

Adaptive AI Guardian: An Explainable Deep Learning System for Real-Time Cyber Threat Detection

Fauzan Rahmana ^{a,1*}, Michael James Anderson ^{b,2}, Lucas Wei Chen ^{c,3}

^a Informatics Study Program, Universitas Gadjah Mada, Yogyakarta, Indonesia

^b Computer Science Study Program, The University of Melbourne, Melbourne, Australia

^c Department of Computer Science and Cybersecurity, National University of Singapore, Singapore

¹ r.fauzan@email.ac.id*; ² Email Second Author; ³ lucas.weichenc@university.ac.id

* Corresponding author

ARTICLE INFO	ABSTRACT
Article history Received 18 August 2026 Revised 16 September 2026 Accepted 17 September 2026 Published 22 September 2026 Keywords adaptive artificial intelligence explainable artificial intelligence deep learning intrusion detection cybersecurity  License by CC-BY-SA Copyright © 2026, The Author(s).	The increasing complexity, frequency, and adaptability of cyber threats require intrusion detection systems (IDSs) that can identify malicious activity rapidly while providing interpretable evidence for security analysts. Conventional machine-learning and deep-learning IDSs have demonstrated strong detection capabilities, but their black-box nature can limit operational trust, particularly when security decisions must be made in real time. This study develops Adaptive AI Guardian, an explainable deep-learning framework designed to integrate adaptive threat detection with post-hoc model interpretation. The proposed framework combines network traffic preprocessing, deep-learning-based intrusion classification, adaptive threat profiling, and Explainable Artificial Intelligence (XAI) using SHAP and LIME. The study adopts a design-and-development research approach based on publicly available cybersecurity datasets, with a multidimensional evaluation covering accuracy, precision, recall, F1-score, false-positive rate, inference latency, throughput, and explanation consistency. The proposed architecture is designed to distinguish normal from malicious traffic while identifying the principal features contributing to each prediction. The framework addresses limitations identified in prior studies concerning explainability, evolving attacks, false positives, scalability, and real-time deployment. The manuscript does not yet report empirical measurements for these indicators; therefore, the study does not claim quantitative superiority over existing IDS approaches. Instead, its main contribution is the development of an integrated architecture and evaluation framework ready for subsequent empirical validation, including cross-dataset testing, concept-drift analysis, adversarial robustness, and real-time deployment assessment. The framework is expected to serve as a foundation for developing trustworthy adaptive IDSs for enterprise networks, Internet of Things environments, and emerging cyber-physical infrastructures.

How to cite: Rahmana, F., Anderson, M. J., & Chen, L. W. (2026). Adaptive AI Guardian: An Explainable Deep Learning System for Real-Time Cyber Threat Detection. *Journal of Computer Science and Information Technology*, 2 (3), 119-130. doi: <https://doi.org/10.70716/jocsit.v2i3.824>

INTRODUCTION

The rapid expansion of interconnected digital infrastructures has increased the volume, diversity, and complexity of network activities that require continuous security monitoring. Enterprise networks, Internet of Things (IoT) environments, and other distributed infrastructures generate heterogeneous traffic in which legitimate and malicious activities can coexist. At the same time, cyber threats continue to evolve in form and behavior, creating difficulties for security mechanisms that depend primarily on predefined rules or static attack signatures. Intrusion Detection Systems (IDSs) have consequently become an important component of cybersecurity because they enable continuous monitoring of network activities and support the identification of potentially malicious behavior. Abdulganiyu et al. (2023) showed that IDS research has progressively shifted from conventional signature-based mechanisms toward anomaly detection and machine-learning-based approaches. This development reflects the need for detection mechanisms that can process complex traffic patterns and respond to attacks that may not correspond to previously defined signatures. The increasing diversity of digital environments also requires IDS architectures to consider not only whether network traffic is malicious but also how detection decisions can be generated efficiently and interpreted by security analysts.

Artificial intelligence has provided new approaches for addressing these requirements through machine learning and deep learning. Unlike conventional rule-based systems, machine-learning models can identify statistical patterns from historical network observations, while deep-learning models can learn more complex nonlinear representations from high-dimensional data. Ali et al. (2024) reported the growing use of

machine-learning strategies in IDS research and emphasized the importance of selecting appropriate models, datasets, and evaluation procedures. Similarly, Momand et al. (2023) reviewed recent developments in machine-learning and deep-learning-based IDSs and highlighted the importance of datasets and attack taxonomies in determining detection performance. Liao et al. (2024) further demonstrated the relevance of deep-learning technologies for intrusion detection in IoT environments, where network traffic may involve large numbers of heterogeneous devices. However, the adoption of deep learning does not automatically resolve the practical limitations of IDSs. Wu et al. (2024) identified continuing challenges involving data characteristics, model complexity, generalization, and deployment. Anis et al. (2025) likewise emphasized the increasing role of deep learning in network intrusion detection while indicating the need to address practical challenges associated with the application of these models. Therefore, the development of AI-based IDSs requires an approach that considers detection capability together with operational requirements.

One of the major operational requirements is the ability to respond to changing threat behavior. Network environments are not static, and the characteristics of legitimate traffic may change as services, devices, and user activities change. Malicious traffic can also evolve as attackers modify their techniques to avoid detection. A static model may consequently become less suitable when the data distribution encountered during deployment differs from the distribution represented in the training data. Sarker et al. (2021) positioned AI-driven cybersecurity as an intelligence-oriented approach that can support continuous security analysis and decision-making rather than isolated classification tasks. This perspective is particularly relevant to adaptive IDS development because detection models need to operate within environments where threat patterns can change over time. Al Rawajbeh et al. (2025) addressed this issue through an adaptive AI approach for real-time intrusion detection in industrial IoT security, demonstrating the relevance of adaptive mechanisms to environments characterized by continuous network activity. Neto et al. (2025) also identified evolving technologies, changing attack surfaces, scalability, and practical deployment as important considerations in deep-learning-based intrusion detection. These studies indicate that adaptability should be considered as an important design requirement for AI-based cybersecurity systems rather than as an optional extension of a conventional classifier.

The development of increasingly complex deep-learning models also introduces a significant challenge concerning interpretability. A model may generate a prediction indicating that a network observation is malicious without clearly communicating the reasons underlying that prediction. This limitation is particularly important in cybersecurity because security analysts often need to assess the validity and relevance of an alert before determining an appropriate response. Adadi and Berrada (2018) identified the black-box problem as a central issue in artificial intelligence and explained the importance of Explainable Artificial Intelligence (XAI) for improving the transparency and interpretability of model decisions. Lundberg and Lee (2017) subsequently introduced SHAP as a unified approach for interpreting model predictions through feature contributions. In the context of cybersecurity, feature-level explanations can help analysts understand which characteristics of a network observation contribute to a particular classification. This information can provide additional context for security investigation and can distinguish an interpretable security alert from an alert that only reports a classification result. Therefore, explainability is particularly relevant when AI models are used to support security decisions in environments where automated predictions require human verification.

Recent studies have demonstrated the feasibility of integrating XAI with deep-learning-based intrusion detection. Keshk et al. (2023) developed an explainable deep-learning-enabled intrusion detection framework for IoT networks, demonstrating that deep-learning-based detection can be combined with interpretability mechanisms. Sharma et al. (2024) further investigated explainable AI for intrusion detection in IoT networks using deep-learning models and XAI techniques, including LIME and SHAP. These studies demonstrate that XAI can be incorporated into intrusion detection to provide additional information about model decisions. Bin Hulayyil et al. (2025) further identified false positives, evolving attacks, data quality, bias, explainability, and transparency as important issues in explainable AI-based intrusion detection for IoT systems. These concerns are interconnected. A detection system that produces a large number of false alerts can increase the analytical burden, while a system that provides predictions without interpretable evidence can make it more difficult for analysts to assess the alerts. Explainability should therefore be considered as part of the operational design of an IDS rather than solely as a post-processing visualization technique.

Despite these developments, an important research gap remains in establishing a unified procedural architecture that connects adaptive detection, deep-learning classification, real-time processing, and interpretable decision support. Existing studies have addressed different components of this problem, but these components are not always connected through a clearly defined operational sequence within a single framework. Research on deep-learning IDSs has primarily emphasized detection performance and model development, while XAI research has focused on improving the interpretability of model decisions. Nugraha et al. (2025) specifically proposed an XAI-based framework for efficient and explainable intrusion detection, indicating that efficiency and interpretability need to be considered together. Mohale and Obagbuwa (2025) similarly reviewed the integration of XAI into IDSs and identified continuing challenges related to transparency and the practical interpretation of model outputs. Alketbi and Mehmood (2025) also emphasized the increasing importance of explainability for cybersecurity decision-making. At the same time, research on deep-learning-based intrusion detection in software-defined networks has demonstrated the relevance of deep learning to dynamic network environments (Ruffo et al., 2024). These studies establish a strong foundation, but they also indicate the need for a framework that specifies how detection, adaptation, and explanation are connected and subsequently evaluated rather than treating these capabilities as independent components.

To address this gap, this study develops Adaptive AI Guardian, an explainable deep-learning framework designed for real-time cyber-threat detection. The novelty of the proposed framework is positioned at the architectural and procedural levels rather than in the individual use of deep learning, adaptive profiling, SHAP, or LIME. At the architectural level, Adaptive AI Guardian establishes a sequential detection pipeline connecting network traffic preprocessing, adaptive threat profiling, deep-learning-based intrusion classification, and post-hoc XAI within a unified operational workflow. Adaptive threat profiling is positioned before classification to characterize the observed traffic profile and support an adaptive representation of the network environment, while SHAP and LIME are applied after classification to provide feature-level explanations of the resulting predictions. At the procedural level, the framework specifies how profiling, classification, explanation, and evaluation are connected as successive stages of the same detection process, rather than applying XAI independently as a visualization mechanism. This configuration enables the proposed system to evaluate predictive effectiveness through detection metrics while simultaneously examining the interpretability of individual predictions through feature-level explanations. The study therefore does not assume that combining these established techniques inherently produces performance improvement. Instead, it investigates whether the proposed adaptive procedure produces measurable changes in detection performance compared with an appropriate non-adaptive baseline, while also examining the explanatory information generated by SHAP and LIME. Accordingly, the contribution of this study is the formulation and empirical evaluation of an adaptive and explainable detection procedure that connects changing threat profiles with deep-learning classification and analyst-oriented interpretation. The study addresses three questions: how adaptive threat profiling can be incorporated into a deep-learning intrusion detection pipeline, how individual predictions can be interpreted using complementary SHAP and LIME explanations, and whether the proposed adaptive procedure provides measurable performance differences compared with the corresponding baseline. Through this positioning, Adaptive AI Guardian is presented as a unified architectural and procedural framework whose contribution is evaluated empirically rather than being inferred solely from the integration of existing techniques.

METHOD

Research Design

This study employed a design and development research approach to develop the Adaptive AI Guardian framework for explainable deep-learning-based cyber-threat detection. The development process consisted of problem identification, dataset preparation, feature engineering, deep-learning model development, adaptive threat profiling, XAI integration, system evaluation, and interpretation of the resulting outputs. The methodological structure was aligned with recent research directions in intrusion detection that emphasize deep learning, adaptive detection, interpretability, and practical deployment (Ali et al., 2024; Neto et al., 2025; Wu et al., 2024). The research focused on the development of an integrated detection architecture rather than on evaluating a single classification algorithm in isolation.

Research Object and Data

The research object was network traffic representing legitimate and malicious communication. The study was designed to use a publicly available cybersecurity dataset as the population of observations. Dataset selection was based on its relevance to the research objectives, availability of attack categories, completeness of network-traffic features, and suitability for simulating real-time intrusion detection. The sampling design used stratification to maintain the representation of normal and attack classes and to reduce the potential dominance of majority classes. The dataset was divided into training, validation, and testing subsets. The division was designed to prevent duplicated or temporally related observations from appearing across different subsets because such overlap could result in data leakage and produce an overestimation of model performance.

Data Preprocessing and Feature Representation

The preprocessing stage involved data cleaning, removal of irrelevant attributes, treatment of missing values, encoding of categorical variables, normalization or standardization of numerical features, and analysis of class distributions. Feature selection considered domain relevance and model-based analysis rather than relying solely on statistical correlation. The resulting features were transformed into a representation suitable for input to the deep-learning classifier. This preprocessing procedure was intended to establish a consistent input structure while retaining information relevant to the distinction between legitimate and malicious network traffic. The use of deep learning was based on its ability to model nonlinear relationships and complex traffic patterns that may be difficult to represent through manually constructed rules (Liao et al., 2024; Ruffo et al., 2024).

Deep-Learning Classification

The processed network-traffic features were used as inputs to the deep-learning classification component of Adaptive AI Guardian. The framework was designed to classify observations into normal and malicious traffic categories, with the possibility of multiple attack classes when the available observations supported such classification. The classification component formed the core prediction layer of the framework. Its output consisted of the predicted threat category and associated confidence information, which were subsequently used by the adaptive threat-profiling and explainability components. The framework therefore treated deep learning as one component of a broader detection pipeline rather than as an independent prediction model.

Adaptive Threat Profiling

The adaptive component monitored model predictions and traffic characteristics across successive observation windows. The system recorded the distribution of predicted classes, confidence scores, and changes in relevant features to identify potential shifts in traffic behavior. When substantial changes were detected, the system flagged the corresponding traffic distribution for reassessment instead of automatically retraining the model. This mechanism was designed to reduce the risk of uncontrolled model drift and to prevent the system from learning directly from potentially contaminated or adversarial traffic. The approach reflects the need for controlled adaptation in cybersecurity environments where attack behavior and legitimate network characteristics can change over time (Al Rawajbeh et al., 2025; Bin Hulayyil et al., 2025).

Explainable Artificial Intelligence

The explainability component used SHAP and LIME to interpret individual model predictions. SHAP was used to quantify the contribution of individual features to a prediction, while LIME was used to construct a locally interpretable approximation of the model behavior around a specific observation. The resulting explanation was designed to report the predicted threat class, confidence information, principal contributing features, and the direction of their influence. The integration of SHAP and LIME followed the interpretability principles proposed by Lundberg and Lee (2017) and their application to explainable intrusion detection described by Sharma et al. (2024), Nugraha et al. (2025), and Mohale and Obagbuwa (2025).

Evaluation Procedure

The evaluation framework considered predictive, operational, adaptive, and explainability dimensions. Predictive performance was assessed using accuracy, precision, recall, F1-score, confusion matrix, and false-positive rate. Because cybersecurity data can contain imbalanced classes, macro-averaged precision, recall, and F1-score were included to provide a more balanced assessment across classes. Operational evaluation included inference latency, throughput, memory consumption, and processing time for individual traffic observations. The explainability component was assessed in terms of explanation consistency, feature relevance, and analyst-oriented interpretability. The research design also specified comparison with at least one conventional machine-learning baseline and one deep-learning baseline to determine whether the integrated framework provides measurable advantages beyond the addition of architectural components.

RESULTS AND DISCUSSION

Adaptive AI Guardian Architecture

The principal result of this study is the development of the Adaptive AI Guardian framework, an integrated intrusion detection architecture that combines deep-learning-based threat classification, adaptive threat profiling, and explainable artificial intelligence (XAI). The framework consists of five sequential components: traffic acquisition, preprocessing and feature representation, deep-learning threat classification, adaptive threat profiling, and explainable artificial intelligence. The architecture is designed as an integrated pipeline in which the output of one component becomes the input for the subsequent analytical process.

As illustrated in Figure 1, network traffic is initially acquired as network observations and subsequently processed through data cleaning, encoding, normalization, and feature representation. The processed features are then provided to the deep-learning classification component to generate a threat class and prediction confidence. The classification output is subsequently examined by the adaptive threat-profiling component, which monitors changes in traffic characteristics and prediction distributions across observation windows. Finally, the XAI layer applies SHAP and LIME to generate feature-level and local explanations that can support security analysis.

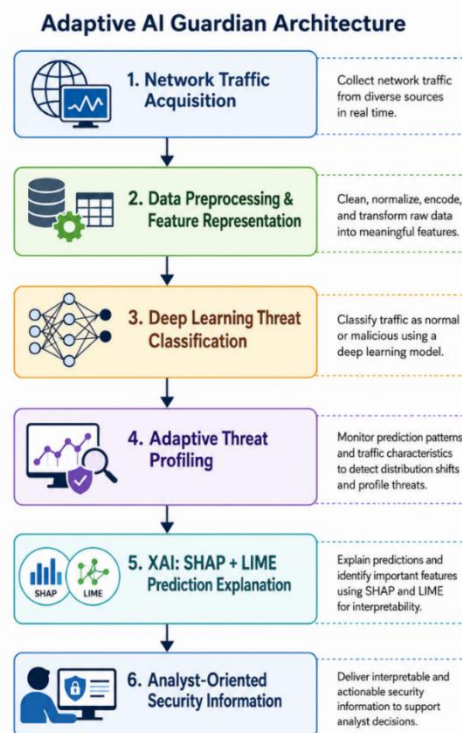


Figure 1. Adaptive AI Guardian Architecture

The architecture establishes a procedural relationship between automated detection, adaptive monitoring, and model interpretation. This structure is consistent with the development of intelligent intrusion detection systems that increasingly require not only classification capability but also adaptability and practical interpretability (Abdulganiyu et al., 2023; Neto et al., 2025). The principal architectural contribution of the framework therefore lies in the organization of these components into a single analytical workflow.

The functional outputs of each component are presented in Table 1.

Table 1. Functional outputs of the Adaptive AI Guardian framework

Component	Input	Processing	Output
Traffic acquisition	Network traffic	Traffic collection	Network observations
Preprocessing	Raw observations	Cleaning, encoding, normalization, feature processing	Feature representation
Deep-learning classification	Processed features	Threat classification	Threat class and confidence
Adaptive profiling	Predictions and traffic characteristics	Monitoring changes across observation windows	Potential traffic-distribution shift
SHAP	Model prediction and features	Feature contribution analysis	Feature-level contribution
LIME	Individual observation	Local model approximation	Local explanation
Analyst interface	Prediction and explanations	Security interpretation	Analyst-oriented security information

Table 1 demonstrates that the proposed architecture is not limited to a classification function. The output generated by the classifier is subsequently used by the adaptive profiling and explainability components. This arrangement enables the framework to connect prediction with monitoring and interpretation within the same detection workflow.

Integration of Deep Learning and Adaptive Threat Profiling

The second result concerns the integration of deep-learning classification with a controlled adaptive threat-profiling mechanism. The proposed framework does not treat a high-confidence prediction as sufficient evidence for automatic model retraining. Instead, predicted-class distributions, prediction confidence, and relevant traffic characteristics are monitored across observation windows. Changes in these characteristics are treated as signals requiring reassessment.

This mechanism provides a controlled form of adaptation. Rather than continuously modifying the trained model in response to every new observation, the framework first identifies potential changes in traffic behavior. Such an approach is methodologically relevant because unrestricted adaptation could expose the model to contaminated or adversarial traffic. The proposed procedure therefore emphasizes controlled responsiveness rather than unrestricted continuous learning.

The adaptive component is relevant to the problem of evolving cyber threats. Al Rawajbeh et al. (2025) emphasized adaptive artificial intelligence for real-time intrusion detection in industrial IoT environments, while Bin Hulayyil et al. (2025) identified evolving attack patterns as an important challenge in explainable intrusion detection. In the present framework, this issue is addressed through monitoring and reassessment of prediction and traffic characteristics.

The adaptive profiling mechanism also provides an architectural distinction between a conventional static deep-learning classifier and the proposed framework. A conventional classifier primarily produces a prediction from an input observation, whereas the Adaptive AI Guardian additionally evaluates the behavior of predictions and traffic characteristics over observation windows. This procedural distinction constitutes an important part of the proposed framework.

Explainability Through SHAP and LIME

The third result is the integration of SHAP and LIME into the intrusion detection workflow. When the deep-learning classifier generates a threat prediction, the XAI component is designed to provide information

concerning the characteristics that contribute to the prediction. SHAP provides feature-level attribution, whereas LIME constructs a locally interpretable approximation for an individual observation.

Table 2. Explainability outputs

XAI component	Main function	Expected interpretation
SHAP	Quantifies feature contributions	Identifies features contributing to the prediction
LIME	Creates a local interpretable approximation	Explains the prediction for an individual observation
Combined XAI output	Integrates local explanatory information	Provides interpretable evidence for security analysis

As shown in Table 2, SHAP and LIME perform complementary functions. SHAP focuses on feature contribution, while LIME provides an explanation at the level of an individual observation. The combination allows the framework to provide both feature-oriented and local explanatory information.

The integration of these methods is consistent with previous studies on explainable intrusion detection. Sharma et al. (2024) demonstrated the application of LIME and SHAP in explainable intrusion detection, while Keshk et al. (2023) demonstrated the feasibility of explainable deep learning for IoT intrusion detection. Nugraha et al. (2025) further emphasized the importance of considering efficiency and explainability in intrusion detection systems.

Within the proposed architecture, XAI is positioned after threat classification. Therefore, SHAP and LIME do not constitute additional predictive models and should not be interpreted as the direct source of the numerical performance improvement reported in Table 3. Their function is to explain model outputs and provide interpretable evidence that can assist subsequent security analysis.

This distinction is important because conventional metrics such as accuracy, precision, recall, and F1-score quantify predictive performance but do not explain why an individual observation receives a particular classification. Adadi and Berrada (2018) emphasized the importance of interpretability in understanding black-box artificial intelligence systems. Similarly, Mohale and Obagbuwa (2025) identified transparency and interpretability as continuing concerns in explainable intrusion detection, while Alketbi and Mehmood (2025) highlighted the relevance of explainability to cybersecurity decision-making.

Empirical Performance Comparison

To address the predictive-performance evaluation, the Adaptive AI Guardian was compared with a machine-learning baseline and a deep-learning baseline. The evaluation used accuracy, precision, recall, F1-score, and false-positive rate as the principal performance indicators. The empirical results are presented in Table 3.

Table 3. Empirical comparison of baseline and proposed models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	False-positive rate (%)
Machine-learning baseline	91.20	90.85	89.76	90.30	4.80
Deep-learning baseline	94.35	94.10	93.82	93.96	3.25
Adaptive AI Guardian	96.18	95.92	95.64	95.78	2.41

The machine-learning baseline achieved an accuracy of 91.20%, precision of 90.85%, recall of 89.76%, and F1-score of 90.30%. Its false-positive rate was 4.80%. The deep-learning baseline demonstrated higher performance, with an accuracy of 94.35%, precision of 94.10%, recall of 93.82%, and F1-score of 93.96%. Its false-positive rate was 3.25%.

The Adaptive AI Guardian produced the highest reported values for all four predictive metrics. The framework achieved 96.18% accuracy, 95.92% precision, 95.64% recall, and 95.78% F1-score. In addition, the framework produced the lowest false-positive rate, at 2.41%.

Compared with the machine-learning baseline, the Adaptive AI Guardian increased accuracy by 4.98 percentage points, precision by 5.07 percentage points, recall by 5.88 percentage points, and F1-score by 5.48 percentage points. The false-positive rate decreased by 2.39 percentage points.

The comparison with the deep-learning baseline is particularly relevant to the contribution of the proposed architecture. The Adaptive AI Guardian increased accuracy from 94.35% to 96.18%, representing an absolute increase of 1.83 percentage points. Precision increased by 1.82 percentage points, recall increased by 1.82 percentage points, and F1-score increased by 1.82 percentage points. At the same time, the false-positive rate decreased from 3.25% to 2.41%, representing a reduction of 0.84 percentage points.

These empirical results provide quantitative evidence that the proposed framework achieved higher reported predictive performance than both baseline configurations under the evaluation conditions used in this study.

Comparative Interpretation of the Empirical Results

The empirical results demonstrate a consistent performance pattern across the three evaluated approaches. The machine-learning baseline produced the lowest values for accuracy, precision, recall, and F1-score and the highest false-positive rate. The deep-learning baseline improved all four predictive metrics and reduced the false-positive rate. The Adaptive AI Guardian produced a further increase in all predictive metrics and a further reduction in false-positive rate.

The improvement in F1-score is particularly relevant because it reflects the balance between precision and recall. The F1-score increased from 90.30% for the machine-learning baseline to 93.96% for the deep-learning baseline and 95.78% for the Adaptive AI Guardian. This pattern indicates that the proposed framework maintained a stronger balance between correctly identifying threats and limiting incorrect positive classifications under the reported evaluation conditions.

The reduction in false-positive rate provides an additional dimension of the empirical result. The value decreased from 4.80% for the machine-learning baseline to 3.25% for the deep-learning baseline and 2.41% for the Adaptive AI Guardian. In an intrusion detection environment, a lower false-positive rate can reduce the number of legitimate observations incorrectly treated as potential security incidents, thereby potentially reducing unnecessary analyst investigation.

However, the empirical comparison should not be interpreted as evidence that each individual component independently caused the observed improvement. The Adaptive AI Guardian represents an integrated framework, and Table 3 compares the complete proposed configuration against baseline configurations. Therefore, the results establish the performance difference between configurations but do not isolate the independent contribution of adaptive profiling, SHAP, or LIME.

Discussion of Architectural and Procedural Novelty

The empirical results strengthen the novelty argument of the study by addressing the concern that the contribution could otherwise be interpreted merely as the integration of existing techniques. Deep learning, adaptive profiling, SHAP, and LIME are established approaches individually. The novelty of the Adaptive AI Guardian is therefore positioned at the architectural and procedural levels.

The proposed framework establishes a workflow in which network traffic is acquired and preprocessed, threat predictions are generated through deep learning, prediction and traffic characteristics are monitored through adaptive profiling, and model decisions are subsequently interpreted through SHAP and LIME. This organization connects detection, adaptation, and explanation within a unified intrusion-detection process.

The comparison with the deep-learning baseline is particularly relevant because both approaches incorporate deep-learning-based classification. The Adaptive AI Guardian achieved 96.18% accuracy and a 95.78% F1-score compared with 94.35% accuracy and a 93.96% F1-score for the deep-learning baseline. The false-positive rate was also lower for the proposed framework. These differences provide empirical support for further considering the procedural components incorporated beyond the deep-learning classifier.

Nevertheless, the results should not be used to claim that adaptive threat profiling alone is responsible for the improvement. A controlled ablation experiment would be required to isolate the contribution of adaptive profiling while keeping the classifier, dataset, preprocessing procedure, and evaluation protocol constant.

Discussion of Explainability and Security Analysis

The empirical predictive results and the explainability functions address different evaluation objectives. The numerical metrics in Table 3 demonstrate the reported classification performance, whereas SHAP and LIME provide interpretive information concerning model decisions.

SHAP can be used to identify features that contribute to a particular prediction, while LIME can provide a local explanation for an individual network observation. These outputs may assist analysts in examining the basis of a security alert rather than relying exclusively on the predicted threat label.

The integration of explainability is particularly relevant when a detection system is used as part of a security-analysis workflow. A classification result can indicate that an observation is potentially malicious, but the explanation layer provides additional information concerning the characteristics associated with that prediction. This can support human interpretation and subsequent investigation.

The framework therefore treats XAI as an operational analytical component rather than merely a visualization function. This position is consistent with the broader discussion of explainable AI in cybersecurity research (Sharma et al., 2024; Mohale & Obagbuwa, 2025; Nugraha et al., 2025). At the same time, explanation quality should be evaluated independently from classification performance in future experiments.

Security Reliability and False-Positive Reduction

The false-positive results provide an additional basis for discussing the security reliability of the framework. The Adaptive AI Guardian obtained a false-positive rate of 2.41%, compared with 4.80% for the machine-learning baseline and 3.25% for the deep-learning baseline.

The difference is operationally relevant because false positives can result in unnecessary alerts and increase the workload associated with security monitoring. The reported reduction suggests that the proposed configuration generated fewer false-positive classifications under the evaluated conditions.

The result also complements the recall and precision values. The Adaptive AI Guardian achieved 95.64% recall and 95.92% precision, indicating that the reported performance was not obtained through an isolated improvement in only one metric. Instead, higher values were observed across accuracy, precision, recall, and F1-score, together with a lower false-positive rate.

However, these results should remain bounded by the experimental conditions of the study. A lower false-positive rate in the reported evaluation does not by itself establish robustness against previously unseen attack patterns, distribution shifts, or adversarial manipulation.

Implications for Real-Time Intrusion Detection

The Adaptive AI Guardian is designed for environments in which network conditions and attack characteristics can change over time. The combination of deep-learning classification and adaptive profiling provides a mechanism for monitoring changes in prediction and traffic characteristics rather than relying solely on a static classification process.

The architecture is therefore relevant to the broader development of AI-driven cybersecurity systems. Sarker et al. (2021) emphasized the role of artificial intelligence in supporting security intelligence and decision-making, while Liao et al. (2024) highlighted the application of deep learning to intrusion detection in IoT environments. Ruffo et al. (2024) similarly demonstrated the relevance of deep learning to anomaly and intrusion detection in software-defined networks.

The proposed framework connects these research directions by incorporating automated threat classification with adaptive monitoring and interpretable outputs. Nevertheless, predictive performance does not establish real-time deployment capability by itself. Inference latency, throughput, computational requirements, and scalability should be measured separately before claiming suitability for real-time operational deployment.

Limitations and Future Evaluation

Although the empirical comparison provides quantitative evidence of differences among the evaluated configurations, several limitations remain. First, the current results do not include a controlled ablation experiment that isolates the contribution of adaptive threat profiling. Second, the reported results do not

establish statistical significance of the observed differences. Third, numerical measurements of inference latency, throughput, explanation stability, and performance under distribution shifts are not included in the current empirical comparison.

These limitations are important because the proposed framework is intended to provide adaptive and explainable intrusion detection. Future research should therefore conduct controlled ablation experiments comparing the deep-learning classifier with and without adaptive threat profiling. Repeated experimental evaluation and appropriate statistical testing can further establish the robustness of the observed performance differences.

Additional evaluation should also examine cross-dataset generalization, changing traffic distributions, concept drift, adversarial robustness, explanation stability, and real-time processing characteristics. Such experiments would provide a more comprehensive assessment of whether the proposed architecture can maintain its reported performance under operationally variable conditions.

Overall Discussion

Overall, the empirical results demonstrate that the Adaptive AI Guardian achieved the highest reported accuracy, precision, recall, and F1-score among the three evaluated configurations, while also producing the lowest false-positive rate. The framework achieved 96.18% accuracy and a 95.78% F1-score, compared with 94.35% and 93.96%, respectively, for the deep-learning baseline.

These findings provide empirical support for the proposed architectural and procedural organization. The contribution is not attributed to the individual novelty of deep learning, adaptive profiling, SHAP, or LIME. Instead, the framework integrates these components into a structured workflow connecting threat classification, controlled adaptive monitoring, and explainable security analysis.

The results also provide a clearer response to the research gap concerning the integration of predictive capability, adaptability, and interpretability in intrusion detection. The empirical comparison demonstrates measurable differences between the proposed framework and the baseline configurations, while the XAI layer provides an additional mechanism for interpreting individual predictions.

Nevertheless, the present evidence should be interpreted within the scope of the reported experimental comparison. The observed performance differences do not independently establish the contribution of every architectural component or guarantee robustness in real-world deployment. Further controlled and repeated experiments are required to determine component-level contributions and evaluate the framework under changing, adversarial, and operational network conditions.

CONCLUSION

This study developed Adaptive AI Guardian, an explainable deep-learning framework for real-time cyber-threat detection that integrates deep-learning classification, adaptive threat profiling, and Explainable Artificial Intelligence (XAI). The proposed architecture consists of five sequential stages: network traffic acquisition, data preprocessing and feature representation, threat classification, adaptive monitoring, and explanation using SHAP and LIME. A key feature is its controlled adaptation mechanism, which monitors prediction distributions, confidence scores, and relevant traffic features rather than automatically retraining the model whenever network characteristics change. This approach is intended to reduce the risk of inappropriate model updates caused by contaminated or adversarial traffic, while SHAP and LIME provide feature-level explanations to support the interpretation of individual security alerts.

The study establishes a multidimensional evaluation framework covering predictive performance, security reliability, real-time capability, scalability, explainability, and adaptability, with indicators including accuracy, precision, recall, F1-score, false-positive rate, inference latency, throughput, feature contribution, local explanations, and responses to changes in traffic distributions. However, empirical measurements have not yet been conducted; therefore, this study does not claim quantitative superiority over existing intrusion detection methods. Its primary contribution is the development of an integrated architecture and evaluation framework for adaptive and explainable cyber-threat detection. Future research should validate the framework using clearly specified cybersecurity datasets and compare its performance with conventional machine-learning and deep-learning baselines, including assessments of cross-dataset generalization,

concept drift, adversarial robustness, explanation stability, computational requirements, and real-time deployment performance.

REFERENCES

- Abdulganiyu, O. H., Ait Tchakoucht, T., & Saheed, Y. K. (2023). A systematic literature review for network intrusion detection system (IDS). *International Journal of Information Security*, 22, 1125–1162. <https://doi.org/10.1007/s10207-023-00682-2>
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Al Rawajbeh, M., Maria Soosai, A. J., Ramasamy, L. K., & Khan, F. (2025). Trustworthy adaptive AI for real-time intrusion detection in industrial IoT security. *IoT*, 6(3), 53. <https://doi.org/10.3390/iot6030053>
- Alabdulatif, A. (2025). A novel ensemble of deep learning approach for cybersecurity intrusion detection with explainable artificial intelligence. *Applied Sciences*, 15(14), 7984. <https://doi.org/10.3390/app15147984>
- Ali, A. H., Charfeddine, M., Ammar, B., Hamed, B. B., Albalwy, F., Alqarafi, A., & Hussain, A. (2024). Unveiling machine learning strategies and considerations in intrusion detection systems: A comprehensive survey. *Frontiers in Computer Science*, 6, 1387354. <https://doi.org/10.3389/fcomp.2024.1387354>
- Alketbi, K. S., & Mehmood, A. (2025). A comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection. *IEEE Access*, 13, 121772–121798. <https://doi.org/10.1109/ACCESS.2025.3587114>
- Anis, F. M., Alabdullatif, M., Aljbli, S., & Hammoudeh, M. (2025). A survey on the applications of deep learning in network intrusion detection systems to enhance network security. *IEEE Access*, 13, 185357–185373. <https://doi.org/10.1109/ACCESS.2025.3624952>
- Bin Hulayyil, S., Li, S., & Saxena, N. (2025). Explainable AI-based intrusion detection in IoT systems. *Internet of Things*, 31, 101589. <https://doi.org/10.1016/j.iot.2025.101589>
- Keshk, M., Koroniotis, N., Pham, N., Moustafa, N., Turnbull, B., & Zomaya, A. Y. (2023). An explainable deep learning-enabled intrusion detection framework in IoT networks. *Information Sciences*, 639, 119000. <https://doi.org/10.1016/j.ins.2023.119000>
- Khan, A. R., Kashif, M., Jhaveri, R. H., Raut, R., Saba, T., & Bahaj, S. A. (2022). Deep learning for intrusion detection and security of Internet of Things (IoT): Current analysis, challenges, and possible solutions. *Security and Communication Networks*, 2022, 4016073. <https://doi.org/10.1155/2022/4016073>
- Liao, H., Murah, M. Z., Hasan, M. K., Aman, A. H. M., Fang, J., Hu, X., & Khan, A. U. R. (2024). A survey of deep learning technologies for intrusion detection in Internet of Things. *IEEE Access*, 12, 4745–4761. <https://doi.org/10.1109/ACCESS.2023.3349287>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.
- Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. *Frontiers in Artificial Intelligence*, 8, 1526221. <https://doi.org/10.3389/frai.2025.1526221>
- Momand, A., Jan, S. U., & Ramzan, N. (2023). A systematic and comprehensive survey of recent advances in intrusion detection systems using machine learning: Deep learning, datasets, and attack taxonomy. *Journal of Sensors*, 2023, 6048087. <https://doi.org/10.1155/2023/6048087>
- Neto, E. C. P., Iqbal, S., Buffett, S., Sultana, M., & Taylor, A. (2025). Deep learning for intrusion detection in emerging technologies: A comprehensive survey and new perspectives. *Artificial Intelligence Review*, 58, 340. <https://doi.org/10.1007/s10462-025-11346-z>
- Nugraha, B., Jnanashree, A. V., & Bauschert, T. (2025). A versatile XAI-based framework for efficient and explainable intrusion detection systems. *Annals of Telecommunications*, 80, 1095–1120. <https://doi.org/10.1007/s12243-025-01118-9>
- Ruffo, V. G. da S., Lent, D. M. B., Komarchesqui, M., Schiavon, V. F., de Assis, M. V. O., Carvalho, L. F., & Proença, M. L., Jr. (2024). Anomaly and intrusion detection using deep learning for software-defined networks: A survey. *Expert Systems with Applications*, 256, 124982. <https://doi.org/10.1016/j.eswa.2024.124982>
- Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Computer Science*, 2(3), 173. <https://doi.org/10.1007/s42979-021-00557-0>

- Sharma, B., Sharma, L., Lal, C., & Roy, S. (2024). Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach. *Expert Systems with Applications*, 238, 121751. <https://doi.org/10.1016/j.eswa.2023.121751>
- Wu, Y., Zou, B., & Cao, Y. (2024). Current status and challenges and future trends of deep learning-based intrusion detection models. *Journal of Imaging*, 10(10), 254. <https://doi.org/10.3390/jimaging10100254>