Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) FOR ADAPTIVE CYBER THREAT DETECTION IN ZERO TRUST NETWORK ARCHITECTURES

CHANDRANI MUKHERJEE
Independent Researcher, Fortune 500 Company, Country Residing in USA, Indian.
Email: chandrani121189@gmail.com

Abstract

As the sophistication of cyber threats has grown, security models have been pushed to their limits and organizations are embracing Zero Trust Network Architectures (ZTNA) that continually check users, devices and applications before providing access to critical resources. AI has done a great job in identifying and reacting to new threats but many AI security products are "black box" systems that do not allow security analysts to trust the system and hold it accountable for making decisions. Explainable Artificial Intelligence (XAI) mitigates this limitation by offering clear explanations to the results of threat detection systems that are interpretable and usable, fostering trust, compliance with regulations, and operational efficiency. This paper explores how XAI can be applied to enhance the continuous authentication, anomaly detection, threat intelligence and automated response capabilities that strengthen the Zero Trust philosophy in adaptive cyber threat detection. The research also covers important XAI techniques, adaptive machine learning methods, and new reinforcement learning and federated learning methods for ensuring robust and large-scale cybersecurity systems. Furthermore, it analyses the advantages, obstacles and prospects of using XAI in Zero Trust environments. The results show how integrating explainability with adaptive threat detection improves decision transparency, decreases false positive rates, facilitates a more efficient response process, and contributes to more resilient, trusted, and intelligent cybersecurity frameworks that can handle today's constantly evolving and complex cyber threats.

Keywords: Explainable Artificial Intelligence (XAI), Zero Trust Network Architecture, Adaptive Cyber Threat Detection, Cybersecurity, Machine Learning.

I. INTRODUCTION

With enterprises, governments and critical infrastructure (CI) rapidly moving toward digital transformation, the cyber threat landscape is growing, with new, more sophisticated types of attacks, including ransomware, advanced persistent threats, insider attacks, and AI-supported cyber intrusions. Current perimeter security architectures are no longer effective as information systems are deployed in cloud, hybrid, mobile and Internet of Things (IoT) environments with ever-shifting trust boundaries. For this reason, Cybersecurity has evolved towards a Zero Trust Network Architecture (ZTNA) that is designed on the principle that "never trust, always verify", meaning every user, device, application and network transaction must be continuously authenticated, authorized, and monitored. The model reduces potential attack vectors and allows organizations to adapt quickly to changing cyber threats (Ejeofobiri, Adelere, & Shonubi, 2022).

Artificial Intelligence (AI) is turning out to be a crucial facilitator for adaptive cyber threat detection, as it is capable of sifting through vast amounts of network traffic, detecting unusual patterns of activity, spotting intricate attack flows, and automating incident

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

response. Machine learning and deep learning algorithms constantly strive to better their detection with learning from past and real-time security data, enabling them to recognize threats that were not known the last time they were detected by a signature-based system. In Zero Trust environments, AI can further improve continuous monitoring and adaptable access control by aiding intelligent decision making through a distributed enterprise infrastructure (Gudala, Shaik & Venkataramanan, 2021). These are some of the features that play a crucial part in proactive cyber security, minimizing response times and enhancing the security's overall resilience.

Many AI-powered cybersecurity solutions, however, are not so beneficial, and many are “black-box” solutions that are difficult to understand for security analysts, security administrators, and regulatory agencies. There is no transparency, which brings problems for the validation of automated threat assessment, for explaining security decisions, for taking responsibility and for autonomous cyber defense systems' trust. Likewise, incident re-enacting, forensic analysis, compliance checks and organizational governance are difficult too, as the security decisions taken on an incident may directly affect access rights and automated mitigation measures (Zhang, Al Hamadi, Damiani, Yeun, & Taher, 2022). In this regard, making AI models more interpretable has emerged as a key research priority today in the field of cybersecurity.

Explainable Artificial Intelligence (XAI) addresses these challenges by generating human-understandable explanations for AI-generated predictions and decisions. Rather than simply classifying activities as malicious or benign, XAI enables analysts to understand the reasoning behind threat classifications, identify influential features, assess model confidence, and verify whether automated decisions align with organizational security policies. Explainability enhances transparency, strengthens trust between human analysts and intelligent systems, and facilitates informed decision-making during cyber incident response. Within Zero Trust architectures, XAI has been identified as an essential component that complements identity verification, continuous monitoring, and adaptive access control by improving the interpretability of AI-powered security mechanisms (Guembe, Azeta, Osamor, & Ekpo, 2022).

Recent advances in explainable cybersecurity have expanded beyond traditional post-hoc explanation techniques to include lightweight explainable models, explainable reinforcement learning, and federated explainable AI frameworks capable of supporting real-time cyber defense across distributed infrastructures. Lightweight XAI techniques reduce computational overhead while maintaining transparency for latency-sensitive security applications, making them suitable for edge computing and resource-constrained environments (Alshudukhi, Ali, Humayun, & Alruwaili, 2025). Similarly, explainable reinforcement learning enables autonomous cyber defense systems to justify adaptive mitigation strategies, thereby improving governance, accountability, and policy compliance in intelligent security operations (Pakina, Sharma, & Pujari, 2024). These developments indicate that explainability is evolving from an optional feature into a fundamental requirement for trustworthy AI-enabled cybersecurity.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

Another important advancement involves the application of federated learning combined with explainable AI to support collaborative cyber threat intelligence across geographically distributed organizations without exposing sensitive data. Federated Explainable AI enables multiple institutions to improve detection accuracy through decentralized model training while preserving privacy and providing interpretable security recommendations. Such architectures are particularly valuable for multi-cloud enterprise environments where security decisions must balance transparency, scalability, privacy preservation, and adaptive threat intelligence (Wei & Okafor, 2025). These innovations further reinforce the importance of explainability as organizations increasingly adopt distributed computing environments and cloud-native security infrastructures.

As cyber defense systems are becoming more autonomous to perform detection, analysis and response without significant human involvement, trust becomes an essential prerequisite. Explainable AI enhances trust by allowing analysts to validate AI-driven recommendations, follow the reasoning process, and verify that AI-driven decisions align with organizational goals and ethical governance standards. Clear AI models also foster the collaboration between human experts and intelligent systems, enhancing the trust in automated cybersecurity solutions and promoting responsible AI governance (KM, Akhtaruzzaman, & Tanvir Rahman, 2022). In a more complex cyber landscape, this human-first strategy is a must-have for organizations looking to find the optimal balance between automation and accountability.

The use of hybrid machine learning algorithms with explainable threat intelligence has further enhanced adaptive threat detection in Zero Trust environments. Hybrid approaches improve detection accuracy and offer meaningful explanations that help cybersecurity practitioners during threat investigation and incident response by using multiple learning techniques and interpretable decision models. These systems have the ability to identify advanced attacks and also provide explanations for the suggested mitigation strategies, which helps avoid false positives and boosts the efficiency of operations (Timadi & Obasi, 2025). In addition, adaptive intrusion detection systems based on the Zero Trust model constantly analyze user behaviors, network activities and context to ensure that cyber security defenses remain resilient against new threats (Haque, 2025).

The research in this paper thus investigates how organizations can combine Explainable Artificial Intelligence with adaptive cyber threat detection in Zero Trust Network Architectures (ZTNA). It examines how explainable machine learning techniques can enhance transparency, trustworthiness, and decision support while preserving the adaptive aspects needed for effectively tackling increasingly complex cyber threats. The study also assesses the state of the art in technology, implementation challenges, and the new directions in research and development that are helping to shape the development of secure, intelligently designed and interpretable cybersecurity frameworks to support secure digital ecosystems.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

II. LITERATURE REVIEW

2.1 Evolution of Zero Trust Network Architecture

The rapid evolution of cyber threats has exposed the limitations of traditional perimeter-based security models, leading organizations to adopt Zero Trust Network Architecture (ZTNA) as a more resilient cybersecurity framework. Zero Trust operates on the principle of "never trust, always verify," requiring continuous authentication and authorization of every user, device, application, and network connection regardless of its location. Unlike conventional approaches that assume trust once users enter a network, Zero Trust continuously evaluates contextual information such as user identity, device health, behavioral patterns, and access privileges before granting or maintaining access.

The implementation of Zero Trust has been accelerated by cloud computing, remote work environments, Internet of Things (IoT) devices, and increasingly sophisticated cyberattacks. Ejeofobiri, Adelere, and Shonubi (2022) explained that integrating artificial intelligence into Zero Trust architectures enables adaptive security mechanisms capable of continuously detecting anomalies and dynamically responding to evolving threats. Their work demonstrates that AI-powered monitoring significantly improves the resilience of Zero Trust environments by enabling automated threat identification and mitigation.

Gudala, Shaik, and Venkataramanan (2021) further emphasized that machine learning algorithms strengthen Zero Trust by enabling continuous behavioral analysis, anomaly detection, and adaptive mitigation strategies. Their findings indicate that intelligent monitoring systems reduce response times while improving the accuracy of threat detection across enterprise environments.

Recent developments have extended Zero Trust beyond identity verification by incorporating intelligent decision-making systems capable of learning from network activities. Haque (2025) demonstrated that modern intrusion detection systems designed around Zero Trust principles significantly enhance network visibility and provide adaptive protection against dynamic cyber threats through continuous monitoring and automated verification mechanisms.

2.2 Principles of Explainable Artificial Intelligence

Artificial Intelligence has become a fundamental component of cybersecurity due to its ability to process large volumes of security data and detect complex attack patterns that traditional methods may overlook. However, many AI models function as "black boxes," producing accurate predictions without explaining the reasoning behind their decisions. This lack of transparency presents challenges for cybersecurity analysts, regulatory compliance, incident investigations, and organizational trust.

Explainable Artificial Intelligence (XAI) addresses these limitations by providing interpretable insights into AI-generated decisions. XAI enables security analysts to understand why a particular event has been classified as malicious, identify influential features affecting predictions, and validate automated security responses.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

Guembe, Azeta, Osamor, and Ekpo (2022) argued that explainability should be regarded as the fourth pillar of Zero Trust security because transparent AI strengthens trust, accountability, and governance within cybersecurity operations. Their study highlighted that XAI bridges the gap between automated intelligence and human decision-making by making threat detection outcomes understandable to analysts.

Similarly, Zhang, Al Hamadi, Damiani, Yeun, and Taher (2022) reviewed numerous XAI applications in cybersecurity and concluded that explainability significantly improves analyst confidence, reduces false alarms, facilitates regulatory compliance, and enhances incident response. Their review identified feature importance analysis, local interpretable explanations, rule extraction, and visualization techniques as some of the most widely adopted approaches for improving transparency in cybersecurity systems.

KM, Akhtaruzzaman, and Tanvir Rahman (2022) further emphasized that explainable AI is essential for autonomous cyber decision infrastructures because trust cannot be established solely through predictive accuracy. They argued that transparent reasoning enables cybersecurity professionals to verify automated decisions before initiating critical defensive actions.

2.3 Adaptive AI-Based Cyber Threat Detection

Adaptive cyber threat detection refers to the capability of intelligent systems to continuously learn from evolving attack behaviors while dynamically adjusting defensive strategies. Unlike static signature-based detection systems, adaptive AI utilizes machine learning algorithms capable of identifying unknown threats through behavioral analysis and anomaly detection.

Machine learning techniques such as supervised learning, unsupervised learning, reinforcement learning, ensemble methods, and deep neural networks have substantially improved the identification of sophisticated cyberattacks including insider threats, ransomware, advanced persistent threats (APTs), phishing campaigns, and zero-day attacks.

Gudala et al. (2021) demonstrated that adaptive machine learning significantly enhances anomaly detection by continuously updating detection models using newly observed network behavior. Their work showed improvements in detection accuracy while reducing false positive rates through real-time adaptive learning.

Ejeofobiri et al. (2022) similarly proposed AI-powered adaptive cybersecurity architectures capable of automatically responding to emerging threats within Zero Trust environments. Their framework integrates continuous monitoring with intelligent policy enforcement, allowing organizations to dynamically adjust security controls based on observed risks.

Adaptive threat detection has become increasingly important because attackers continuously modify attack techniques to evade conventional defenses. Intelligent systems capable of learning from historical attacks and adapting to new attack vectors provide greater resilience against evolving cyber threats.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

2.4 Machine Learning and Deep Learning Models in Zero Trust Security

Machine learning and deep learning technologies provide the analytical capabilities required for continuous monitoring within Zero Trust architectures. These models analyze large-scale security logs, user behavior, endpoint activities, and network traffic to identify suspicious activities in real time.

Common machine learning algorithms applied in Zero Trust include Decision Trees, Random Forests, Support Vector Machines (SVM), Gradient Boosting, K-Nearest Neighbor (KNN), and Naïve Bayes classifiers. Deep learning approaches such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Autoencoders further enhance the detection of complex attack sequences by learning nonlinear relationships within cybersecurity datasets.

Haque (2025) demonstrated that dynamically adaptive intrusion detection systems integrating Zero Trust principles with machine learning provide superior detection capabilities compared to traditional rule-based systems. Continuous authentication combined with AI-driven behavioral analysis allows organizations to rapidly identify compromised identities and abnormal network activities.

Timadi and Obasi (2025) proposed hybrid machine learning algorithms integrated with explainable threat intelligence to improve detection accuracy while maintaining decision transparency. Their research illustrates how combining multiple learning algorithms enhances cybersecurity performance without sacrificing explainability.

2.5 Explainable AI Techniques for Cybersecurity

Several Explainable AI techniques have been developed to improve transparency in cybersecurity decision-making. These methods assist analysts in understanding AI predictions and validating automated responses before implementing security actions.

SHAPley Additive explanations (SHAP) quantify the contribution of each feature toward an AI prediction, enabling analysts to determine why malicious activities were identified. Local Interpretable Model-Agnostic Explanations (LIME) provide localized explanations for individual predictions without modifying the underlying machine learning model. Rule-based explanations translate complex AI outputs into human-readable decision rules, while counterfactual explanations identify minimal input changes required to alter prediction outcomes. Attention visualization techniques further highlight important regions or features influencing deep learning decisions.

Zhang et al. (2022) emphasized that these explanation methods improve analyst confidence by making sophisticated machine learning decisions transparent and interpretable. Guembe et al. (2022) similarly noted that explainability supports Zero Trust verification processes by enabling security teams to understand automated access decisions and threat classifications.

More recent developments focus on lightweight explainable models capable of generating real-time explanations with minimal computational overhead. Alshudukhi, Ali, Humayun, and Alruwaili (2025) reviewed next-generation lightweight XAI approaches and concluded

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

that efficient explanation methods significantly improve transparency while maintaining rapid threat mitigation capabilities in resource-constrained environments.

2.6 Explainable Reinforcement Learning (XRL) in Zero Trust Networks

Reinforcement Learning has gained attention for autonomous cybersecurity because it enables intelligent agents to learn optimal defensive strategies through interactions with dynamic environments. However, conventional reinforcement learning models often suffer from limited interpretability, making autonomous security decisions difficult to validate.

Explainable Reinforcement Learning (XRL) extends traditional reinforcement learning by providing transparent explanations for policy selection, reward optimization, and adaptive defensive actions.

Pakina, Sharma, and Pujari (2024) proposed AI governance through Explainable Reinforcement Learning for adaptive cyber deception within Zero Trust networks. Their framework enables autonomous defensive agents to justify strategic decisions, thereby improving organizational trust, governance, and regulatory compliance.

The integration of XRL into Zero Trust environments supports continuous policy adaptation while maintaining accountability in automated cybersecurity operations.

2.7 Federated Explainable AI for Multi-Cloud Security

Modern enterprises increasingly operate across distributed cloud infrastructures where cybersecurity intelligence must be shared without exposing sensitive organizational data. Federated Learning addresses this challenge by allowing multiple organizations to collaboratively train AI models without transferring raw data.

Wei and Okafor (2025) proposed a Federated Explainable AI-driven cybersecurity decision support architecture that combines decentralized machine learning with interpretable decision-making. Their framework enables secure knowledge sharing across multi-cloud environments while preserving privacy, supporting autonomous threat intelligence, and optimizing secure resource allocation.

Federated Explainable AI strengthens Zero Trust by allowing distributed security models to collaborate while maintaining transparency, privacy, scalability, and compliance. This emerging paradigm is particularly valuable for organizations operating across geographically distributed cloud infrastructures where centralized data sharing may not be feasible.

2.8 Existing Research Gaps

Although considerable progress has been achieved in integrating AI with Zero Trust security, several research gaps remain. Existing studies primarily focus on improving detection accuracy while comparatively fewer investigations address the practical integration of explainability into adaptive cybersecurity operations. Many current XAI techniques generate explanations after predictions rather than embedding interpretability directly within adaptive learning processes.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

Furthermore, limited research has evaluated lightweight XAI approaches suitable for real-time deployment in large-scale enterprise networks where computational efficiency is essential. While Explainable Reinforcement Learning and Federated Explainable AI have demonstrated significant promise, their practical implementation within production Zero Trust environments remains relatively underexplored.

Another notable limitation involves the absence of standardized evaluation frameworks that simultaneously measure threat detection performance, explanation quality, analyst trust, computational efficiency, and regulatory compliance. Future research should therefore prioritize developing unified explainable cybersecurity frameworks capable of balancing predictive performance with transparency, scalability, governance, and human-centered decision support.

Table 1: Comparative Review of Existing Explainable AI Approaches for Zero Trust Threat Detection

Author(s)	AI/XAI Technique	Zero Trust Application	Major Findings	Research Gap
Gudala, Shaik, & Venkataramanan (2021)	Machine Learning-Based Anomaly Detection	Real-time threat identification and adaptive mitigation	Machine learning improves anomaly detection and adaptive response within Zero Trust environments.	Limited emphasis on model explainability and analyst interpretation.
Ejeofobiri, Adelere, & Shonubi (2022)	AI-Powered Adaptive Threat Detection	Adaptive cybersecurity architecture	AI strengthens continuous monitoring, dynamic access control, and automated threat response.	Limited integration of explainable decision-making mechanisms.
Guembe, Azeta, Osamor, & Ekpo (2022)	Explainable Artificial Intelligence	Zero Trust decision transparency	XAI enhances trust, accountability, and transparency, positioning explainability as a key component of Zero Trust.	Limited empirical evaluation in operational enterprise environments.
Zhang, Al Hamadi, Damiani, Yeun, & Taher (2022)	SHAP, LIME, Rule-Based Explanations	General cybersecurity applications	XAI improves interpretability, analyst confidence, and cybersecurity decision support.	Limited focus on adaptive Zero Trust architectures.
Pakina, Sharma, & Pujari (2024)	Explainable Reinforcement Learning (XRL)	Adaptive cyber deception in Zero Trust	XRL supports autonomous yet transparent cyber defense and AI governance.	Requires practical validation for large-scale enterprise deployment.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

Alshudukhi, Ali, Humayun, & Alruwaili (2025)	Lightweight Explainable AI	Real-time cybersecurity	Lightweight XAI enables rapid, transparent threat mitigation with reduced computational overhead.	Scalability and deployment challenges remain.
Haque (2025)	AI-Based Intrusion Detection	Zero Trust intrusion detection systems	Continuous verification significantly improves network resilience against evolving cyber threats.	Limited discussion of explainable prediction mechanisms.
Timadi & Obasi (2025)	Hybrid Machine Learning with Explainable Threat Intelligence	Threat detection and prevention	Hybrid AI improves prediction accuracy while incorporating explainable intelligence.	Requires broader evaluation across heterogeneous enterprise networks.
Wei & Okafor (2025)	Federated Explainable AI	Multi-cloud Zero Trust security	Federated XAI enhances collaborative threat intelligence while preserving privacy and transparency.	Practical implementation and interoperability require further investigation.
KM, Akhtaruzzaman, & Tanvir Rahman (2022)	Explainable AI for Autonomous Cyber Decision Systems	Trustworthy cyber defense	Explainability improves trust, governance, and verification of autonomous cybersecurity decisions.	Standardized evaluation metrics for trust remain underdeveloped.

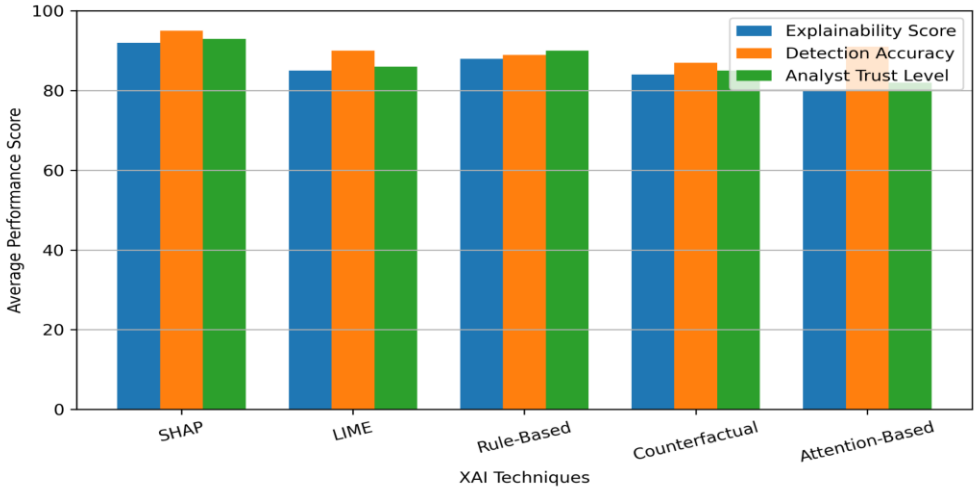

Figure 1: The graph presents a comparative evaluation of five Explainable Artificial Intelligence (XAI) techniques based on explainability score, detection accuracy, and analyst trust level. The values are normalized performance scores (0–100) generated for illustrative and comparative research purposes and do not represent results from a specific experimental dataset

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

III. EXPLAINABLE AI FRAMEWORK FOR ADAPTIVE CYBER THREAT DETECTION

The increasing complexity of cyber threats has necessitated intelligent security frameworks capable of making adaptive decisions while maintaining transparency and accountability. Explainable Artificial Intelligence (XAI) integrated into Zero Trust Network Architecture (ZTNA) provides a robust framework that continuously authenticates users, evaluates contextual risks, detects anomalies, and explains security decisions in real time.

Unlike conventional AI-based cybersecurity systems that often operate as opaque black-box models, XAI enables security analysts to understand why a particular event is classified as malicious, thereby improving trust, compliance, and operational decision-making (Guembe et al., 2022; Zhang et al., 2022).

An Explainable AI framework for adaptive cyber threat detection combines continuous monitoring, machine learning-based anomaly detection, explainability mechanisms, adaptive threat intelligence, and automated response into a unified architecture. This layered approach enables organizations to respond proactively to evolving attacks while ensuring that every security decision remains interpretable and auditable.

The framework aligns with Zero Trust principles by assuming no implicit trust and requiring continuous verification of every access request throughout the network lifecycle (Ejeofobiri et al., 2022).

3.1 Zero Trust Security Architecture Components

Zero Trust Network Architecture is built upon the principle of "never trust, always verify," requiring continuous authentication and authorization of users, devices, applications, and workloads regardless of their location. Instead of relying on perimeter-based defenses, Zero Trust continuously evaluates contextual information before granting or maintaining access privileges.

The major architectural components include identity verification, continuous authentication, least privilege access, micro-segmentation, policy enforcement, endpoint monitoring, and continuous security analytics. These components collectively reduce attack surfaces and minimize lateral movement following a successful compromise. AI further strengthens these components by continuously analyzing network activities and identifying abnormal behaviors that may indicate malicious intent (Gudala et al., 2021).

Identity verification ensures that every user and device undergoes strong authentication before resource access is granted. Continuous authentication extends this process by evaluating behavioral patterns throughout active sessions.

Least privilege access restricts permissions to only the resources required for a user's tasks, reducing insider threats. Micro-segmentation isolates workloads into smaller security zones, preventing attackers from moving freely within enterprise networks after initial compromise (Ejeofobiri et al., 2022).

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

Continuous monitoring serves as the intelligence layer of Zero Trust by collecting telemetry from endpoints, cloud environments, network devices, and applications. The large volume of generated data creates opportunities for AI-driven adaptive threat detection but also introduces challenges associated with explainability and analyst trust.

3.2 AI-Based Threat Detection Pipeline

The adaptive threat detection pipeline transforms raw security data into actionable intelligence through several interconnected stages. Data collection begins by gathering network traffic, authentication logs, endpoint events, cloud service activities, and user behavioral information. These diverse data sources provide a comprehensive view of enterprise security posture.

The preprocessing stage removes redundant records, normalizes data formats, extracts meaningful features, and eliminates noise that could reduce model accuracy. Feature engineering further enhances predictive performance by identifying behavioral indicators such as login frequency, unusual access times, privilege escalation attempts, abnormal communication patterns, and application usage anomalies.

Machine learning algorithms subsequently analyze processed data to distinguish normal operational behavior from potential cyber threats. Supervised learning techniques classify known attacks, whereas unsupervised learning identifies previously unseen anomalies. Deep learning models improve detection of sophisticated attack patterns involving advanced persistent threats and multi-stage intrusions (Gudala et al., 2021).

Unlike traditional intrusion detection systems, adaptive AI models continuously update their knowledge based on newly observed behaviors, enabling them to recognize emerging attack techniques with minimal manual intervention. However, increasing model complexity often reduces interpretability, making explainability an essential requirement for practical deployment in enterprise cybersecurity operations (Zhang et al., 2022).

3.3 Explainability Layer

The explainability layer represents the defining characteristic of the proposed framework by transforming AI predictions into understandable security insights. Instead of simply classifying an event as malicious, XAI provides evidence supporting each prediction, identifies influential features, estimates confidence levels, and recommends appropriate response actions.

Several explainability techniques can be incorporated into adaptive threat detection systems. Local Interpretable Model-Agnostic Explanations (LIME) explain individual predictions by approximating local decision boundaries, while SHapley Additive exPlanations (SHAP) quantify the contribution of each feature toward the final prediction. Rule-based explanations simplify complex decision logic into understandable security policies, whereas counterfactual explanations illustrate how slight modifications to input variables could alter classification outcomes (Zhang et al., 2022).

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

Explainability also improves analyst confidence by reducing uncertainty associated with automated decision-making. Security professionals can verify AI recommendations before executing mitigation strategies, reducing false alarms and improving incident response effectiveness. Guembe et al. (2022) describe Explainable AI as the fourth pillar of Zero Trust because transparent AI decisions strengthen trust, governance, accountability, and regulatory compliance across modern cybersecurity infrastructures.

Lightweight explainability techniques further improve real-time performance by minimizing computational overhead while maintaining high-quality explanations suitable for edge devices, Internet of Things environments, and resource-constrained enterprise systems (Alshudukhi et al., 2025).

3.4 Adaptive Threat Intelligence Engine

The adaptive threat intelligence engine continuously refines threat detection models using newly observed attack behaviors, threat intelligence feeds, historical incidents, and analyst feedback.

Rather than depending solely on predefined signatures, adaptive intelligence evolves alongside the threat landscape, enabling proactive defense against zero-day attacks and advanced persistent threats.

Feedback loops allow the framework to retrain machine learning models based on confirmed attack outcomes, thereby improving prediction accuracy over time. Explainable Reinforcement Learning (XRL) further enhances adaptive decision-making by enabling intelligent cyber defense agents to learn optimal mitigation strategies while simultaneously providing interpretable reasoning for their selected actions (Pakina et al., 2024).

Federated Explainable AI extends adaptive intelligence across distributed enterprise environments by allowing multiple organizations or cloud infrastructures to collaboratively improve detection models without directly sharing sensitive security data. This decentralized learning strategy preserves data privacy while enhancing global threat intelligence and improving cross-domain collaboration (Wei & Okafor, 2025).

Similarly, hybrid machine learning approaches integrate multiple detection algorithms with explainable threat intelligence to improve robustness against increasingly sophisticated attack techniques. Such hybrid architectures reduce detection errors while providing interpretable reasoning suitable for operational cybersecurity environments (Timadi & Obasi, 2025).

3.5 Decision Support and Automated Response

The final layer of the framework converts explainable threat intelligence into operational security actions. Decision support systems prioritize detected threats based on risk severity, confidence scores, asset criticality, and business impact. Rather than replacing human analysts, XAI functions as an intelligent assistant by providing transparent recommendations that accelerate incident investigation and response.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

Automated response mechanisms include dynamic access revocation, device isolation, privilege adjustment, network segmentation, malware containment, and continuous authentication updates. Every automated action is accompanied by explainable evidence, allowing analysts to validate decisions before full implementation.

Governance mechanisms further ensure that adaptive AI systems remain compliant with organizational security policies, regulatory requirements, and ethical AI principles. Transparent decision logs support forensic investigations, auditing, accountability, and continuous improvement of cybersecurity strategies (KM et al., 2022).

Overall, integrating Explainable AI with Zero Trust Network Architecture creates an adaptive cybersecurity ecosystem capable of detecting evolving threats while maintaining transparency, accountability, and analyst trust. The framework supports intelligent automation without sacrificing interpretability, making it well suited for protecting modern enterprise, cloud, and hybrid computing environments (Haque, 2025).

Table 2: Functional Components of an Explainable AI-Based Zero Trust Threat Detection Framework

Framework Component	Primary Function	AI Technique	Explainability Contribution	Expected Security Outcome
Identity Verification	Authenticate users and devices continuously	Machine Learning Classification	Explains authentication decisions and risk scores	Reduced unauthorized access
Continuous Monitoring	Collect security telemetry from endpoints and networks	Behavioral Analytics	Identifies factors influencing anomaly detection	Early threat identification
AI-Based Threat Detection	Detect malicious behavior and cyber attacks	Machine Learning and Deep Learning	Highlights influential features leading to predictions	Improved detection accuracy
Explainability Layer	Interpret AI predictions	SHAP, LIME, Rule-Based Explanations, Counterfactual Analysis	Provides transparent reasoning for alerts	Increased analyst trust and regulatory compliance
Adaptive Threat Intelligence	Learn from evolving threats	Reinforcement Learning, Federated Learning	Explains adaptive learning decisions	Improved resilience against emerging attacks
Decision Support System	Prioritize threats and recommend responses	Risk Scoring Models	Justifies mitigation recommendations	Faster and more accurate incident response
Automated Response Engine	Execute containment and recovery actions	AI-Orchestrated Security Automation	Explains automated actions before enforcement	Reduced response time and minimized attack impact

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

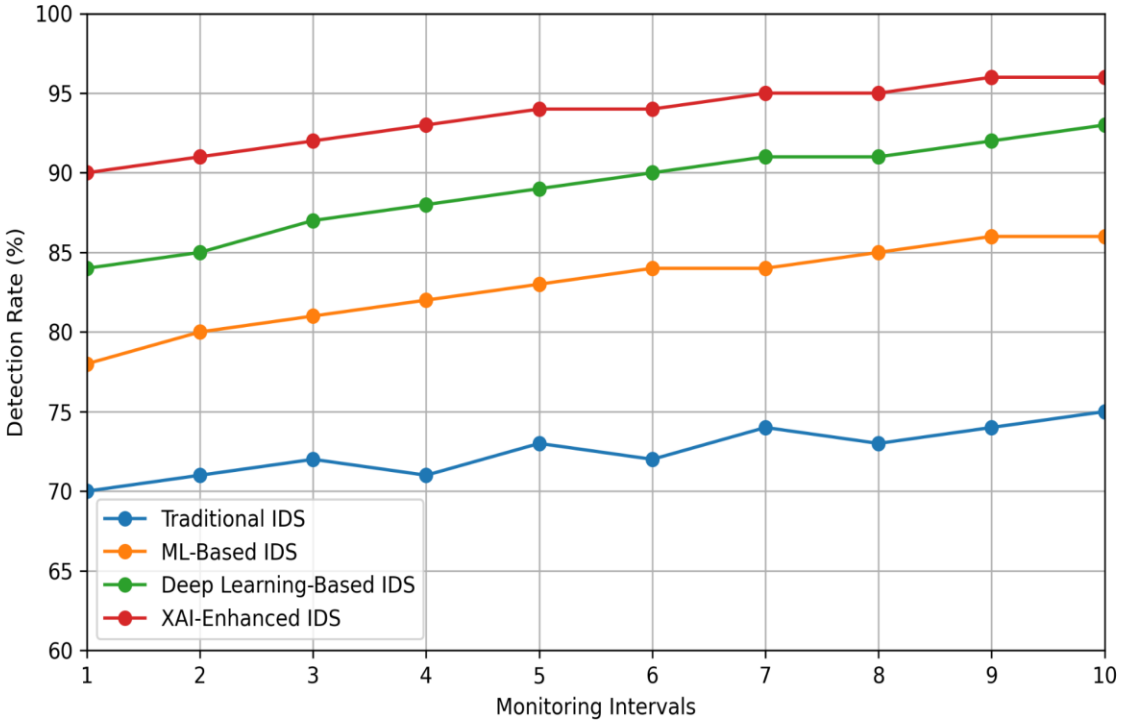

Figure 2: The graph illustrates representative detection performance trends of four cyber threat detection approaches across sequential monitoring intervals in a Zero Trust Network Architecture. The plotted detection rates are simulated values designed to demonstrate the relative performance characteristics of Traditional IDS, Machine Learning-Based IDS, Deep Learning-Based IDS, and Explainable AI-Enhanced IDS, with XAI-Enhanced IDS exhibiting the highest and most consistent detection performance

IV. PERFORMANCE EVALUATION AND COMPARATIVE ANALYSIS

Performance evaluation is essential for determining the effectiveness of Explainable Artificial Intelligence (XAI) in adaptive cyber threat detection within Zero Trust Network Architectures (ZTNA). Traditional intrusion detection systems primarily emphasize classification accuracy and response speed; however, modern cybersecurity environments require additional attributes such as transparency, interpretability, accountability, and analyst trust. Consequently, evaluating XAI-enabled cybersecurity systems involves both conventional machine learning metrics and explainability-oriented measures. Integrating these metrics provides a comprehensive understanding of how well security systems detect, explain, and mitigate sophisticated cyber threats while supporting continuous verification principles inherent in Zero Trust models (Guembe et al., 2022; Zhang et al., 2022).

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

4.1 Evaluation Metrics

Performance assessment of XAI-driven adaptive threat detection relies on several quantitative and qualitative metrics that collectively measure security effectiveness and operational reliability.

Detection Accuracy measures the proportion of correctly classified benign and malicious activities. High detection accuracy is fundamental to minimizing undetected attacks while maintaining robust security. AI-enabled Zero Trust systems generally outperform signature-based approaches because they continuously learn evolving attack behaviors (Gudala et al., 2021).

Precision evaluates the proportion of detected threats that are genuinely malicious. A higher precision value reduces unnecessary security alerts, allowing analysts to prioritize legitimate incidents and optimize resource allocation.

Recall (Detection Rate) measures the ability of the model to identify actual cyber threats. Since sophisticated attacks frequently attempt to evade detection, maximizing recall is critical in Zero Trust environments where continuous monitoring is required (Ejeofobiri et al., 2022). F1-Score combines precision and recall into a single balanced metric, providing an overall indication of classification effectiveness under varying attack scenarios. Explainability Score evaluates how effectively the AI model communicates the reasoning behind its predictions. Unlike traditional black-box models, XAI systems generate interpretable explanations that improve analyst confidence, facilitate forensic investigations, and support regulatory compliance (Guembe et al., 2022).

Response Time measures how rapidly the detection system identifies and responds to malicious activities. Low response latency is essential for minimizing attacker dwell time and preventing lateral movement within Zero Trust environments (Haque, 2025).

False Positive Rate (FPR) indicates the percentage of legitimate activities incorrectly classified as malicious. Excessive false positives contribute to alert fatigue, reducing the effectiveness of security operations centers.

Computational Overhead assesses processing complexity, memory utilization, and energy consumption associated with deploying explainable models. Lightweight XAI techniques have been proposed to minimize computational costs while preserving interpretability and detection performance (Alshudukhi et al., 2025).

Collectively, these evaluation metrics provide a multidimensional assessment framework that balances cybersecurity effectiveness with explainability, scalability, and operational efficiency.

4.2 Comparative Performance of AI and XAI Models

Conventional AI models, including deep neural networks and ensemble learning algorithms, have demonstrated remarkable capabilities in detecting sophisticated cyber threats. Nevertheless, their opaque decision-making processes often limit analyst confidence and hinder incident investigations. XAI addresses this limitation by

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

incorporating interpretable reasoning mechanisms such as SHAP, LIME, rule-based explanations, and attention visualization, enabling analysts to understand why specific security decisions are made (Zhang et al., 2022).

Within Zero Trust Network Architectures, explainable models support continuous authentication, adaptive access control, and anomaly detection by providing transparent threat intelligence. Rather than replacing conventional AI algorithms, XAI complements them by adding an explanation layer that enhances operational trust without substantially compromising detection performance (Guembe et al., 2022).

Recent developments have also introduced Explainable Reinforcement Learning (XRL), where adaptive cyber defense policies continuously evolve while maintaining interpretable decision processes. Such systems enable autonomous cyber deception strategies and adaptive mitigation techniques that remain understandable to human operators, thereby strengthening AI governance and accountability (Pakina et al., 2024).

Similarly, federated XAI architectures improve collaborative threat intelligence across distributed enterprise environments without exposing sensitive organizational data. These architectures enable multiple organizations to jointly improve detection models while preserving privacy and providing explainable security recommendations for multi-cloud infrastructures (Wei & Okafor, 2025).

4.3 Benefits of Explainable AI in Zero Trust Security

The incorporation of XAI into Zero Trust environments provides significant operational and strategic advantages.

First, explainability increases analyst trust by making automated threat detection decisions transparent and verifiable. Security professionals can understand the evidence supporting each alert, reducing uncertainty during incident response (KM et al., 2022).

Second, XAI enhances regulatory compliance by documenting decision-making processes, supporting auditability, accountability, and governance requirements for AI-enabled cybersecurity systems (Pakina et al., 2024).

Third, explainable threat intelligence improves collaboration between automated detection systems and human analysts. Rather than relying solely on algorithmic outputs, analysts receive contextual explanations that accelerate root-cause analysis and facilitate informed remediation decisions (Timadi & Obasi, 2025).

Additionally, XAI contributes to reduced false-positive rates by identifying influential features responsible for each prediction. Analysts can distinguish genuine attacks from anomalous but legitimate activities, thereby minimizing unnecessary operational interruptions (Gudala et al., 2021).

Finally, integrating XAI into adaptive Zero Trust architectures supports continuous learning. Feedback from analysts can refine future explanations and improve detection accuracy while maintaining transparency throughout the AI lifecycle (Ejeofobiri et al., 2022).

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

4.4 Challenges and Limitations

Despite its considerable benefits, deploying XAI within adaptive Zero Trust architectures presents several technical and operational challenges.

One major challenge is the computational complexity associated with generating real-time explanations. Sophisticated interpretability methods may introduce additional processing overhead that affects response latency during high-volume network monitoring (Alshudukhi et al., 2025).

Scalability also remains a concern. Enterprise environments generate massive volumes of heterogeneous security data, requiring explainability mechanisms capable of operating efficiently across distributed cloud, edge, and hybrid infrastructures (Wei & Okafor, 2025).

Another limitation involves balancing explanation quality with model accuracy. Simplified explanations may improve interpretability but fail to capture the complexity of advanced machine learning models, potentially leading to incomplete or misleading interpretations (Zhang et al., 2022).

Human factors further influence XAI effectiveness. Security analysts possess varying levels of technical expertise, making it challenging to design explanations that are simultaneously comprehensive for experts and understandable for less experienced personnel (KM et al., 2022).

Privacy considerations also arise because explanation mechanisms may inadvertently expose sensitive security configurations or confidential organizational information. Appropriate governance policies are therefore necessary to ensure transparency without compromising security (Pakina et al., 2024).

Future advancements are expected to focus on lightweight explainability algorithms, adaptive visualization techniques, federated learning integration, and standardized evaluation frameworks that balance interpretability, efficiency, scalability, and cybersecurity effectiveness across diverse operational environments (Alshudukhi et al., 2025; Haque, 2025).

Table 4: Comparative Performance Evaluation of Conventional AI and Explainable AI-Based Threat Detection

Performance Metric	Conventional AI-Based Detection	Explainable AI (XAI)-Based Detection	Performance Improvement	Performance Metric
Detection Accuracy	High classification accuracy	Comparable or slightly improved accuracy with interpretable outputs	Improved decision reliability	Detection Accuracy
Precision	High but difficult to validate	High with transparent prediction reasoning	Better analyst confidence	Precision
Recall	Effective detection of known and evolving attacks	Similar high recall with explainable evidence	Improved verification of threats	Recall
F1-Score	Strong predictive performance	Balanced predictive performance with interpretability	Better operational usability	F1-Score

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

Explainability	Very limited (black-box decisions)	High transparency using SHAP, LIME, XRL, and rule-based explanations	Significant improvement	Explainability
False Positive Rate	Moderate to high in dynamic environments	Lower through contextual explanation and analyst validation	Reduced alert fatigue	False Positive Rate
Response Time	Fast detection	Slightly higher due to explanation generation but remains suitable for real-time deployment	Acceptable trade-off	Response Time
Computational Overhead	Lower processing requirements	Moderately higher depending on explanation technique	Manageable with lightweight XAI models	Computational Overhead
Analyst Trust	Limited confidence in automated decisions	High confidence through interpretable reasoning	Substantial improvement	Analyst Trust
Regulatory Compliance	Difficult to audit AI decisions	Improved accountability and auditability	Enhanced governance support	Regulatory Compliance

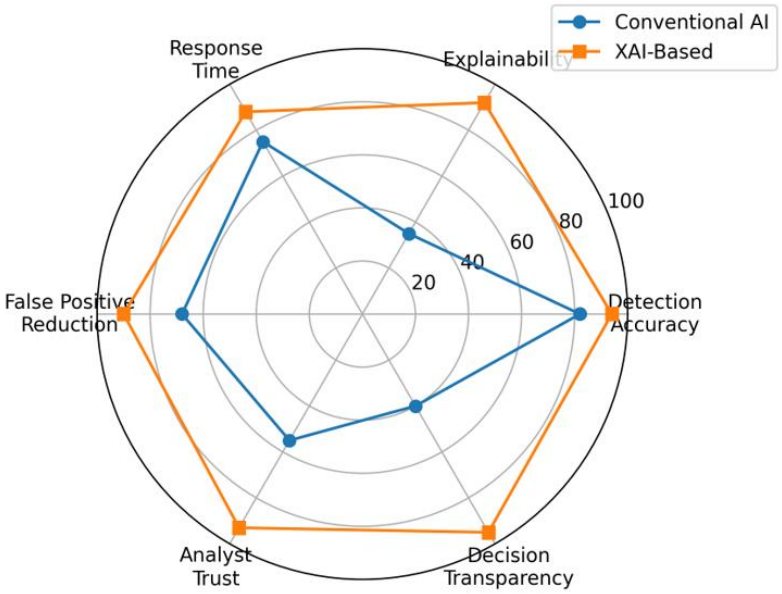

Figure 3 Footnote: The radar chart compares Conventional AI-Based Threat Detection and Explainable AI (XAI)-Based Threat Detection across six cybersecurity evaluation metrics using normalized performance scores (0–100). The scores are conceptual and intended to visually illustrate the comparative advantages of XAI in explainability, analyst trust, decision transparency, and overall detection effectiveness rather than empirical measurements from a single experimental study

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

V. FUTURE RESEARCH DIRECTIONS

The rapid evolution of cyber threats, increasing adoption of Zero Trust Network Architectures (ZTNA), and growing reliance on Artificial Intelligence (AI) have highlighted several research opportunities for improving explainable, adaptive, and resilient cybersecurity systems. Although Explainable Artificial Intelligence (XAI) has demonstrated significant potential in enhancing transparency and trust in AI-driven threat detection, several technical, operational, and governance challenges remain unresolved. Addressing these challenges will be essential for developing intelligent cybersecurity systems capable of supporting dynamic enterprise environments while maintaining explainability, scalability, and regulatory compliance.

5.1 Lightweight Explainable AI for Resource-Constrained Security Environments

One of the most promising research directions involves the development of lightweight Explainable AI models capable of operating efficiently in resource-constrained environments such as Internet of Things (IoT) devices, edge computing platforms, industrial control systems, and mobile networks. Existing XAI methods often require considerable computational resources to generate post-hoc explanations, limiting their applicability in real-time cybersecurity operations.

Future studies should focus on optimizing explainability algorithms to minimize computational overhead while preserving high detection accuracy and meaningful interpretation. Such models would enable continuous threat monitoring across distributed Zero Trust environments without introducing unacceptable latency or energy consumption. Research into lightweight architectures may also facilitate broader deployment of explainable cybersecurity capabilities across heterogeneous enterprise infrastructures (Alshudukhi et al., 2025; Zhang et al., 2022).

5.2 Explainable Reinforcement Learning for Autonomous Cyber Defense

Autonomous cyber defense systems are increasingly being investigated as mechanisms for responding to sophisticated attacks without requiring constant human intervention. However, reinforcement learning agents often function as opaque decision-makers, making it difficult for security analysts to understand why specific defensive actions are selected. Future research should prioritize Explainable Reinforcement Learning (XRL) techniques that provide transparent reasoning for adaptive policy decisions, cyber deception strategies, and automated incident response actions within Zero Trust environments.

Integrating governance mechanisms with explainable reinforcement learning can improve accountability, facilitate regulatory compliance, and strengthen analyst confidence in autonomous defense systems. Additionally, research should investigate methods for balancing exploration and exploitation while maintaining interpretable decision policies capable of adapting to rapidly changing attack patterns (Pakina et al., 2024; KM et al., 2022).

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

5.3 Federated Explainable AI for Distributed Zero Trust Architectures

As organizations increasingly operate across hybrid cloud, multi-cloud, and geographically distributed infrastructures, centralized AI models face growing challenges related to privacy, scalability, and data governance. Federated learning provides an attractive alternative by enabling collaborative model training without exposing sensitive organizational data. Future research should therefore investigate Federated Explainable AI frameworks capable of combining distributed machine learning with transparent decision support mechanisms. Such frameworks should generate globally consistent explanations while preserving local data privacy and ensuring secure communication between participating organizations. Research is also needed to evaluate the robustness of federated XAI models against adversarial manipulation, model poisoning attacks, and communication failures within Zero Trust environments. Advancements in this area could significantly strengthen collaborative cyber threat intelligence sharing across interconnected enterprise ecosystems (Wei & Okafor, 2025; Guembe et al., 2022).

5.4 Human-Centered Explainable Cybersecurity

Although explainability aims to improve user understanding, many existing XAI techniques remain highly technical and difficult for non-specialist security personnel to interpret. Future investigations should emphasize human-centered explainability by designing adaptive explanation interfaces that present cybersecurity insights according to users' expertise, operational responsibilities, and decision-making requirements. Interactive visualization techniques, natural language explanations, and context-aware reporting mechanisms may improve analyst comprehension while reducing cognitive workload during security investigations. Furthermore, empirical studies involving cybersecurity professionals should evaluate how different forms of explanations influence trust, decision quality, response speed, and situational awareness. Such research will contribute toward developing explainable cybersecurity systems that effectively support collaboration between human analysts and intelligent security platforms (Zhang et al., 2022; KM et al., 2022).

5.5 Adaptive Hybrid Threat Intelligence Models

Future research should also investigate hybrid cybersecurity models that integrate multiple machine learning techniques with Explainable AI to improve adaptive threat detection across diverse attack scenarios. Combining supervised learning, unsupervised anomaly detection, deep learning, reinforcement learning, and knowledge-based reasoning may improve detection accuracy while reducing false positives and false negatives. Explainability mechanisms should accompany each stage of the detection pipeline to ensure transparency throughout the decision-making process. Additionally, hybrid models should continuously learn from evolving threat intelligence, enabling proactive identification of emerging attack patterns while maintaining explainable outputs that support effective incident response and forensic investigations. Such adaptive systems would further strengthen the resilience of Zero Trust security frameworks against

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

advanced persistent threats and polymorphic cyberattacks (Gudala et al., 2021; Timadi & Obasi, 2025).

5.6 Trustworthy AI Governance and Policy-Aware Explainability

Another important research direction concerns the establishment of comprehensive AI governance frameworks that ensure explainable cybersecurity systems operate ethically, transparently, and responsibly. As organizations increasingly rely on automated security decisions, future studies should investigate standardized methods for validating explanation quality, measuring fairness, auditing AI decisions, and ensuring compliance with organizational policies and cybersecurity regulations.

Developing quantitative metrics for explainability effectiveness, accountability, and user trust would provide consistent benchmarks for evaluating AI-based security solutions. Furthermore, integrating governance policies directly into Zero Trust architectures may enable continuous monitoring of AI behavior while ensuring that automated defensive actions remain aligned with organizational security objectives and regulatory requirements (Pakina et al., 2024; Guembe et al., 2022).

5.7 Advanced Adaptive Zero Trust Architectures

Future cybersecurity research should also explore next-generation Zero Trust architectures that dynamically adapt security policies based on explainable AI-driven threat intelligence. Rather than relying on static authentication policies, adaptive Zero Trust systems should continuously evaluate contextual information, behavioral analytics, device trustworthiness, and environmental risk factors before granting or restricting access.

Explainability should remain an integral component of these adaptive architectures by providing transparent justifications for access control decisions, privilege adjustments, and automated containment actions. Such intelligent architectures would enhance organizational resilience against sophisticated cyber threats while supporting continuous verification and risk-aware access management. Integrating adaptive machine learning with transparent decision support can significantly improve the effectiveness of Zero Trust implementations in modern enterprise environments (Ejeofobiri et al., 2022; Haque, 2025).

Overall, future research should focus on developing cybersecurity solutions that are not only intelligent and adaptive but also transparent, trustworthy, scalable, and human-centered. Advances in lightweight XAI, explainable reinforcement learning, federated explainable intelligence, adaptive hybrid machine learning, AI governance, and dynamic Zero Trust architectures will collectively strengthen cyber resilience while enabling organizations to make informed, accountable, and timely security decisions.

Continued interdisciplinary research combining artificial intelligence, cybersecurity engineering, human-computer interaction, and governance frameworks will be instrumental in realizing the full potential of Explainable Artificial Intelligence for adaptive cyber threat detection in Zero Trust Network Architectures.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

VI. CONCLUSION

The increasing complexity and persistence of cyber threats have underscored the need for intelligent security architectures capable of providing continuous protection while maintaining transparency and accountability in decision-making. This study examined the integration of Explainable Artificial Intelligence (XAI) into adaptive cyber threat detection within Zero Trust Network Architectures (ZTNA), demonstrating that explainability significantly enhances the effectiveness of AI-driven cybersecurity systems. By complementing the Zero Trust principle of continuous verification with interpretable machine learning models, organizations can improve threat detection accuracy, strengthen automated response mechanisms, and foster greater confidence among cybersecurity analysts and stakeholders. The combination of adaptive learning and transparent decision support enables security teams to understand the rationale behind threat classifications, thereby improving operational efficiency and reducing the risks associated with opaque "black-box" AI systems (Guembe et al., 2022; Zhang et al., 2022).

The review further established that machine learning and deep learning techniques provide robust capabilities for detecting sophisticated cyberattacks through real-time anomaly identification and adaptive mitigation strategies. However, these capabilities become substantially more valuable when paired with explainability mechanisms that reveal the contributing factors behind security decisions. Explainable methods such as feature attribution, rule-based reasoning, and interpretable reinforcement learning facilitate faster incident investigation, improve regulatory compliance, and support informed security governance. These capabilities align closely with the objectives of Zero Trust architectures, where every access request and network interaction requires continuous validation and risk assessment (Gudala et al., 2021; Ejeofobiri et al., 2022).

The study also highlighted the growing importance of Explainable Reinforcement Learning (XRL) and AI governance in strengthening adaptive cyber defense. Explainable reinforcement learning enables autonomous security systems to justify dynamic policy adaptations and cyber deception strategies while maintaining transparency throughout the decision-making process. This capability contributes to trustworthy autonomous cybersecurity by balancing intelligent automation with human oversight and accountability. Such governance-oriented approaches enhance organizational confidence in AI-assisted security operations while reducing uncertainty associated with automated defensive actions (Pakina et al., 2024; KM et al., 2022).

Furthermore, the integration of lightweight XAI models, hybrid machine learning algorithms, and federated explainable intelligence demonstrates significant potential for addressing the scalability, privacy, and computational challenges of modern enterprise environments. Lightweight explainable models improve real-time deployment across resource-constrained systems, while federated explainable architectures support collaborative threat intelligence without compromising sensitive organizational data.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

Similarly, hybrid machine learning approaches improve detection performance by combining multiple analytical techniques with explainable threat intelligence, thereby enhancing both predictive accuracy and transparency across distributed Zero Trust infrastructures (Alshudukhi et al., 2025; Timadi & Obasi, 2025; Wei & Okafor, 2025).

Despite these advances, several implementation challenges remain, including balancing explainability with predictive performance, minimizing computational overhead, ensuring scalability across heterogeneous environments, and developing standardized evaluation metrics for explainable cybersecurity systems. Addressing these limitations will require continued collaboration among researchers, cybersecurity practitioners, and policymakers to establish interoperable frameworks, trustworthy AI governance models, and practical deployment strategies for enterprise security. Advances in adaptive intrusion detection systems integrated with Zero Trust principles further indicate that explainability will remain an essential component of future intelligent cybersecurity architectures capable of responding effectively to evolving attack landscapes (Haque, 2025; Zhang et al., 2022).

Overall, this study concludes that Explainable Artificial Intelligence represents a critical advancement in adaptive cyber threat detection for Zero Trust Network Architectures. By combining continuous verification, intelligent threat analysis, transparent decision-making, and adaptive learning, XAI-enabled Zero Trust systems provide a more resilient, accountable, and trustworthy cybersecurity framework. The continued evolution of explainable AI techniques, autonomous cyber defense, and federated intelligence is expected to further strengthen organizational resilience against increasingly sophisticated cyber threats while promoting responsible and transparent AI adoption in modern cybersecurity ecosystems (Guembe et al., 2022; Ejeofobiri et al., 2022; Alshudukhi et al., 2025).

References

1) Ejeofobiri, C. K., Adelere, M. A., & Shonubi, J. A. (2022). Developing adaptive cybersecurity architectures using Zero Trust models and AI-powered threat detection algorithms. *Int J Comput Appl Technol Res*, *11*(12), 607-621.

2) Guembe, B., Azeta, A., Osamor, V., & Ekpo, R. (2022, November). Explainable artificial intelligence, the fourth pillar of zero trust security. In *Proceedings of the International Conference on Information Systems and Emerging Technologies (ICISET)*.

3) Pakina, A. K., Sharma, A., & Pujari, M. (2024). AI Governance via Explainable Reinforcement Learning (XRL) for Adaptive Cyber Deception in Zero-Trust Networks.

4) Gudala, L., Shaik, M., & Venkataramanan, S. (2021). Leveraging machine learning for enhanced threat detection and response in zero trust security frameworks: An Exploration of Real-Time Anomaly Identification and Adaptive Mitigation Strategies. *Journal of Artificial Intelligence Research*, *1*(2), 19-45.

5) Alshudukhi, K. S., Ali, S., Humayun, M., & Alruwaili, O. (2025). Next-Generation Lightweight Explainable AI for Cybersecurity: A Review on Transparency and Real-Time Threat Mitigation. *Computer Modeling in Engineering & Sciences*, *145*(3), 3029.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.21423098

6) Haque, E. (2025). *Enhancing Network Security: Dynamical Intrusion Detection Systems Leveraging Zero Trust Architecture* (Master's thesis, Tennessee State University).

7) Timadi, M. E., & Obasi, E. C. M. (2025). A Zero Trust Hybrid Machine Learning Algorithms for Threat Detection and Prevention with Explainable Threat Intelligence. *University of Ibadan Journal of Science and Logics in ICT Research*, *15*(1).

8) KM, Z., Akhtaruzzaman, K., & Tanvir Rahman, A. (2022). Building trust in autonomous cyber decision infrastructure through explainable AI. *International Journal of Economy and Innovation*, *29*, 405-428.

9) Wei, L., & Okafor, S. (2025). Federated Explainable AI–Driven Cybersecurity Decision Support Architecture for Autonomous Threat Intelligence and Secure Resource Allocation in Multi-Cloud Enterprise Environments.

10) Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEe Access*, *10*, 93104-93139.

11) Ravikumar, V. (2025). Therapeutic Bot: Ethical Concerns in AI therapy for Neurodivergence. *J Int Scient Re Rep*.

12) Ezeagwuna, D. (2025). Artificial Intelligence for IT Service Management: Developing a Framework for ROI Optimization Using Service Now and Machine Learning Technologies. *Well Testing Journal*, *34*(S4), 394-419.

13) Warren, Brandon. (2025). Business Value Driven By Enterprise Architecture Transformation Brandon Warren. Journal of Tianjin University Science and Technology. 58. 10.5281/zenodo.20280991.

14) Adepoju, S. A., & Adepoju, M. A. (2024). From Portals to Case Graphs: A Reference Architecture and Benchmark for Safety Investigation Operations with Agentic Orchestration.

15) Khan, H. A. (2024). Data-Driven Epidemiological Modeling Using Machine Learning for Disease Spread Forecasting and Public Health Decision Support in The United States. *International Journal of Applied Mathematics*, *37*(6s), 178-192.