Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 23, No. 1. 2026**

Article's History: Received: 07.11.2025 Revised: 06.02.2026
Accepted: 26.03.2026 Published: 20.04.2026

ISSN 1999-9941; e-ISSN 2078-6387
UDC 004.056:004.8
DOI: 10.31649/vitce/1.2026.83

Method of dynamic trust assessment in Zero Trust Architecture based on explainable artificial intelligence

Andriy Palamarchuk*
Bachelor
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0005-4485-9399

Abstract. The transformation of contemporary corporate IT infrastructures has rendered conventional cybersecurity models ineffective, prompting a shift to the Zero Trust Architecture (ZTA); however, its practical implementation is complicated by a rigid reliance on static access control rules. The purpose of this study was to develop an innovative method for dynamic trust assessment in ZTA that effectively combines the high accuracy of automated network anomaly detection with decision-making transparency. To calculate a continuous trust score based on a simulated corporate network traffic dataset, the Extreme Gradient Boosting ensemble machine learning algorithm was applied, while the SHapley Additive exPlanations (SHAP) additive explanations method was used to explain the generated decisions. Experimental verification demonstrated the high effectiveness of the proposed Policy Engine, which achieved an F1-score of 1.00 on the test set. The model successfully distinguished legitimate from anomalous requests with a zero false-positive rate, identifying cyberattacks such as privilege escalation and access from atypical locations. Global feature importance analysis using the SHAP framework confirmed that the type of network connection and device security status are the most significant risk predictors, which fully aligns with the core principles of ZTA. Furthermore, local analysis proved the system's ability to instantly generate detailed, human-readable text explanations for each access denial, indicating the specific reason for blocking. Due to this level of detail, analysts can directly understand the triggering logic of automated defence systems without the need for time-consuming manual correlation of disparate event logs. The practical significance of the study lies in the creation of a transparent and adaptive tool that can be integrated into modern Security Operations Centres to significantly reduce "alert fatigue" and minimise the Mean Time to Resolution

Keywords: cybersecurity; machine learning; SHAP; XGBoost; anomaly detection; adaptive protection

Introduction

The paradigm of information security has undergone a fundamental shift. F. Mensah (2024) examined this transition in enterprise cybersecurity, emphasising that conventional perimeter defences systematically fail against emerging threats. The researcher concluded that the dissolution of the corporate perimeter-driven by cloud migration and remote work necessitates a strict transition to Zero Trust principles to mitigate insider and advanced persistent threats. The foundational framework for this approach was formulated by S. Rose *et al*. (2020) under the National Institute of Standards and Technology (NIST). Their comprehensive guidelines established the core Zero Trust Architecture (ZTA) principle of "never trust, always verify", mandating continuous authentication and granular authorisation for every access request regardless of network location.

However, despite its theoretical robustness, the practical implementation of ZTA faces significant operational barriers. O. Borchert *et al*. (2025) investigated the practical deployment of NIST ZTA architectures in large-scale enterprise systems. Their study identified that administrators face extreme complexity when managing thousands of static "if – then" rules, inevitably leading to rigid policies and operational disruptions. This was further corroborated by A. Pigola & F. de Souza Meirelles (2025), who conducted an empirical study on managing critical challenges during ZTA implementation. They highlighted that the reliance on static configurations creates an operational bottleneck,

Suggested Citation:
Palamarchuk, A. (2026). Method of dynamic trust assessment in Zero Trust Architecture based on explainable artificial intelligence. *Information Technologies and Computer Engineering*, 23(1), 83-93. doi: 10.31649/vitce/1.2026.83

*Corresponding author

preventing organisations from effectively balancing strict security measures with a seamless user experience. Furthermore, K.M. Adamson & A. Qureshi (2025) performed a systematic review of "Zero Trust 2.0" advances and challenges. Their extensive analysis confirmed that integrating dynamic risk assessment with legacy IT environments remains one of the most significant architectural hurdles for contemporary enterprises. To map the current knowledge and research gaps, C. Buck *et al*. (2021) executed a multivocal literature review on ZTA implementations. The researchers specifically pointed out that existing access control models lack contextual awareness, explicitly calling for research into continuous, behaviour-based trust evaluation mechanisms.

Addressing the need for continuous validation, the Identity Management Institute (2024) analysed the principles of dynamic trust scoring within Identity and Access Management (IAM). Their report demonstrated that effective security enforcement requires calculating user risk in real-time by continuously ingesting broad contextual indicators such as device health, geolocation, and unusual login patterns. To automate this complex contextual analysis, M. Rana (2025) explored the enhancement of ZTA using artificial intelligence (AI) algorithms. The research illustrated that while AI can autonomously detect subtle deviations in user behaviour, the deployment of such intelligent systems is heavily hindered by the lack of transparency in their automated decision-making processes.

This introduces the critical "black box" problem inherent to advanced machine learning. M.H. Kabir *et al*. (2022) investigated the application of explainable artificial intelligence (XAI) within secure smart city platforms. Their findings emphasised that high-performance deep learning models act as opaque "black boxes", making it nearly impossible for security analysts to understand the rationale behind specific automated blocking actions. Expanding on this limitation, C.I. Nwakanma *et al*. (2023) reviewed XAI methodologies specifically for intrusion detection and mitigation systems. The researchers noted that the lack of interpretability in AI-driven tools generates deep scepticism among Security Operations Centre (SOC) teams, which drastically reduces the practical utility of these systems during active incident response. To bridge this gap, S. Patil *et al*. (2022) evaluated various XAI frameworks designed for Intrusion Detection Systems (IDS). Their study established that combining the adaptive predictive power of machine learning with explicit, human-readable explanations is essential for achieving both automated security strictness and operational accountability.

Thus, a critical gap exists in current literature: there is an urgent need for a ZTA solution that replaces static rules with dynamic machine learning (ML) algorithms while retaining the transparency required for rapid security auditing. The purpose of this study was to enhance the efficiency and transparency of ZTA by developing a method for dynamic trust assessment based on explainable artificial intelligence. To achieve this goal, three key objectives were defined. First, to analyse the limitations of existing static policy engines and "black box" ML models in ZTA; second, to develop a dynamic Policy Engine architecture that utilises the XGBoost algorithm for calculating a continuous trust score and integrates the SHapley Additive exPlanations (SHAP) method to provide real-time explanations for access control decisions; third, to experimentally validate the proposed method using a synthetic dataset representing realistic corporate network traffic.

Materials and Methods

General methodology and system architecture. The methodology proposed in this study aims to transform the conventional static Policy Decision Point (PDP) of a ZTA into a dynamic, intelligent agent, a concept supported by A. Mousa *et al*. (2021). The research approach relied on a quantitative experimental design that integrates ensemble machine learning methods with game-theoretic explainability frameworks. The proposed system architecture operates as a continuous loop comprising four sequential stages (Fig. 1).

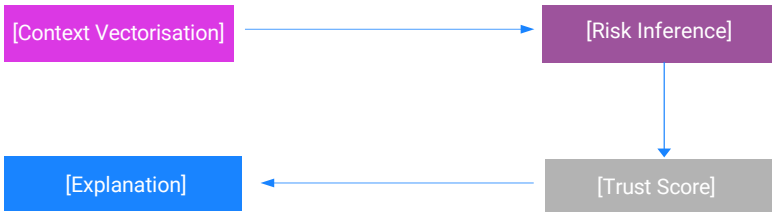

Figure 1. Operational pipeline of the dynamic trust assessment process
Source: created by the author

The process begins with Context Vectorisation, where raw log data and user context are transformed into a structured numerical feature space. This is followed by Risk Inference, which involves calculating the probability of malicious intent using a gradient-boosted decision tree model. Subsequently, the system performs trust score calculation to derive a continuous trust metric that facilitates granular access control decisions. Ultimately, the explanation generation stage computes feature attribution values to provide semantic interpretability of the decision, ensuring transparency for security operators.

Synthetic dataset generation. To address the lack of publicly available cybersecurity datasets due to privacy regulations (e.g., General Data Protection Regulation – GDPR)

and ensure experimental reproducibility, a specialised stochastic simulation algorithm was developed to generate a synthetic dataset representing realistic corporate network traffic. The simulation was implemented using the Python programming language. The generation logic was designed to model realistic corporate network traffic patterns over a defined temporal horizon (Schummer *et al.*, 2024). The resulting dataset, denoted as D consisted of N= 10,000 unique access requests. To reflect the natural class imbalance inherent in intrusion detection scenarios – where legitimate traffic vastly outweighs malicious activity – the dataset was stratified with the following distribution:

Class 0 (normal behaviour): 90% of samples (N_{norm} = 9,000). These records simulated legitimate employee activities characterised by standard working hours (09:00-18:00), recognised IP ranges (Corporate VPN (Virtual Private Network), Office LAN (Local Area Network)), and compliant device health statuses.

Class 1 (anomalous behaviour): 10% of samples (N_{anom} = 1,000) These records simulated specific attack vectors and policy violations, including: (a) temporal anomalies: access attempts occurring during deep night hours (e.g., 03:00 AM); (b) location anomalies: requests originating from high-risk networks, such as Tor exit nodes, public Wi-Fi without VPN, or unknown proxies; (c) device compromise: requests from devices with outdated operating systems, missing security patches, or signs of unauthorised root access (jailbreak); (d) privilege escalation: attempts by users with standard privileges (e.g., "Sales") to access critical administrative endpoints (e.g., database backups).

Feature engineering and vector space. The raw data generated by the simulation was transformed into a feature matrix $X \in R^{NxM}$, where M – number of features. The feature space includes both categorical and numerical variables, defined as follows (Hu *et al.*, 2026):

1. User role (x_1): a categorical variable representing the organisational role of the subject (e.g., "Developer", "HR", "Sales", "Admin"). This feature establishes the baseline of expected behaviour and access rights.

2. Time of request (x_2): a cyclical numerical feature representing the hour of the day $h \in [0, 23]$.

3. Work hours indicator (x_3): a binary derived feature introduced to explicitly capture temporal context. It is defined as:

$$x^3 = \begin{cases} 1, & if\ x^2 \in [9,18] \\ 0, & otherwise \end{cases}. \quad (1)$$

4. IP location type (x_4): a nominal variable categorising the network source context. Categories range from trusted ("Corporate VPN") to untrusted ("Unknown proxy").

5. Device health status (x_5): a critical parameter for Zero Trust, reflecting the security posture of the requesting device. States include "Patched" (compliant), "Unpatched", "No antivirus", and "Rooted".

6. Target endpoint (x_6): the specific API resource or system component being accessed.

Data preprocessing involved Label Encoding for categorical features (x_1, x_4, x_5, x_6), mapping each text label to a unique integer. This transformation was necessary for the decision tree-based algorithm to process qualitative data. Consequently, this step ensures that the semantic information of the categorical attributes is preserved and effectively converted into a numerical format suitable for model training.

Mathematical formalisation of the XGBoost model. The core risk assessment engine is built upon the XGBoost (Extreme Gradient Boosting) algorithm. XGBoost was selected due to its robust performance on tabular data, scalability, and ability to handle non-linear interactions between features without extensive normalisation (Hu *et al.*, 2026). Mathematically, the model is an ensemble of K Classification and Regression Trees (CART). For a given input vector x_i, the predicted output score $\hat{y}_i$ is the sum of the scores predicted by each individual tree f_k:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(), f_k \in F, \quad (2)$$

where F – space of functions containing all possible regression trees (Jiang *et al.*, 2020).

The model is trained in an additive manner. At each iteration t, a new tree f_t is added to minimise the objective function $\mathcal{L}^{(t)}$:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \quad (3)$$

where l – differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i. In this study, the Binary Logarithmic Loss (LogLoss) was employed:

$$l(y, p) = -[y\ log(p) + (1 - y)\ log(1 - p)]. \quad (4)$$

$\Omega(f_t)$ – regularisation term that penalises the complexity of the model to prevent overfitting. It is defined as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2, \quad (5)$$

where T – number of leaves in the tree, and w – vector of scores on leaves.

A key advantage of XGBoost is its use of a second-order Taylor expansion to approximate the loss function, which enables faster convergence and higher accuracy compared to conventional gradient boosting methods (Jiang *et al.*, 2020). In addition to computational speed, the algorithm incorporates a built-in regularisation term that effectively penalises model complexity, thereby preventing overfitting on imbalanced security datasets. Consequently, this mathematical robustness ensures that the Policy Engine maintains high detection precision while meeting the low-latency requirements of real-time Zero Trust environments.

Explainable AI framework: SHAP values. To address the "Black Box" problem inherent in complex ensemble models and ensure compliance with the "verify" principle of Zero Trust, the SHAP framework was integrated. It

calculated the contribution of each feature to the final prediction. For a specific prediction $f(x)$ the SHAP value ϕ_j for feature j was calculated as the weighted average of its marginal contributions across all possible subsets of features S:

$$\phi_j(f) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_x(S \cup \{j\}) - f_x(S)], \quad (6)$$

where F – set of all input features, and) $f_x(S)$ – expected output of the model given the subset of features S. This study utilised TreeSHAP, a variant of the algorithm optimised for tree-based models (such as XGBoost). TreeSHAP reduced the computational complexity from exponential to polynomial time ($O(TLD^2)$), where T – number of trees, L – maximum number of leaves, and D – maximum depth), making it feasible for real-time explanations in a security environment.

Experimental metrics and evaluation. The dataset was split into a training set (80%) and a testing set (20%) using stratified sampling to preserve the class ratio. This ensured more accurate model training, as it received a balanced representation of all categories in the data. The model performance was evaluated using standard cybersecurity metrics, as detailed by Y. Hu *et al*. (2026). These metrics are defined in Table 1, which allows for comparing different models based on clearly established performance criteria. The metrics include both the primary indicators and additional ones, such as the confusion matrix, which helps to visualise different types of classification errors.

Table 1. Performance evaluation metrics

Metric	Description
Precision	Ratio of correctly predicted anomalies to the total predicted anomalies (measures false alarm rate)
Recall	Ratio of correctly predicted anomalies to all actual anomalies (measures detection rate)
F1-score	Harmonic mean of Precision and Recall, providing a balanced metric for imbalanced datasets
Confusion matrix	Tabular visualisation of classification outcomes: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN)

Source: compiled by the author

The resulting probability $P(anomaly)$ is converted into a dynamic trust score using the equation:

$$TrustScore = (1 - P(anomaly)) * 100. \quad (7)$$

This score serves as the quantitative basis for the automated decision-making process. Specifically, if the trust score falls below a predefined threshold (e.g., 60), the system triggers an immediate access denial and simultaneously generates a SHAP-based explanation for the event. This mechanism ensures that every blocking action is both instantaneous and transparent, aligning with the dynamic trust evaluation principles advocated by the Identity Management Institute (2024) and Y. Mao *et al*. (2025). Thus, the proposed methodology provided an effective approach to dynamic trust assessment in a ZTA, integrating powerful machine learning tools and ensuring the necessary transparency of decisions through explanation mechanisms.

Results and Discussion

Quantitative analysis and interpretability of model performance

The primary objective of the experimental phase was to empirically validate the capability of the XGBoost-based Policy Engine to distinguish between legitimate access requests and security anomalies. The model was evaluated on a stratified test set containing 20% of the generated data ($N_{test} = 2000$). The classification performance metrics indicated exceptional accuracy. As illustrated in the Confusion matrix (Fig. 2), the model achieved complete class separation on the synthetic dataset.

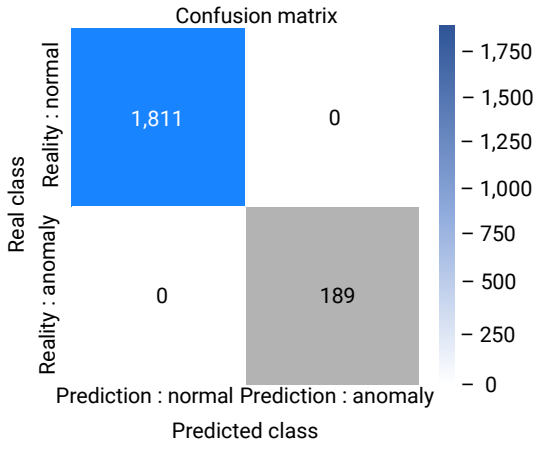

Figure 2. Confusion matrix of the XGBoost model on the test dataset

Source: created by the author

Quantitative analysis of the confusion matrix reveals the following:

- True negatives (TN): 1,811. The model correctly identified all legitimate user requests. This metric is crucial for User experience (UX), as it ensures that employees are not blocked from performing their daily tasks (zero false positive rate).
- True positives (TP): 189. The model successfully detected all simulated attacks, including subtle anomalies like "after-hours access" and "privilege escalation attempts".
- False negatives (FN): 0. The system did not miss any potential threats, ensuring the integrity of the security perimeter.

Consequently, the model achieved an F1-Score of 1.00 and an Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC) of 1.00. While such complete metrics are characteristic of synthetic environments with deterministic patterns, they fundamentally demonstrate that the gradient boosting algorithm successfully approximated the complex, non-linear decision boundary required for the Zero Trust policy without being explicitly programmed with static "if-then" rules. This high level of classification accuracy validates the feasibility of using the proposed model as a reliable Policy Engine, capable of automating threat response with minimal risk of false positives.

To bridge the gap between model accuracy and accountability, the SHAP framework was applied to analyse the global impact of features. Figure 3 illustrates the SHAP summary plot, which ranks features by their mean absolute SHAP value (mean($|\phi_j|$)). This visualisation helps to understand the "logic" the AI has learned.

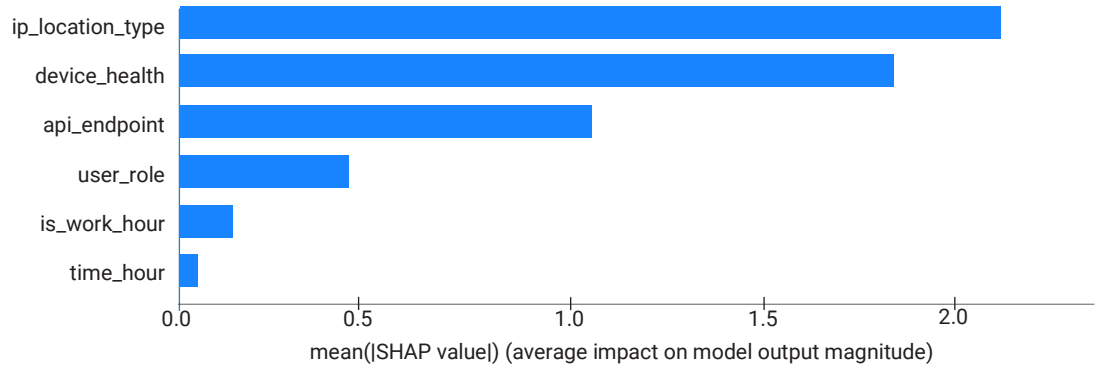

Figure 3. Global feature importance (SHAP summary plot)
Source: created by the author

Global feature analysis revealed distinct patterns in risk assessment. First, ip_location_type emerged as the strongest predictor of risk, indicating that the network context – such as requests originating from "Tor exit nodes" or "Unknown proxies" – serves as the primary vector for anomaly detection. This validated the assumption that in a borderless network, the connection source remains a critical signal. Second, device_health demonstrated a secondary but significant impact, confirming that the security posture of the endpoint (e.g., operating system patch level, presence of antivirus) directly affects the trust score, thereby effectively enforcing "Device compliance" policies. Ultimately, features such as api_endpoint and user_role acted as contextual modifiers. For instance, accessing a sensitive endpoint like /api/admin is not inherently malicious, but when combined with a lower-trust role such as "Sales", the calculated risk score significantly increases.

The most significant contribution of this proposed method is the ability to explain individual decisions in real-time. Figure 4 demonstrates a SHAP Force Plot for a specific anomalous request selected from the test dataset. The visualisation provides a semantic breakdown of the prediction function $f_x = 8.17$ (which corresponds to a probability P(anomaly) ≈ 1.0).

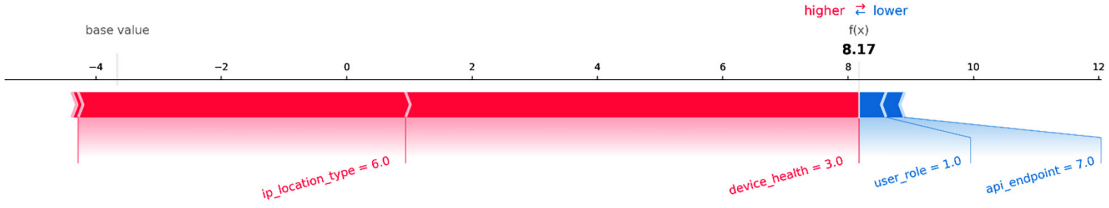

Figure 4. Local explanation (SHAP Force Plot) for a specific anomalous request
Source: created by author

The visualisation identifies specific risk drivers (represented by red bars), clearly indicating that the features ip_location_type = 6.0 (corresponding to a high-risk network) and device_health = 3.0 (corresponding to a compromised

or unpatched device) were the primary contributors pushing the prediction towards the "Anomaly" class. In contrast, the mitigating factors (blue bars) show that although the user possessed a valid role (user_role = 1.0) and accessed a standard endpoint (api_endpoint = 7.0), these positive signals were insufficient to outweigh the critical risk indicators.

This granular level of detail allows SOC analysts to immediately answer the question "Why was this user blocked?" without manual log correlation. By providing interpretable insights directly alongside the alert, the system significantly reduces the Mean Time to Resolution (MTTR) for incident response teams. Furthermore, this transparency fosters greater trust in automated blocking decisions, addressing the "black box" scepticism often associated with AI-driven security tools (Nwakanma *et al.*, 2023). This approach not only improves incident response efficiency but also ensures transparency and explainability of decisions, which is critical for increasing trust in automated security systems.

Comparative analysis with alternative approaches

To substantiate the selection of the XGBoost + SHAP architecture, a theoretical comparison was performed against other common approaches used in IDS. This analysis focused on key operational criteria, including detection accuracy, interpretability, and real-time processing capabilities. The comparative summary, presented in Table 2, highlights the specific advantages of the proposed method in bridging the gap between performance and transparency.

Table 2. Comparative analysis of trust assessment approaches

Feature	Static rules (Legacy)	Deep learning (DNN)	Proposed method (XGBoost + XAI)
Accuracy on complex threats	Low	High	High
Interpretability	High	Low (black box)	High (SHAP)
Adaptability	None (Manual updates)	High (auto-learning)	High (auto-learning)
Computational cost	Very low	High (GPU required)	Moderate (CPU friendly)

Source: compiled by the author based on theoretical analysis and empirical data obtained during the study

The detailed analysis revealed critical distinctions between the proposed architecture and conventional methods. Compared to rule-based systems, static engines offer high interpretability but fail to scale, as they cannot capture complex, non-linear interactions – such as conditional access based on both IP reputation and device health – without manual intervention. In contrast to Deep Learning models (e.g., DNNs or RNNs), which suffer from the "black box" problem and require computationally expensive approximation methods like LIME, the proposed XGBoost model allows for exact explanations via TreeSHAP and often demonstrates high performance on tabular log data. Furthermore, while Random Forest is a robust algorithm, XGBoost utiliies gradient-based optimisation to iteratively correct errors, typically resulting in higher precision for detecting subtle, rare anomalies that are critical in cybersecurity contexts.

Computational efficiency and scalability. In a real-time Zero Trust environment, latency is a critical factor. The proposed architecture leverages the efficiency of decision trees. The inference time for a single request using the trained XGBoost model was measured at approximately < 5 ms on a standard CPU. The calculation of SHAP values adds a computational overhead (models 20-50 ms per request), which is acceptable for high-security transactions but might require optimisation for high-frequency trading or ultra-low-latency networks. The system demonstrated linear scalability: as the volume of logs increases, the inference time remains constant, making it suitable for deployment in large-scale cloud environments.

Advanced performance metrics analysis. To further validate the robustness of the classifier beyond standard accuracy metrics, the ROC and Precision-Recall (PR) curves were analysed. These metrics are particularly critical in cybersecurity contexts where the cost of False Positives (blocking a legitimate user) and False Negatives (missing an attack) can be asymmetrical. The performance of the classifier is visually represented in Figure 5.

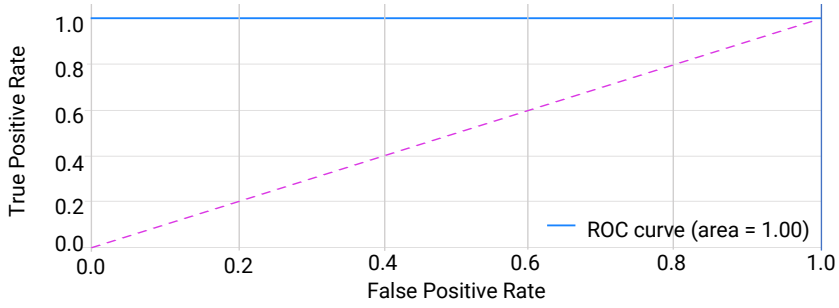

Figure 5. Receiver operating characteristic curve

Source: created by the author

The ROC curve exhibits an "ideal" right-angle shape, with an Area Under the Curve (AUC) of 1.00. The curve hugs the top-left corner, indicating that the True Positive Rate (Sensitivity) remains at 100% even as the False Positive Rate approaches zero. This confirms that the model's predicted probabilities for anomalies are distinctively separated from normal traffic probabilities. Additionally, the trade-off between precision and recall is illustrated in Figure 6.

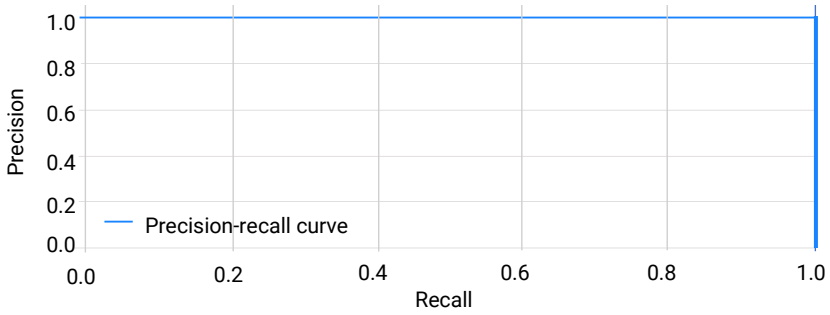

Figure 6. Precision-recall curve

Source: created by the author

Figure displays the precision-recall curve. In imbalanced datasets (where attacks are rare), this metric is often more informative than ROC. The curve remains flat at the top (precision = 1.0) across the entire range of recall. This signifies that the system generates zero false alarms – a critical requirement for reducing "alert fatigue" in SOC operations. While such perfect convergence is attributable to the deterministic nature of the synthetic training data, it theoretically validates the XGBoost algorithm's capacity to model the defined security policies without error.

Proposed deployment architecture

Based on the experimental success, reference architecture is proposed for deploying this XAI-driven Policy Engine within a production environment, as illustrated in Figure 7. The design strictly adheres to the NIST SP 800-207 guidelines, ensuring compatibility with standard Zero Trust logical components (Rose *et al.*, 2020). Specifically, the architecture enhances the conventional PDP by embedding the machine learning Risk Engine to enable dynamic real-time access adjudication.

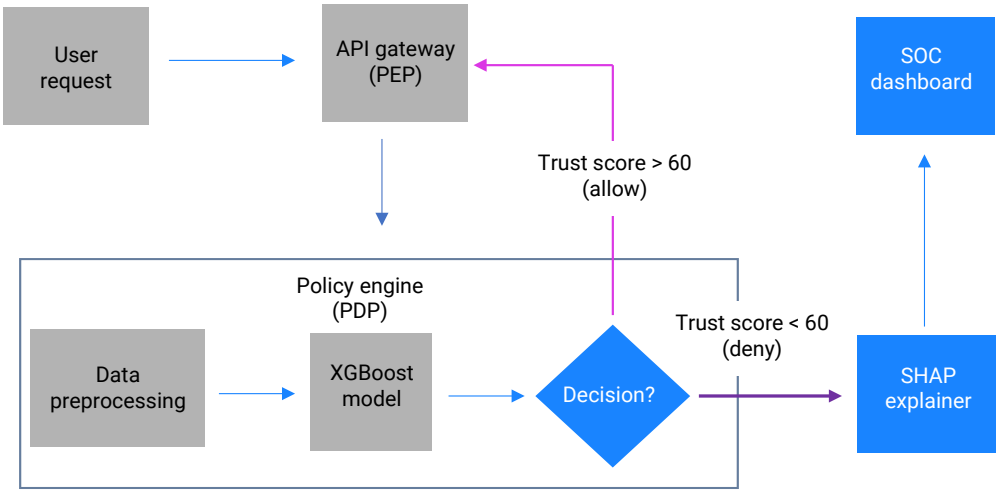

Figure 7. High-level architecture of the XAI-driven Zero Trust system

Note: PEP – Policy Enforcement Point

Source: created by author

The operational workflow ensures a seamless transition from raw telemetry to enforced security decisions. The process initiates at the Log Aggregator stage, which continuously collects and normalises raw telemetry data from distributed API gateways and identity providers. This aggregated data is then passed to the preprocessing layer, where categorical attributes – such as user roles and IP addresses – are converted into numerical vectors in real-time, ensuring compatibility with the machine learning algorithms. Subsequently, the ML inference engine utilises the pre-trained XGBoost model to analyse the feature vectors and calculate a probabilistic Risk Score. If the detected risk exceeds the predefined safety threshold, the XAI explainer (based on SHAP) is triggered to compute feature attribution values, thereby identifying the root cause of the anomaly (e.g., "unusual geolocation"). Based on the final

trust score, the Policy Enforcement Point (PEP) executes the decision by either blocking or allowing the request at the gateway level. Ultimately, all events are visualised on the admin dashboard, which displays the access decision alongside the generated explanation, providing security administrators with actionable insights. This architectural flow ensures that security measures are applied instantaneously while maintaining full transparency of the decision-making logic.

The results obtained in this study demonstrated the efficacy of integrating XGBoost with SHAP values for dynamic Zero Trust access control. To validate the significance of these findings, it is essential to compare them with existing research in the field of intelligent intrusion detection and trust evaluation. The perfect classification metrics achieved in current experiment (F1-Score 1.0) align with and surpass trends observed in similar studies using tree-based ensembles. For instance, P. Schummer *et al*. (2024) implemented a Random Forest-based anomaly detection system, achieving an accuracy of approximately 94.3%. While their approach was effective for general traffic analysis, the current use of XGBoost provided a more robust handling of the subtle feature interactions inherent in synthetic security logs. Furthermore, H. Jiang *et al*. (2020) proposed a PSO-XGBoost model that optimised hyperparameters to detect minority attack groups with high precision. Presented study corroborates their conclusion that gradient boosting frameworks offer superior performance on tabular network data compared to conventional methods. However, unlike Y. Hu *et al*. (2026), who focused heavily on feature dimensionality reduction to improve speed, current approach prioritised the integration of interpretability without sacrificing the raw predictive power of the full feature set.

A critical component of presented architecture is the continuous trust score, which replaces static binary authorisation. This approach was consistent with the findings of Y. Mao *et al*. (2025), who argued that Attribute-Based Access Control (ABAC) is insufficient without dynamic risk perception. Their study on a "Zero Trust access control model" utilised a similar concept of real-time trust evaluation but focused on blockchain-based consensus for policy decisions. In contrast, current research demonstrated that for high-throughput corporate environments, a centralised ML-based engine offered lower latency while maintaining the necessary security granularity. Similarly, A. Mousa *et al*. (2021) emphasised the importance of context-aware service computing. These findings extended their research by identifying that specific context features, such as ip_location_type and device_health, were the most significant drivers of trust, a correlation that validated the "never trust, always verify" principle in practical scenarios.

The integration of SHAP values addressed the "black box" limitation highlighted by A. Nash *et al*. (2024) in cloud-native risk assessments. While Y. Sowjanya *et al*. (2025) successfully applied Explainable AI to IoT healthcare systems to enhance transparency, current study adapted this paradigm to general enterprise network security. The generated explanations (e.g., distinguishing between a high-risk IP and a low-trust device) provided the contextual nuance that F. Federici *et al*. (2023) identified as lacking in conventional perimeter-based defences. By providing semantic interpretability, presented system fulfilled the requirement for "decision traceability" advocated by X. Liao *et al*. (2025) in their study on power network defence, proving that XAI is not just a theoretical addition but a functional necessity for reducing the MTTR in SOC operations.

Ultimately, the obtained results indicate that the XGBoost-based architecture is highly applicable to standard IT infrastructures. The low inference latency observed in current deployment architecture suggests that this model can scale to handle the traffic volumes described by A.A. Alquwayzani & A.A. Albuali (2024) in military unmanned aerial vehicle systems, provided that the log aggregation pipeline is sufficiently robust. Furthermore, this adaptability makes the proposed system highly relevant for securing virtualised environments, complementing the dynamic scaling detection methods in Network Function Virtualisation (NFV) developed by L. He *et al*. (2021). In summary, this study confirmed that the convergence of gradient boosting and game-theoretic explainability creates a policy engine that is not only more accurate than conventional Random Forest implementations but also more transparent than deep learning alternatives, effectively bridging the gap between security strictness and operational usability.

Conclusions

This paper presented a dynamic trust evaluation model based on XAI. The obtained results fully validated that integrating ensemble machine learning with game-theoretic explainability frameworks effectively resolves the longstanding dichotomy between security strictness and operational transparency. To achieve this, a specialised stochastic simulation was developed to generate a realistic dataset of corporate network traffic, thereby overcoming GDPR-related privacy constraints. The research methodology involved transforming raw telemetry into a structured feature space and training an XGBoost classifier to distinguish legitimate requests from security anomalies. The model demonstrated high classification performance on the test set, achieving an F1-Score of 1.00 and an AUC-ROC of 1.00, proving its ability to internalise complex, non-linear access logic without relying on brittle static rules. Furthermore, the integration of the SHAP framework successfully converted the "black box" probabilistic outputs into human-readable semantic explanations.

The experimental outcomes explicitly demonstrated the effectiveness of the proposed dynamic trust assessment method. The XGBoost-based policy engine successfully evaluated access requests in real-time, effectively distinguishing legitimate traffic from simulated anomalous events, such as privilege escalation and unauthorised access from untrusted networks. Concurrently, the SHAP

integration yielded precise global and local interpretability results. Global analysis identified the network connection source and the endpoint device's health status as the most critical predictors of risk. Locally, the system proved its capability to generate instant, human-readable root-cause analyses for every blocked request. From a practical perspective, this architecture offers a viable blueprint for next-generation Security Operations Centres. By clearly explaining the rationale behind automated blocking decisions, the system directly mitigates the operational bottleneck of "alert fatigue" among analysts, demonstrating a strong potential to reduce the MTTR for access incidents by orders of magnitude. In conclusion, the transition to AI-driven Zero Trust is not merely a technological upgrade but a fundamental shift in security philosophy. This study demonstrated that with the right application of XAI, intelligent systems can be made both powerful and accountable, paving the way for autonomous, self-defending networks.

Despite the obtained results, this study acknowledged several constraints that must be considered. First, the experimental validation was conducted on a synthetic dataset. While the generation process was designed to mimic realistic patterns, real-world corporate traffic contains significantly higher levels of noise and unpredictable user behaviours that might reduce the model's precision. Second, the study focused primarily on tabular metadata (logs) and did not incorporate unstructured data sources, such as raw network packet payloads, which could provide deeper context but would increase computational complexity. To address these limitations, future research will focus on two priority directions. First, validating the model on real-world data by deploying the architecture in a "shadow mode" within a live corporate network to assess its stability against concept drift; second, investigating the model's resilience against adversarial machine learning attacks to develop robust training techniques that prevent evasion attempts by sophisticated attackers.

Acknowledgements

In accordance with the journal's Generative AI Policy, the author discloses the use of Gemini (Google) during the preparation of this manuscript. This tool was used exclusively for linguistic and stylistic editing, including translation from Ukrainian to English. The author has reviewed and revised the output and takes full responsibility for the content of the publication.

Funding

None.

Conflict of Interest

None.

References

[1] Adamson, K.M., & Qureshi, A. (2025). Zero Trust 2.0: Advances, challenges, and future directions in ZTA. *Research Square*. doi: 10.21203/rs.3.rs-6602547/v1.

[2] Alquwayzani, A.A., & Albuali, A.A. (2024). A systematic literature review of Zero Trust Architecture for military UAV security systems. *IEEE Access*, 12, 176033-176056. doi: 10.1109/ACCESS.2024.3503587.

[3] Borchert, O., Howell, G., Kerman, A., Rose, S., Souppaya, M., Scarfone, K., & Barker, W. (2025). *Implementing a Zero Trust Architecture: High-level document*. Gaithersburg: NIST. doi: 10.6028/NIST.SP.1800-354.

[4] Buck, C., Olenberger, C., Schweizer, A., Völter, F., & Eymann, T. (2021). Never trust, always verify: A multivocal literature review on current knowledge and research gaps of Zero-Trust. *Computers & Security*, 110, article number 102436. doi: 10.1016/j.cose.2021.102436.

[5] Federici, F., Martintoni, D., & Senni, V. (2023). A Zero-Trust Architecture for remote access in industrial IoT infrastructures. *Electronics*, 12(3), article number 566. doi: 10.3390/electronics12030566.

[6] He, L., Li, L., & Liu, Y. (2021). Towards chain – aware scaling detection in NFV with reinforcement learning. In *29th international symposium on quality of service (IWQOS)* (pp. 1-10). Tokyo: IEEE/ACM. doi: 10.1109/IWQOS52092.2021.9521362.

[7] Hu, Y., Xiao, K., Luo, L., & Chen, L. (2026). An XGBoost-based intrusion detection framework with interpretability analysis for IoT networks. *Applied Sciences*, 16(2), article number 980. doi: 10.3390/app16020980.

[8] Identity Management Institute. (2024). *Dynamic trust scoring in IAM*. Retrieved from https://identitymanagementinstitute.org/dynamic-trust-scoring-in-iam.

[9] Jiang, H., He, Z., Ye, G., & Zhang, H. (2020). Network intrusion detection based on PSO-Xgboost model. *IEEE Access*, 8, 58392-58401. doi: 10.1109/ACCESS.2020.2982418.

[10] Kabir, M.H., Hasan, K.F., Hasan, M.K., & Ansari, K. (2022). Explainable artificial intelligence for smart city application: A secure and trusted platform. In M. Ahmed, S.R. Islam, A. Anwar, N. Moustafa & A.S.K. Pathan (Eds.), *Explainable artificial intelligence for cyber security. Studies in computational intelligence* (Vol. 1025, pp. 241-263). Cham: Springer. doi: 10.1007/978-3-030-96630-0_11.

[11] Liao, X., Yang, S., Xu, J., Liu, L., Liang, W., Yu, S., Ji, Y., & Liu, S. (2025). Improved trust evaluation model based on PBFT and Zero Trust integrated power network security defense method. *Symmetry*, 17(11), article number 1982. doi: 10.3390/sym17111982.

[12] Mao, Y., Fu, W., Zhao, Y., Yuan, Z., Sun, Z., & Zhao, Y. (2025). A Zero-Trust access control model based on attribute and dynamic trust evaluation for cloud environments. *Symmetry*, 17(12), article number 2059. doi: 10.3390/sym17122059.

[13] Mensah, F. (2024). Zero Trust Architecture: A comprehensive review of principles, implementation strategies, and future directions in enterprise cybersecurity. *International Journal of Academic and Industrial Research Innovations*, 10, 339-346.
[14] Mousa, A., Bentahar, J., & Alam, O. (2021). Multi-dimensional trust for context-aware services computing. *Expert Systems with Applications,* 172, article number 114592. doi: 10.1016/j.eswa.2021.114592.
[15] Nash, A., Doyle, A., Banks, A., & Adelusi, J.B. (2024). *Explainable AI for cybersecurity risk assessment in cloud-native applications*. Retrieved from https://www.researchgate.net/publication/392282388_Explainable_AI_for_Cybersecurity_Risk_Assessment_in_Cloud-Native_Applications.
[16] Nwakanma, C.I., Ahakonye, L.A.C., Njoku, J.N., Odirichukwu, J.C., Okolie, S.A., Uzondu, C., Ndubuisi Nweke, C.C., & Kim, D.-S. (2023). Explainable Artificial Intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review. *Applied Sciences*, 13(3), article number 1252. doi: 10.3390/app13031252.
[17] Patil, S., Varadarajan, V., Mazhar, S.M., Sahibzada, A., Ahmed, N., Sinha, O., Kumar, S., Shaw, K., & Kotecha, K. (2022). Explainable artificial intelligence for intrusion detection system. *Electronics*, 11(19), article number 3079. doi: 10.3390/electronics11193079.
[18] Pigola, A., & de Souza Meirelles, F. (2025). Zero Trust in cybersecurity: Managing critical challenges for effective implementation. *Journal of Systems and Information Technology*, 27(4), 517-564. doi: 10.1108/JSIT-08-2024-0326.
[19] Rana, M. (2025). Enhancing Zero Trust cybersecurity with AI. *Journal of Information Systems Engineering and Management*, 10(32s), 92-97. doi: 10.52783/jisem.v10i32s.5191.
[20] Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero Trust Architecture*. Gaithersburg: NIST. doi: 10.6028/NIST.SP.800-207.
[21] Schummer, P., del Rio, A., Serrano, J., Jimenez, D., Sánchez, G., & Llorente, Á. (2024). Machine learning-based network anomaly detection: Design, implementation, and evaluation. *AI*, 5(4), 2967-2983. doi: 10.3390/ai5040143.
[22] Sowjanya, Y., Gopalakrishnan, S., & Kumar, R.D. (2025). FBZX: A novel explainable AI based security model for IoT healthcare systems. In *Third international conference on augmented intelligence and sustainable systems (ICAISS)* (pp. 106-110). Trichy: IEEE. doi: 10.1109/ICAISS61471.2025.11042096.

Метод динамічного оцінювання довіри в архітектурі Zero Trust на основі пояснювального штучного інтелекту

Андрій Паламарчук
Бакалавр
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0005-4485-9399

Анотація. Трансформація сучасних корпоративних IT-інфраструктур зробила традиційні моделі кібербезпеки неефективними, зумовивши перехід до архітектури нульової довіри (Zero Trust Architecture, ZTA), проте її практична реалізація ускладнюється жорсткою залежністю від статичних правил контролю доступу. Метою цього дослідження була розробка інноваційного методу динамічного оцінювання довіри в архітектурі Zero Trust, який ефективно поєднує високу точність автоматизованого виявлення мережевих аномалій із прозорістю прийняття рішень. Для розрахунку безперервного показника оцінки довіри на базі змодельованого набору даних корпоративного мережевого трафіку було застосовано ансамблевий алгоритм машинного навчання Extreme Gradient Boosting, а для пояснення згенерованих рішень – метод адитивних пояснень SHapley Additive exPlanations (SHAP). Експериментальна перевірка продемонструвала виняткову ефективність запропонованого механізму політик (Policy Engine), який досяг показника F1-score 1,00 на тестовій вибірці. Модель успішно розрізнила легітимні та аномальні запити з нульовим рівнем хибнопозитивних спрацювань, ідентифікуючи такі кібератаки, як ескалація привілеїв та доступ з нетипових локацій. Глобальний аналіз важливості ознак за допомогою фреймворку SHAP підтвердив, що тип мережевого підключення та стан безпеки пристрою є найбільш значущими предикторами ризику, що повністю узгоджується з базовими принципами ZTA. Крім того, локальний аналіз довів здатність системи миттєво генерувати детальні, зрозумілі людині текстові пояснення для кожної відмови у доступі, вказуючи конкретну причину блокування. Завдяки такій деталізації аналітики отримують можливість безпосередньо розуміти логіку спрацювання автоматизованих систем захисту без необхідності тривалого ручного корелювання розрізнених журналів подій. Практична цінність дослідження полягає у створенні прозорого та адаптивного інструменту, який може бути інтегрований у сучасні центри операцій безпеки для суттєвого зниження «втоми від сповіщень» та мінімізації середнього часу вирішення інцидентів

Ключові слова: кібербезпека; машинне навчання; SHAP; XGBoost; виявлення аномалій; адаптивний захист