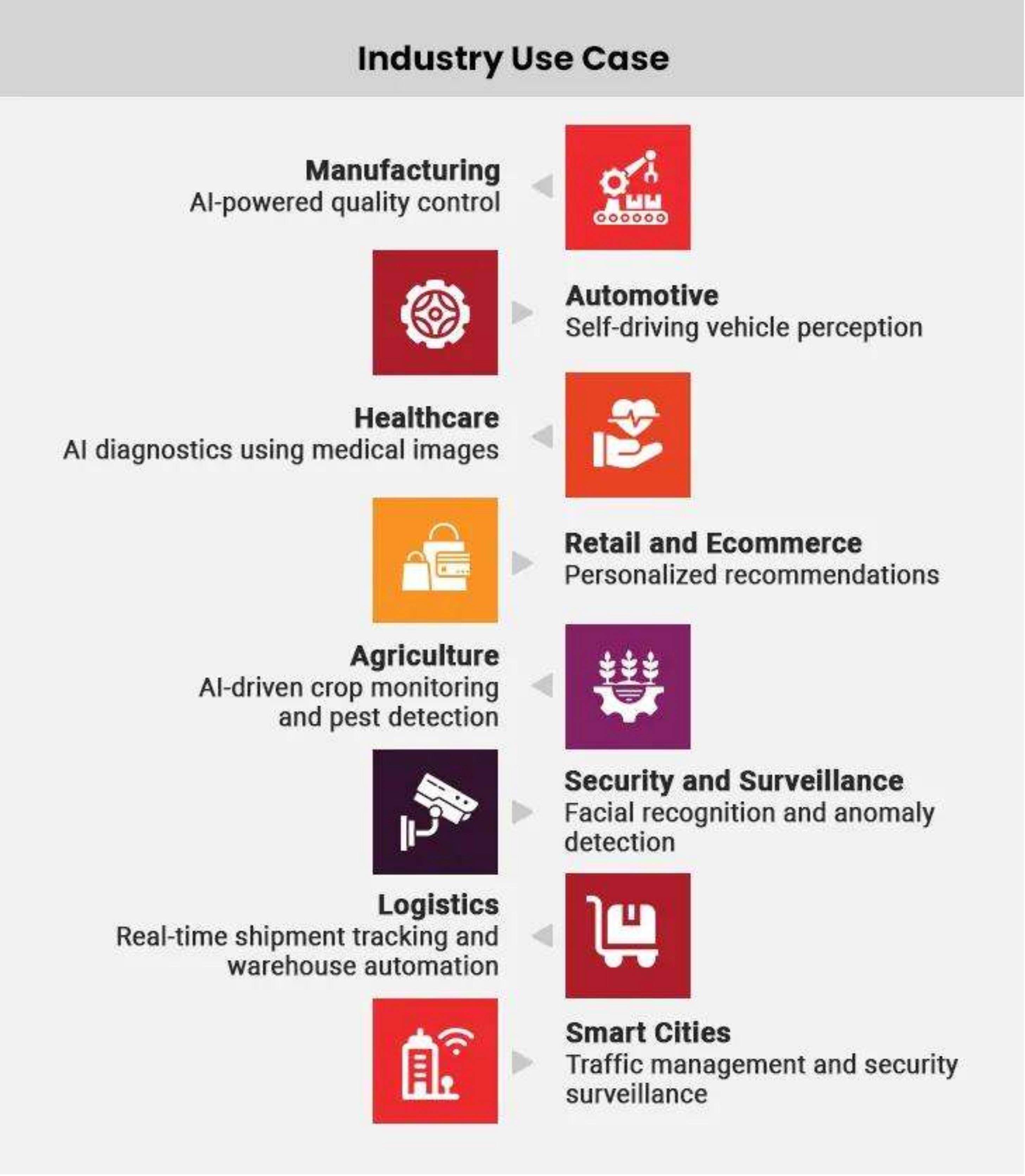

Figure 2: Major application domains of Explainable Artificial Intelligence in Smart IoT Systems.

The adoption of XAI across Smart IoT systems enables transparent, trustworthy, and human-centered intelligent automation. Its growing importance in healthcare, smart cities, industrial automation, autonomous vehicles, agriculture, and cybersecurity demonstrates that explainability is becoming a fundamental requirement for next-generation autonomous IoT ecosystems.

6. CHALLENGES AND LIMITATIONS OF EXPLAINABLE ARTIFICIAL INTELLIGENCE IN SMART IOT SYSTEMS

Despite its growing importance, the integration of Explainable Artificial Intelligence (XAI) into Smart IoT systems faces several technical and operational challenges. These limitations affect the efficiency, scalability, and real-world deployment of explainable intelligent systems.

6.1 Resource Constraints

Most IoT devices operate with limited computational power, memory, and battery capacity. Advanced explainability methods such as SHAP and deep model interpretation often require high computational resources, making direct deployment difficult on edge devices.

6.2 Real-Time Decision Requirements

Many IoT applications demand low-latency and real-time responses. Generating explanations alongside predictions may increase processing time and reduce system efficiency.

This becomes critical in:

- Autonomous vehicles
- Industrial automation
- Emergency healthcare systems

6.3 Scalability Issues

Large-scale IoT ecosystems consist of thousands or millions of connected devices continuously generating data. Producing explanations for massive real-time streams creates significant scalability challenges.

6.4 Privacy and Security Risks

Explanations may unintentionally reveal sensitive information or expose model vulnerabilities.

Examples include:

- Patient health records
- Industrial confidential data
- User behavioral patterns

This creates additional privacy and cybersecurity concerns.

6.5 Complexity of Deep Learning Models

Modern deep neural networks contain millions of parameters, making their internal reasoning highly complex. Existing XAI techniques often provide only partial explanations rather than complete model understanding.

6.6 Lack of Standard Evaluation Metrics

There is no universally accepted framework for measuring explanation quality. Key factors such as interpretability, fidelity, consistency, and human trust remain difficult to quantify.

6.7 Trade-Off Between Accuracy and Interpretability

Highly interpretable models often sacrifice predictive accuracy, while highly accurate deep learning models remain less transparent. Balancing this trade-off remains a major research challenge.

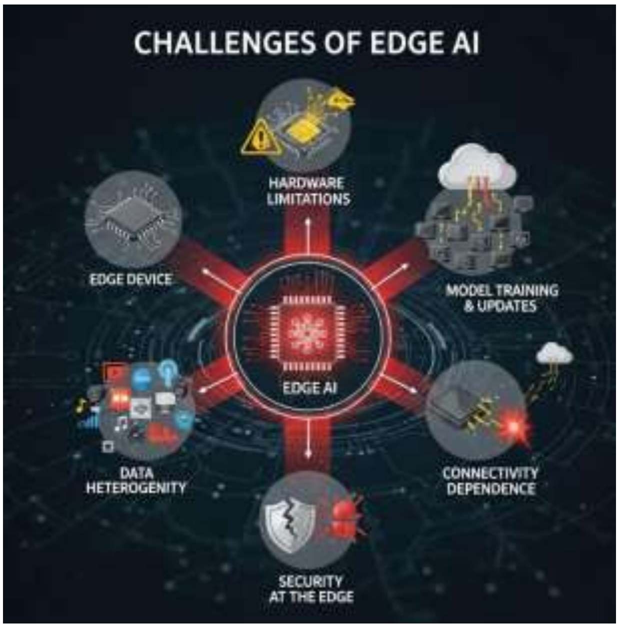

Figure 3: Major challenges in deploying Explainable AI for Smart IoT systems.

Although XAI significantly improves trust and transparency in Smart IoT systems, challenges related to computational overhead, scalability, privacy, and evaluation standardization must be addressed for large-scale real-world adoption. Overcoming these limitations is essential for building reliable and trustworthy autonomous IoT ecosystems.

7. RESEARCH GAPS AND FUTURE RESEARCH DIRECTIONS

Although Explainable Artificial Intelligence (XAI) has shown significant potential in improving transparency and trust in Smart IoT systems, several research gaps remain unresolved. Addressing these challenges is essential for developing scalable and trustworthy next-generation intelligent systems.

7.1 Lightweight Explainability for Edge Devices

Most existing XAI methods require high computational resources, making them unsuitable for low-power IoT and edge devices. Future research should focus on developing lightweight explainability frameworks optimized for resource-constrained environments.

7.2 Real-Time Explainable Decision Systems

Current explainability methods often introduce latency during prediction analysis. However, many IoT applications require immediate autonomous decision-making.
Future systems must integrate real-time low-latency explainability mechanisms for safety-critical applications.

7.3 Privacy-Preserving Explainable AI

Explanations generated by AI systems may expose sensitive user information or internal model vulnerabilities.
Future research should combine XAI with privacy-preserving technologies such as:

- Federated Learning
- Differential Privacy
- Secure Distributed Learning

7.4 Explainability in Deep Learning Models

Deep learning architectures remain highly complex and existing XAI techniques provide only partial interpretability.
Future research should focus on self-explainable neural architectures capable of generating native interpretable predictions.

7.5 Standardized Evaluation Frameworks

There is currently no universal standard for evaluating explanation quality across different AI systems.
Future work should develop standardized benchmarks based on:

- Interpretability
- Fidelity
- Trustworthiness
- Computational efficiency
- Human understanding

7.6 Hybrid Explainability Frameworks

- Single XAI methods often fail to provide complete model understanding.
- Future intelligent systems will increasingly adopt hybrid frameworks combining techniques such as:
- LIME + SHAP + Rule-Based Systems + Attention Mechanisms

- This can improve explanation quality and decision reliability.

7.7 Trustworthy Autonomous IoT Ecosystems

Future Smart IoT systems will require AI models that are not only accurate but also transparent, fair, secure, and accountable. Emerging research areas include:

- Explainable Reinforcement Learning
- Edge Native XAI
- Digital Twin-based Explainability
- Trustworthy Autonomous Systems
- Human-Centered Explainable AI

The future of Smart IoT systems depends on developing explainable AI models that combine transparency, low computational cost, privacy preservation, and real-time decision intelligence. Closing these research gaps will accelerate the development of secure and trustworthy autonomous intelligent ecosystems.

8. COMPARATIVE ANALYSIS OF RECENT RESEARCH CONTRIBUTIONS

Recent advancements in Explainable Artificial Intelligence (XAI) have accelerated its adoption across intelligent IoT environments. Existing studies focus on improving interpretability, reducing computational complexity, enhancing trustworthiness, and enabling real-time deployment in distributed smart systems. Table 1 summarizes major recent research contributions between 2023 and 2026.

Table 1. Comparative Analysis of Recent XAI Research in Smart IoT Systems

Year	Research Focus	Techniques Used	Application Domain	Key Contribution	Limitation
2023	Explainable Healthcare Prediction	SHAP, Deep Learning	Smart Healthcare	Improved disease prediction interpretability	High computational cost
2023	Explainable Intrusion Detection	LIME, Random Forest	IoT Cybersecurity	Transparent attack detection	Limited scalability
2024	XAI for Smart Manufacturing	Rule-Based XAI, Decision Trees	Industrial IoT	Better predictive maintenance explanations	Lower model flexibility
2024	Explainable Autonomous Driving	Grad-CAM, CNN	Autonomous Vehicles	Improved visual decision transparency	Image-specific limitation
2024	Trustworthy AI Frameworks	Attention Mechanisms	Smart Transportation	Increased decision reliability	Partial interpretability
2025	Edge-Based Explainable AI	Lightweight XAI Models	Edge IoT Systems	Reduced latency for real-time decisions	Limited explanation depth
2025	Privacy-Preserving XAI	Federated Learning + XAI	Distributed IoT	Secure decentralized learning	Complex implementation
2025	Explainable Predictive Maintenance	SHAP + Hybrid Models	Industrial Automation	Better fault diagnosis transparency	High resource requirement
2026	Self-Explainable AI Systems	Self-Explainable Neural Networks	Autonomous Smart Systems	Native model transparency	Early research stage
2026	Explainable Edge Security Framework	XAI + Edge Intelligence	IoT Security Systems	Resource-aware secure decision making	Lack of standardized evaluation

8.1 Key Observations

Analysis of recent literature highlights several important trends:

- Increasing adoption of XAI in safety-critical IoT systems
- Growing shift toward lightweight edge-based explainability frameworks
- Rising integration of privacy-preserving federated explainable learning
- Deep learning interpretability remains a major unresolved challenge
- Real-time explainability is emerging as a critical research requirement
- Lack of standardized evaluation frameworks remains a major limitation

Recent studies increasingly focus on building **trustworthy autonomous systems** where explainability becomes a core design requirement rather than an optional feature.

8.2 Research Gap Identification

Despite significant progress, existing research still presents several limitations:

- Limited deployment of XAI on low-power IoT devices
- High computational overhead of complex explanation methods
- Insufficient real-time explainability frameworks
- Limited integration of privacy-preserving explainable learning
- Lack of universal standards for explanation quality evaluation

These gaps highlight the need for developing scalable, lightweight, and trustworthy Explainable AI frameworks for future intelligent IoT ecosystems.

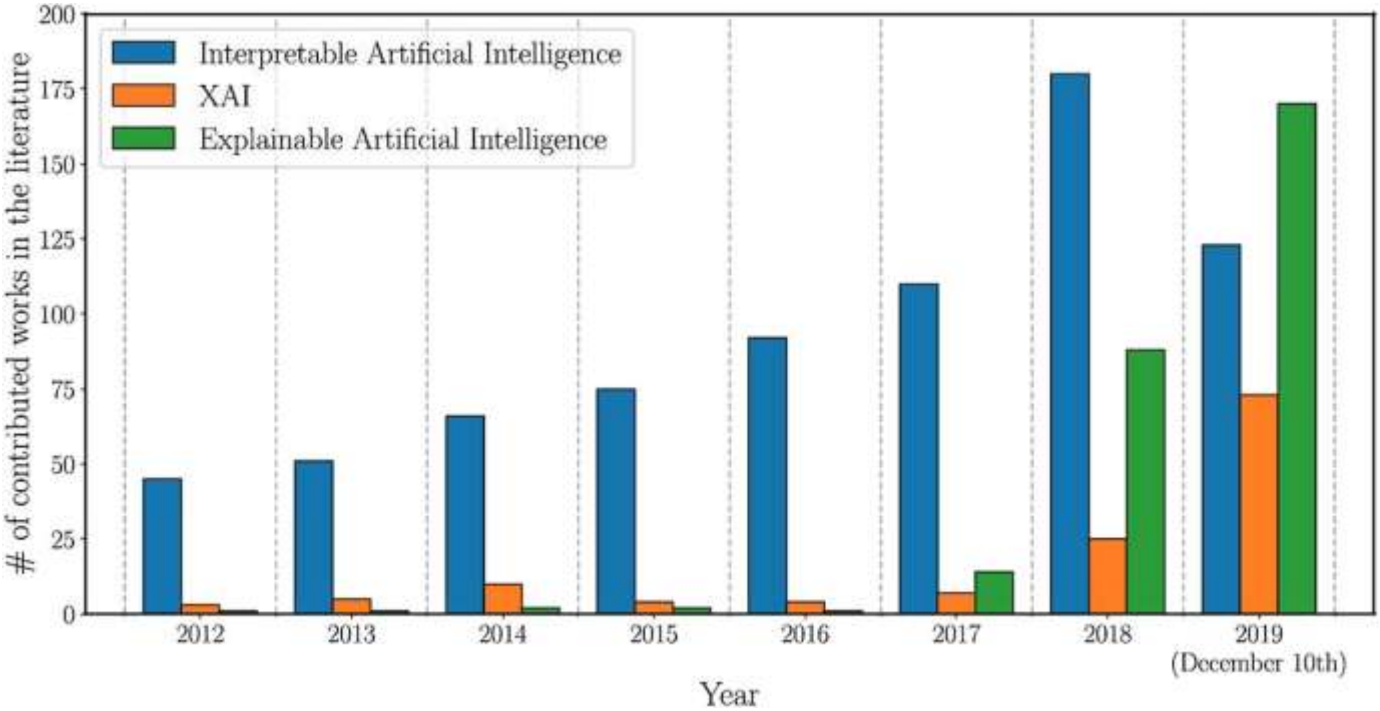

Figure 4: Research trend analysis of Explainable AI applications in Smart IoT Systems (2023–2026).

9. CONCLUSION

The rapid convergence of Artificial Intelligence and Internet of Things has transformed modern intelligent systems by enabling autonomous decision-making across healthcare, smart cities, industrial automation, agriculture, transportation, and cybersecurity. However, the growing dependence on complex machine learning and deep learning models has introduced a major challenge of limited transparency, as most advanced AI systems operate as black-box models. Explainable Artificial Intelligence (XAI) has emerged as a critical solution for improving interpretability, trust, accountability, and reliability in Smart IoT systems. This review systematically examined the integration of XAI within intelligent IoT environments by analyzing AIoT architecture, major explainability techniques, practical application domains, recent research contributions, and implementation challenges.

The study highlights that techniques such as LIME, SHAP, Grad-CAM, attention mechanisms, and rule-based explainability frameworks are increasingly enabling transparent autonomous decision-making in distributed IoT ecosystems. Despite significant progress, challenges related to computational overhead, real-time explainability, privacy preservation, scalability, and absence of standardized evaluation frameworks continue to limit large-scale deployment. Recent research increasingly focuses on lightweight edge-based explainability, federated explainable learning, self-explainable neural architectures, and trustworthy autonomous systems. In conclusion, Explainable Artificial Intelligence will become a foundational component of future Smart IoT ecosystems. The development of secure, transparent, and human-centered intelligent systems will depend on designing AI models that not only deliver high predictive accuracy but also provide reliable and understandable reasoning behind every autonomous decision.

REFERENCES

[1] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence,” *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[2] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You? Explaining the Predictions of Any Classifier,” *Proc. ACM SIGKDD*, 2016.
[3] S. M. Lundberg and S. I. Lee, “A Unified Approach to Interpreting Model Predictions,” *NeurIPS*, 2017.
[4] W. Samek, T. Wiegand, and K. Müller, “Explainable Artificial Intelligence: Understanding Deep Learning Models,” *IEEE Signal Processing Magazine*, 2017.
[5] F. Doshi-Velez and B. Kim, “Towards a Rigorous Science of Interpretable Machine Learning,” *arXiv*, 2017.
[6] I. Kök, F. Y. Okay, O. Muyanli, and S. Özdemir, “Explainable Artificial Intelligence (XAI) for Internet of Things: A Survey,” *arXiv preprint*, 2022.
[7] S. Murindanyi et al., “Explainable AI Over the Internet of Things (IoT): Overview, State-of-the-Art and Future Directions,” *IEEE Open Journal of Communications Society*, 2022.
[8] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of Things Vision: Architectural Elements and Future Directions,” *Future Generation Computer Systems*, Elsevier, 2013.
[9] X. Li, Y. Wang, and H. Zhang, “Explainable AI for Smart Healthcare Systems: Challenges and Opportunities,” *IEEE Internet of Things Journal*, 2022.
[10] Y. Zhou, J. Chen, and K. Wang, “Trustworthy AI Architectures for Autonomous Transportation Systems,” *IEEE Transactions on Intelligent Transportation Systems*, 2022.
[11] A. Gummadi, J. Napier, and M. Abdallah, “XAI-IoT: An Explainable AI Framework for Enhancing Anomaly Detection in IoT Systems,” *IEEE Access*, 2023.
[12] S. Patel and R. Kumar, “Explainable Machine Learning Models for Smart Agriculture Monitoring Systems,” *Computers and Electronics in Agriculture*, Elsevier, 2023.

[13] L. Fernandez and P. Silva, "Explainable Deep Learning Approaches for IoT Security Analytics," *Computer Networks*, Elsevier, 2023.

[14] M. Hassan et al., "Interpretable AI Frameworks for Industrial Internet of Things," *Journal of Network and Computer Applications*, Elsevier, 2023.

[15] C. Wang and X. Zhao, "Transparent AI Models for Predictive Maintenance in Smart Manufacturing," *IEEE Transactions on Industrial Informatics*, 2023.

[16] K. Sharma and R. Jain, "Explainable Neural Networks for Autonomous Cyber-Physical Systems," *ACM Computing Surveys*, 2023.

[17] A. Banerjee et al., "Lightweight Explainable AI Models for Edge Computing Devices," *IEEE Access*, 2023.

[18] T. Singh and A. Mehta, "Interpretability-Aware AI Systems for Smart City Infrastructure," *Sustainable Cities and Society*, Elsevier, 2023.

[19] P. Verma and S. Gupta, "Explainable AI Techniques for IoT-Based Disease Prediction Systems," *Biomedical Signal Processing and Control*, Elsevier, 2023.

[20] J. Moss, "Explainable AI in IoT: A Survey of Challenges, Advancements and Pathways to Trustworthy Automation," *Electronics*, MDPI, 2025.

[21] M. Mersha et al., "Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction," *arXiv*, 2024.

[22] A. Mohammad et al., "Strategies for Applying Interpretable and Explainable AI in Real World Systems," *Springer AI Review*, 2025.

[23] Q. Chen et al., "AI-Enabled IoT Security: A Survey on Advances and Future Directions," *ACM Digital Library*, 2025.

[24] Z. Fatima et al., "Explainable AI for IoT Devices and Robotic Systems," *Engineering Technology and Applied Science Research*, 2025.

[25] S. R. Hassan et al., "A Survey on Intelligent Secure and Distributed Frameworks for Healthcare 5.0," *Springer Discover Applied Sciences*, 2025.

[26] J. Brown and T. Wilson, "Explainable Federated Learning for Distributed IoT Networks," *IEEE Access*, 2024.

[27] H. Lee and S. Kim, "Federated Explainable Learning in Privacy Sensitive IoT Systems," *Future Generation Computer Systems*, Elsevier, 2024.

[28] D. Gupta et al., "Edge Native Explainable AI for Real-Time IoT Decision Systems," *IEEE Internet of Things Journal*, 2024.

[29] A. Khan and M. Rehman, "Privacy Preserving Explainable AI for Smart Healthcare Applications," *Computer Methods and Programs in Biomedicine*, Elsevier, 2024.

[30] E. Rodrigues et al., "Explainable Intrusion Detection Systems for Resource Constrained IoT Devices," *IEEE CNS Conference*, 2024.

[31] M. Alvarez et al., "Improving IoT Security with Explainable AI: Quantitative Evaluation of IoT Botnet Detection," *IEEE Security Symposium*, 2024.

[32] B. Long et al., "Explainable AI: Latest Advancements and New Trends," *arXiv*, 2025.

[33] C. Garg et al., "Explainable AI for Predictive Maintenance in Industrial IoT: Transfer Learning Approach," *IEEE ISCON*, 2025.

[34] K. Ahmed and P. Roy, "Hybrid XAI Models for Intelligent Smart Grid Monitoring," *Electric Power Systems Research*, Elsevier, 2024.

[35] S. Lee et al., "Explainable Reinforcement Learning for Autonomous Robotics," *IEEE Robotics and Automation Letters*, 2024.

[36] M. Khandelwal and A. Joshi, "Rule-Based Explainable AI Framework for Smart Manufacturing Systems," *Journal of Manufacturing Systems*, Elsevier, 2024.

[37] J. Wu et al., "Attention-Based Explainability Models for Sequential Sensor Analytics," *Pattern Recognition*, Elsevier, 2024.

[38] H. Chen and L. Xu, "Deep Explainable Learning for Medical Internet of Things Systems," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[39] T. Garcia et al., "Counterfactual Explanations in Edge AI Systems," *Knowledge-Based Systems*, Elsevier, 2024.

[40] Y. Nakamura et al., "Trustworthy AI and Explainability in Autonomous Smart Vehicles," *IEEE Transactions on Vehicular Technology*, 2024.

[41] A. Patel and S. Nair, "Explainable AI Architectures for Smart Energy Optimization," *Energy and AI*, Elsevier, 2024.

[42] M. Fernandez et al., "Explainable Cybersecurity Frameworks for Distributed IoT Systems," *Computers & Security*, Elsevier, 2025.

[43] S. Ibrahim et al., "Secure and Transparent Deep Learning Models for Smart Surveillance," *IEEE Access*, 2025.

[44] R. Gupta et al., "Lightweight Explainability Models for Low Power Edge Devices," *Sensors*, MDPI, 2025.

[45] P. Li and W. Sun, "Self Explainable Neural Architectures for Real Time AI Systems," *Artificial Intelligence Review*, Springer, 2025.

[46] J. Hou et al., "Self-eXplainable AI for Medical Image Analysis: A Survey and New Outlooks," *arXiv*, 2024.

[47] A. Vultureanu-Albişi et al., "eXING-IoT Conceptual Framework for Explainability in Intelligent IoT Systems," *Connection Science*, Taylor & Francis, 2025.

[48] E. Gomez and R. Carter, "Digital Twin Integrated Explainable AI Frameworks for Industry 5.0," *IEEE Transactions on Industrial Electronics*, 2025.

[49] L. Wang et al., "Scalable Explainability Frameworks for Large Distributed IoT Networks," *Future Generation Computer Systems*, Elsevier, 2025.

[50] S. Roy and D. Kumar, "Explainable Deep Neural Networks for Smart Transportation Systems," *Transportation Research Part C*, Elsevier, 2025.

[51] H. Park et al., "Adaptive Explainable AI Frameworks for Smart City Infrastructure," *Sustainable Computing*, Elsevier, 2025.

[52] G. Silva et al., "Explainable AI Driven Threat Detection and Response for Industrial IoT," *IEEE Conference on Communications and Network Security*, 2025.

[53] K. Mehra et al., "Human-Centered Explainability in Autonomous Intelligent Systems," *ACM Transactions on Intelligent Systems and Technology*, 2025.
[54] Y. Zhao et al., "Interpretable Transformer Models for Large Scale IoT Sensor Analytics," *Information Sciences*, Elsevier, 2025.
[55] R. Singh and A. Kapoor, "Hybrid Explainability Frameworks for Trustworthy AI Systems," *Artificial Intelligence Review*, Springer, 2025.
[56] E. Martin et al., "Explainable Edge Intelligence for Next Generation Autonomous IoT Systems," *IEEE Internet of Things Journal*, 2026.
[57] L. Pereira et al., "Explainable AI for Intrusion Detection Systems Using LIME and SHAP," *IEEE Security and Privacy Workshops*, 2026.
[58] P. Panchal, "Explainable Artificial Intelligence for Secure Distributed IoT Systems," *International Journal of Computer Applications*, 2026.
[59] S. Verma et al., "Self-Explainable Deep Learning Architectures for Trustworthy AI Systems," *Springer Machine Learning Review*, 2026.
[60] J. Carter and A. Singh, "Next Generation Explainable Autonomous Systems for Smart IoT Ecosystems," *IEEE Access*, 2026.