International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ENHANCING WAZUH-BASED INTRUSION DETECTION WITH MACHINE LEARNING, DEEP LEARNING, AND LARGE LANGUAGE MODELS

Muhammad Claudio Delvin[1,*], Rojali

[1,2]Computer Science, Binus University, Jakarta, Indonesia

*Corresponding email: muhammad.delvin@binus.ac.id

Abstract

Cybersecurity threats are increasing in complexity and frequency, requiring intelligent intrusion detection systems (IDS) capable of real-time anomaly detection and interpretation. Security Information and Event Management (SIEM) platforms, such as Wazuh, offer robust log collection and monitoring capabilities but rely heavily on static, rule-based detection, making them less effective against evolving attack vectors. This study proposes an integrated anomaly detection and interpretation framework that combines machine learning (ML), deep learning (DL), and large language models (LLMs) to address these limitations. Four predictive models Random Forest, XGBoost, LSTM, and a hybrid RF + LSTM were developed and evaluated using Wazuh agent logs containing both normal and anomalous events. Quantitative evaluation metrics included accuracy, precision, recall, and F1-score, while qualitative metrics assessed semantic relevance, inference latency, and operational usefulness. Results showed that Random Forest achieved the highest accuracy for structured data, LSTM excelled in capturing temporal dependencies, and the hybrid RF + LSTM model provided the most balanced performance. To enhance interpretability, four LLMs—GPT-3.5-Turbo, GPT-4, GPT-4-Turbo, and GPT-4o were integrated to generate contextual, actionable explanations of detected anomalies. GPT-4o emerged as the optimal choice, offering high-quality interpretations with minimal latency, enabling near real-time decision support in Security Operations Centers (SOCs). This study demonstrates the practical and theoretical value of combining ML, DL, and LLM technologies for SIEM log analysis, delivering both improved detection accuracy and operationally relevant insights. The proposed framework offers a deployable blueprint for enhancing the responsiveness and efficiency of modern SOC environments.

Keywords: Wazuh; Intrusion Detection System; Machine Learning; Deep Learning; Large Language Models

1 Introduction

Cybersecurity has become a pivotal concern in the digital era, particularly as cyberattacks grow in both complexity and frequency. According to the Verizon, 83% of breaches involved human-related factors, while 32% involved advanced persistent threats (APTs) targeting system vulnerabilities through stealthy, multi-stage operations [1]. In Indonesia, the National Cyber and Crypto Agency (BSSN) reported over 403 million cybersecurity incidents in 2023, with a significant portion targeting endpoint infrastructure and system logs [2]. These developments underscore the increasing urgency of intelligent intrusion detection systems (IDS), particularly

Received: August 02, 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

those that can analyze host-based and time-series data to detect anomalous behaviors in real-time.

Security Information and Event Management (SIEM) platforms are central to enterprise security operations, providing log aggregation, correlation, and alerting mechanisms. Among open-source SIEM platforms, Wazuh stands out due to its high configurability, rule-based detection engine, and real-time log monitoring capabilities. Its modular architecture allows distributed deployment across diverse network environments, and it supports various data sources, from operating system logs to application telemetry. According to Wazuh Community Analytics “However, despite its utility, Wazuh’s reliance on static rule sets and signature-based detection often renders it ineffective against unknown or evolving attack vectors” [3].

The increasing volume and heterogeneity of log data pose significant challenges to manual log analysis. Security analysts face alert fatigue, high false positive rates, and delayed incident response times due to the limitations of static detection mechanisms. While some efforts have integrated basic machine learning (ML) into Wazuh, these models are generally limited to threshold-based anomaly detection or simple clustering, which lack contextual awareness. This highlights a critical gap: the need for an adaptive, intelligent detection mechanism that can dynamically learn from patterns in log data and produce actionable, interpretable insights.

A 2025 study specifically evaluated Wazuh's ability to detect APT attacks through real-time log analysis. This study confirmed Wazuh's effectiveness in detecting attack patterns such as brute force attacks and unauthorized directory access, and highlighted the potential for machine learning integration to improve automated and predictive detection in the future. However, this study did not directly implement ML/DL in Wazuh, but instead recommended further exploration in this direction [4]. Various surveys and comparative studies have shown that ML and DL have improved the accuracy of intrusion and anomaly detection in IDS systems in general, with deep learning models such as DNNs, CNNs, and autoencoders demonstrating high performance on various IDS datasets [5]–[10].

Several studies have developed hybrid models integrating LSTM and Random Forest for intrusion detection. Typically, LSTM is used to extract sequence features from network data, and the results are then classified by Random Forest to distinguish between normal activity and attacks and identify specific attack types. These models were tested on popular IDS datasets such as NSL-KDD, UNSW-NB-15, and CSE-CIC-IDS2018, achieving very high accuracy (up to 99.8%) and outperforming single models. Although not explicitly mentioned for direct integration into specific SIEM platforms, the RF-LSTM hybrid model is highly relevant for adoption in SIEM environments due to its ability to analyze large amounts of log data and detect anomalies in real-time. 123 These studies emphasize the importance of hybrid models for improving the detection of attacks hidden among normal traffic, a key challenge in SIEM systems [11]–[13].

Recent literature highlights both the promise and the limitations of ML approaches in cybersecurity. According to Venturi while ML techniques such as ANN, SVM, fuzzy logic, and evolutionary computation have shown success in prior studies, they often underperform when confronted with high data complexity or sequential dependencies [14]. Sudyana reinforces this point by highlighting the generalizability of ML-based IDSs to detect previously

Received: August 02, 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

unseen attacks, yet acknowledges the need for more robust temporal modeling to enhance detection accuracy [15].

As a response to ML limitations, deep learning has gained traction in the IDS domain. Shone introduced an unsupervised deep autoencoder model that significantly improved detection rates on benchmark datasets, showcasing the capability of DL to learn hidden representations of attacks [16]. Similarly, Vinayakumar developed a scalable deep neural network (DNN) framework that outperformed classical ML classifiers in detecting dynamic, evolving threats using diverse datasets such as NSL-KDD and CICIDS2017 [17].

Despite these promising results, many DL-based systems still lack interpretability making them less suitable for real-time decision-making in Security Operation Centers (SOCs). Classical explainable AI (XAI) techniques such as SHAP and LIME offer feature attribution but fail to provide contextual narratives or mitigation insights. This gap opens a promising path for the adoption of Large Language Models (LLMs) such as GPT-4 and GPT-4o, which have demonstrated potential in converting structured log data into semantically rich explanations and actionable recommendations.

However, most existing studies separate detection and interpretation processes, focusing solely on improving accuracy while neglecting practical usability, latency, and operational integration key aspects required in high-stakes environments. Moreover, there is a lack of comprehensive research that evaluates the combined use of ML, DL, and LLMs within a unified framework for SIEM log analysis, particularly in the context of Wazuh agent logs.

To address these gaps, this study proposes an integrated anomaly detection and interpretation framework that leverages four predictive models Random Forest, XGBoost, LSTM, and a hybrid Random Forest–LSTM architecture alongside four state-of-the-art LLMs for interpreting Wazuh agent logs. The framework is evaluated comprehensively using both quantitative metrics (accuracy, precision, recall, and F1-score) and qualitative metrics (semantic relevance, inference latency, and operational usefulness of generated explanations). By combining machine learning, deep learning, and large language model capabilities within a single, unified system, this research delivers a novel contribution to SIEM log analysis: the first empirical study, to the best of our knowledge, that jointly optimizes anomaly detection accuracy and interpretability through an ML–DL–LLM architecture specifically applied to Wazuh-based environments.

2 Research Method

To address the limitations of static rule-based detection in SIEM platforms and the lack of explainability in conventional machine learning pipelines, this study adopts a multi-phase methodological framework that integrates machine learning (ML) [4], deep learning (DL), and large language models (LLMs). The approach is designed to simultaneously optimize detection performance and interpretability for Wazuh agent logs. The methodology comprises six key stages: data collection, data preprocessing and feature engineering, model development, model evaluation, LLM-based log interpretation, and experimental implementation.

Data Collection

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Log data were collected from Wazuh agents deployed across multiple Linux-based systems configured with a variety of monitoring modules, including file integrity checking, authentication auditing, and system command monitoring [18]. The logs were extracted primarily from Wazuh log data, containing records of authentication attempts, privilege escalations, system warnings, and network activity. Two distinct sets of log data representing both normal and anomalous scenarios were merged to create a comprehensive dataset that simulates real-world security threats and operational conditions [19].

Figure 1. Data Collection Workflow

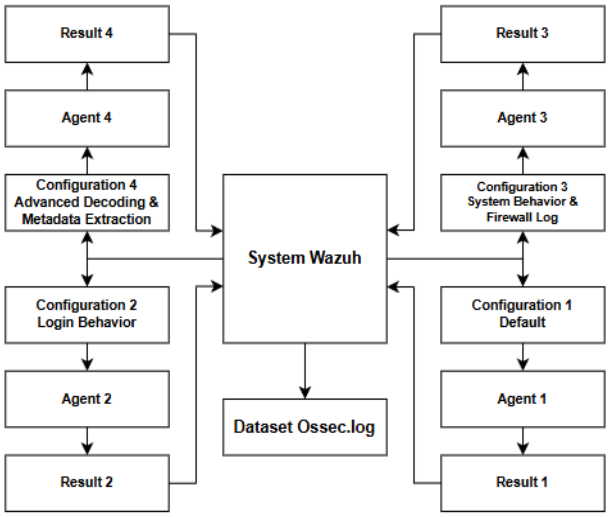

Figure 1 presents the workflow for collecting Wazuh log data used in this study. The process begins with multiple data sources, including various system agents deployed across Linux-based environments, each configured with monitoring capabilities such as authentication auditing, privilege escalation tracking, system command monitoring, and file integrity verification. Log events from these agents are aggregated by the Wazuh system, forming a centralized dataset that records diverse security-related activities, such as login attempts, system warnings, and network traffic anomalies. The dataset integrates logs from two primary sources: normal operational activities and artificially simulated attack scenarios. This combination ensures the resulting dataset captures a wide spectrum of benign and malicious events, enabling robust training and evaluation of anomaly detection models in realistic operational contexts.

To ensure robust model generalization, two types of data were curated: logs from routine operations and logs from artificially simulated attack scenarios. This hybrid dataset simulates realistic operational environments while capturing varied patterns of both benign and malicious behaviors. All data were anonymized prior to processing to preserve confidentiality [20].

Data Preprocessing and Feature Engineering

Raw log entries were cleaned, parsed, and transformed into structured formats using regular expression-based parsing and log normalization techniques. From each entry, attributes such as timestamp, rule ID, alert level, agent name, action type, source IP, and message text were

Received: August 02, 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

extracted. For classical ML models (Random Forest, XGBoost), categorical variables were encoded via label encoding, and numerical attributes were normalized to reduce scale variance. For DL-based models, particularly LSTM and hybrid configurations, textual log messages were tokenized, padded into equal-length sequences, and embedded using trainable word embeddings to capture semantic representations. Labels were heuristically assigned based on rule ID classes and alert severity levels, following predefined thresholds adapted from the alert taxonomy. Logs indicating brute-force login, unauthorized access, or privilege misuse were labeled as anomalies. This dual preprocessing path ensured that both tabular and sequential features could be effectively utilized depending on the architectural demands of the models applied [21].

Model Development

Four models were developed to comparatively evaluate detection capabilities:

Random Forest (RF): A non-sequential ensemble classifier optimized for structured tabular input. Configured with 100 decision trees and Gini impurity, the model prioritizes interpretability and computational efficiency in processing discrete features.

Extreme Gradient Boosting (XGBoost): A boosting-based ensemble classifier renowned for its accuracy in imbalanced datasets. Hyperparameters were optimized through randomized search and early stopping to mitigate overfitting.

Long Short-Term Memory (LSTM): A recurrent neural network architecture suited for sequence modeling. The LSTM model consisted of an embedding layer, one bidirectional LSTM layer, and fully connected output layers. It was trained using the Adam optimizer and binary cross-entropy loss.

RF + LSTM Hybrid: This two-stage architecture first uses RF to generate intermediate probability scores from tabular features, which are then concatenated with LSTM inputs derived from tokenized log messages. The fused features are processed jointly in the LSTM network, enabling multimodal learning from both structured and sequential inputs.

All models were trained on 80% of the dataset and tested on the remaining 20%. Sampling techniques SMOTE and random undersampling were applied selectively to correct for class imbalance, ensuring fair evaluation of precision- and recall-sensitive metrics [12].

Figure 2. Hybrid Model Architecture

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

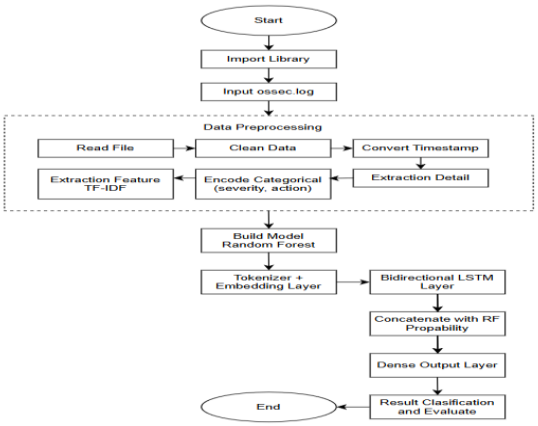

Figure 2 illustrates the architecture of the proposed hybrid Random Forest (RF) and Long Short-Term Memory (LSTM) model. The process begins with the initialization and loading of required libraries, followed by the ingestion of Wazuh log data. The data preprocessing stage involves reading and cleaning the raw log entries, converting timestamps into a standardized format, extracting detailed attributes, encoding categorical variables such as severity and action types, and performing TF-IDF feature extraction. The Random Forest model is first trained on structured tabular features to generate intermediate probability scores. In parallel, tokenized and embedded log message sequences are processed through a bidirectional LSTM layer to capture temporal dependencies. The outputs from the LSTM are concatenated with the RF probability scores, enabling multimodal learning from both structured and sequential inputs. Finally, the combined features are passed through a dense output layer to produce anomaly classifications, which are then evaluated using standard performance metrics. This architecture leverages the complementary strengths of RF for tabular feature learning and LSTM for sequential pattern recognition, resulting in enhanced detection accuracy and generalization capability.

Model Evaluation

Model performance was quantitatively assessed using standard metrics: accuracy, precision, recall, F1-score, and confusion matrix. Additionally, training convergence and generalization capability were visualized through learning curves for both accuracy and loss over epochs.

Among the baseline models, XGBoost demonstrated strong precision and recall in the presence of class imbalance. LSTM, despite lower accuracy, was more effective in capturing complex temporal correlations in logs. Notably, the RF + LSTM hybrid outperformed both individual models, achieving higher F1-scores by leveraging the complementary strengths of tabular and sequence learning. This justifies the integration of cross-domain features in real-world anomaly detection scenarios.

Learning curve analysis (Figure 4, not shown) revealed that the hybrid model converged efficiently and demonstrated minimal overfitting, further confirming its robustness.

LLM-Based Log Interpretation

Received: August 02, 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Most existing anomaly detection pipelines stop at classification and offer no operational insight into detected threats. To address this, we integrated LLMs to semantically interpret anomalous log entries and generate contextual recommendations. Four OpenAI models were evaluated: GPT-3.5-Turbo, GPT-4, GPT-4-Turbo, and GPT-4o [15].

Each model received an anomaly log prompt and generated a narrative explanation, including likely attack cause, implications, and mitigation strategies. Evaluation criteria included:

- Relevance of generated summary to log content
- Clarity and actionability of mitigation instructions
- Inference time, measured in seconds
- Cost-efficiency, estimated from API usage

Among these, GPT-4o offered the most optimal trade-off between latency and interpretive quality, producing high-relevance summaries with detailed response steps within real-time constraints. This step moves beyond classical "black box" ML and demonstrates the feasibility of combining AI explainability with security analytics.

Experimental Implementation

The entire pipeline was implemented using Python 3.11, with Scikit-learn for RF and XGBoost, and TensorFlow/Keras for LSTM and hybrid models. Experiments were conducted on a local workstation equipped with NVIDIA RTX 4060 GPU (16GB VRAM) and 32GB RAM. LLM inference was accessed via OpenAI's official API using anonymized log prompts.

To ensure reproducibility and scalability, all models and workflows were containerized using Docker and tracked with Git version control. Ethical compliance was maintained through log anonymization and local sandboxing of sensitive data.

3 Results

This section presents the findings obtained from implementing the six-stage methodological pipeline described in the previous section, which integrates machine learning (ML), deep learning (DL), and large language models (LLMs) for Wazuh log anomaly detection and interpretation. The results are organized into two main parts: (1) evaluation of the predictive performance of ML and DL models, and (2) assessment of interpretability and operational relevance through LLM-based log analysis [4], [19], [20].

Performance of Machine Learning and Deep Learning Models

Each model was trained and evaluated using a labeled dataset derived from Wazuh's Wazuh log data. The dataset was preprocessed using regular expressions and feature engineering tailored to the needs of each model, such as one-hot encoding for RF/XGBoost and tokenized sequences for LSTM.

Tabel 1 Performance Comparison of Anomaly Detection Models

Model	Accuracy	Precision	Recall	F1-Score
RF	96%	0.93	0.96	0.95

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

LSTM	67%	0.60	0.70	0.64
XGBoost	94%	0.91	0.94	0.92
RF + LSTM	92%	0.89	0.93	0.91

The Random Forest model achieved the highest overall accuracy (96%), demonstrating strong performance on structured, tabular features. However, its inability to model temporal dependencies limited its effectiveness in detecting time-sensitive anomalies. The LSTM model, although producing a lower overall accuracy (67%), excelled in identifying complex sequential patterns that static models often missed, making it suitable for detecting multi-stage attacks.
XGBoost offered a balanced profile, with high F1-score (0.92) and robustness against class imbalance. The hybrid RF + LSTM model achieved a competitive accuracy of 92% while providing superior recall compared to RF alone, illustrating the benefit of combining RF's structured feature learning with LSTM's temporal sequence modeling. This hybrid synergy is particularly valuable in real-world SIEM deployments, where anomalies often exhibit both structured and time-dependent characteristics.
The confusion matrix for the hybrid RF + LSTM model (Figure 3) confirms its ability to minimize both false positives and false negatives, while its learning curves (Figure 4) show stable convergence with no signs of overfitting, validating its suitability for operational deployment

Figure 3. Confusion Matrix – RF + LSTM Model

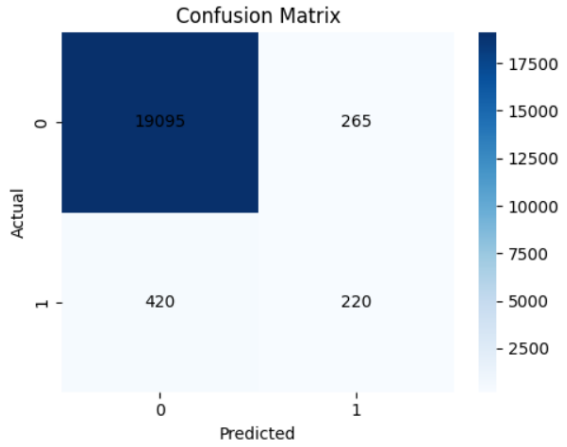

Figure 3 illustrates the confusion matrix for the RF + LSTM hybrid model on the test dataset. The model achieved a high number of true positives (18,093) and true negatives (220), while maintaining relatively low false positive (265) and false negative (403) counts. This indicates that the hybrid model effectively distinguishes between normal and anomalous events, reducing both missed detections and unnecessary alerts. The balanced distribution of correct classifications across both classes reflects the model's ability to generalize well to unseen data, supporting its suitability for deployment in operational Security Information and Event Management (SIEM) environments.

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 4. Accuracy and Loss Curve – RF + LSTM

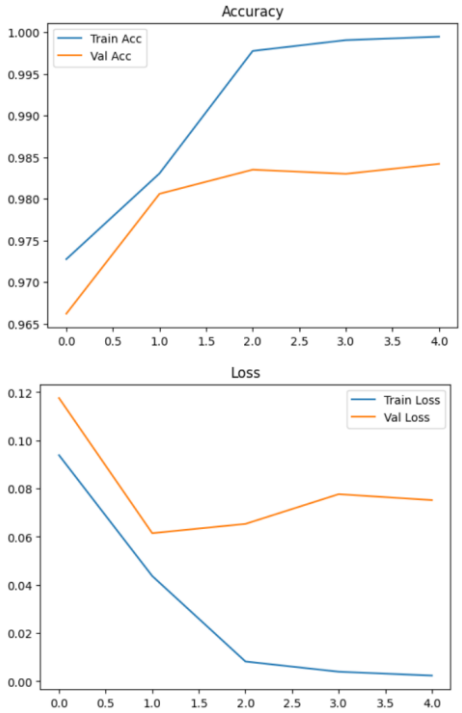

Figure 4 depicts the training and validation accuracy and loss curves for the hybrid RF + LSTM model across five training epochs. The accuracy curve shows consistent improvement for both training and validation sets, with training accuracy reaching nearly 100% and validation accuracy stabilizing around 98.4%, indicating strong generalization capabilities. The loss curve demonstrates a steady decline in training loss, while the validation loss exhibits slight fluctuations after the second epoch, but remains relatively low. The convergence behavior with minimal gap between training and validation performance confirms the absence of overfitting, further validating the model's reliability and robustness for real-world deployment

Inference-Based Interpretation with Large Language Models (LLM)

While achieving high detection accuracy is critical, operational cybersecurity teams also require interpretability clear, actionable explanations of detected anomalies to make rapid and informed decisions. To address this need, four Large Language Models (LLMs) from OpenAI GPT-3.5-Turbo, GPT-4, GPT-4-Turbo, and GPT-4o were integrated into the detection pipeline. This integration was designed to transform raw anomaly classifications into semantically rich, human-readable narratives, aligning with the interpretability objectives outlined in the methodology.

Each anomaly flagged by the detection models was converted into a structured prompt containing key contextual attributes: timestamp, source IP, rule ID, severity level, and event message. The LLMs were then tasked to produce:

1. A concise and accurate summary of the event.
2. Identification of the likely threat type and root cause.
3. Actionable mitigation recommendations.

Performance was evaluated against the four criteria established in the methodology:

Received: August 02, 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Semantic relevance – accuracy in reflecting the log's actual content and context.
- Clarity and actionability – practicality and completeness of recommended actions.
- Inference time – speed of generating responses, important for SOC latency requirements.
- Cost-efficiency – viability for real-time deployment in high-volume alert environments.

Tabel 2. LLM Comparison for Anomaly Interpretation

Model	**Inference Time**	**Summary Quality**	**Actionability**
GPT-3.5-Turbo	1.74s	Good	Moderate
GPT-4	2.89s	Excellent	High
GPT-4-Turbo	2.53s	Excellent	High
GPT-4o	1.97s	Excellent	High

From the comparative analysis, GPT-4o demonstrated the optimal balance between interpretive quality and inference speed. Its 1.97-second average response time met the near real-time requirements of SOC workflows, while consistently producing contextually accurate summaries and precise, multi-tier remediation strategies. For example (*Figure 5*), GPT-4o identified a sequence of repeated failed login attempts from IP address 192.168.1.100 as a brute-force attack. It recommended:

- Immediate actions: block the IP and lock affected accounts.
- Investigative steps: correlate additional logs, verify host integrity, and assess for lateral movement.
- Long-term measures**:** enforce MFA, strengthen password policies, and conduct user awareness training.

This performance aligns closely with the methodology's evaluation framework: GPT-4o maximized semantic relevance and clarity while maintaining latency suitable for high-throughput environments. By contrast, GPT-4 and GPT-4-Turbo offered equally high interpretive quality but with slower inference, potentially limiting scalability in large SOCs. GPT-3.5-Turbo, although fastest, delivered less detailed and less actionable recommendations, reducing its operational value.

The results confirm that integrating GPT-4o within the ML-DL detection pipeline effectively bridges the gap between anomaly detection and human-centered decision-making, enabling security teams to move from detection to remediation planning within a single, streamlined workflow.

Figure 5. GPT-4o Interpretation Output

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Model: gpt-4o
Inference Time: 9.87 seconds
Interpretation:
Based on the security log provided, here's an interpretation of what is happening:

1. Unauthorized Access Attempt: The log indicates that there has been an unauthorized access attempt detected from the IP address 192.168.1.100. The note "[Details: Source IP mismatch]" suggests that the access attempt might be coming from an IP address that does not match the expected or authorized source IP for the user or service trying to access the system.

2. Multiple Failed Login Attempts: In addition to the unauthorized access attempt, there have been multiple failed login attempts from the same host (192.168.1.100). This pattern is often indicative of a brute force attack, where an attacker is trying various passwords in an attempt to gain access to the system.

Actions to be Taken:

1. Immediate Response:

- Block the IP Address: Temporarily block the IP address 192.168.1.100 at the firewall or intrusion prevention system (IPS) level to prevent further access attempts.

- Account Lockout: If specific user accounts are being targeted, consider locking those accounts temporarily to prevent unauthorized access.

2. Investigation:

- Review Logs: Analyze additional logs to determine if there are any other

Received: August 02, 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

suspicious activities associated with this IP address or if other IPs are involved.

- Verify Source: Determine if 192.168.1.100 is a legitimate internal host or an external attacker. If internal, check for any signs of compromise on that machine.

- Check for Vulnerabilities: Ensure that the system is patched and that there are no known vulnerabilities that could be exploited.

3. Long-term Measures:

- Strengthen Authentication: Implement stronger authentication mechanisms such as multi-factor authentication (MFA) to reduce the risk of unauthorized access.

- Monitor and Alert: Set up alerts for similar unauthorized access attempts and failed login attempts in the future to ensure a quick response.

- User Education: Educate users about the importance of strong passwords and recognizing phishing attempts that could lead to credential compromise.

"Unauthorized access attempt detected from 192.168.1.100. Recommend immediate account lockdown, source IP investigation, and enabling MFA policies for user authentication." This layer of interpretation is crucial in operationalizing detection results and bridging the gap between raw log analysis and actionable insights for SOC teams.

4 Discussion

The experimental results of this study reaffirm findings from previous research that different model architectures offer distinct advantages in intrusion detection. As observed by Venturi et al (2022) and Sudyana et al (2024), the Random Forest model demonstrated strong performance in handling structured, high-dimensional tabular data, achieving the highest accuracy of 96%. However, its inability to capture temporal dependencies limited its capacity to detect multi-stage or time-sensitive attack patterns, a limitation also highlighted in earlier studies such as [8]. In contrast, the LSTM model, despite attaining a lower overall accuracy of 67%, excelled in modeling sequential patterns and detecting complex, multi-step intrusion

Received: August 02, 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

behaviors. This aligns with the findings of Shone et al (2018) and Vinayakumar et al (2019),which emphasized that deep learning models can learn latent temporal representations from sequential security data, enabling the identification of sophisticated attack chains.

The hybrid RF + LSTM model in this study successfully bridged the strengths of both approaches. By combining the structured feature learning capability of Random Forest with the sequential modeling power of LSTM, it achieved higher recall and a more balanced F1-score compared to the individual models. This outcome supports the relevance of prior works by Harwahyu et al (2024) and Xue & Chen (2022), who demonstrated that hybrid architectures can outperform standalone models by leveraging multiple feature domains. Importantly, our research extends these findings by applying the hybrid RF + LSTM model to real Wazuh agent logs within a SIEM environment, which, to the best of our knowledge, has not been comprehensively evaluated in existing literature.

A distinctive contribution of this study lies in the integration of Large Language Models (LLMs) for real-time interpretation of anomaly detection results. While explainable AI techniques such as SHAP and LIME [7] have provided feature-level attributions, they have fallen short in producing operationally actionable narratives for security analysts in Security Operations Centers (SOCs). Our experiments demonstrated that GPT-4o achieved the optimal trade-off between interpretive quality and inference latency, delivering contextually accurate summaries and detailed mitigation recommendations within an average of 1.97 seconds. This capability directly addresses a critical gap identified in previous research: the disconnect between anomaly detection and decision support [20]. By converting raw detection outputs into actionable insights, the integrated ML–DL–LLM framework enhances the operational readiness of SOCs and streamlines incident response workflows.

In comparison to earlier works that predominantly focused on improving detection accuracy [5], [6], our approach adds a crucial operational dimension ensuring that the outputs of detection models can be immediately applied to real-world cybersecurity operations. This positions the proposed framework not merely as an academic prototype but as a practical, deployable solution for enterprise-scale SIEM environments. The results directly address the research gaps identified in the Introduction: although prior studies such as Wibowo et al (2025) confirmed Wazuh's capability in detecting advanced persistent threats, they did not implement an integrated ML or DL pipeline, nor did they explore LLM-based interpretation within SIEM contexts. Similarly, while RF–LSTM hybrid models have been validated on benchmark datasets such as NSL-KDD, UNSW-NB-15, and CSE-CIC-IDS2018, their adaptation and evaluation using real-world Wazuh logs remained underexplored.

By evaluating Random Forest, XGBoost, LSTM, and a hybrid RF + LSTM on actual Wazuh logs, and enhancing their utility with LLM-based interpretability, this study delivers a unified framework that advances both detection accuracy and actionable explainability. The novelty of this work lies in its dual contribution: first, providing empirical evidence of the performance benefits of hybrid RF–LSTM for SIEM log anomaly detection; and second, demonstrating the operational viability of LLMs as an interpretability layer in security analytics. These contributions offer a methodological blueprint for future research and practical deployments, combining ML, DL, and LLMs into a single framework capable of supporting real-time SOC operations with both precision and clarity.

Received: August 02, 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

5 Conclusion

This study presented an integrated anomaly detection and interpretation framework for Wazuh agent logs, combining the strengths of machine learning (ML), deep learning (DL), and large language models (LLMs) to enhance both detection accuracy and operational interpretability. Four predictive models Random Forest, XGBoost, LSTM, and a hybrid RF + LSTM were implemented and evaluated on a dataset comprising both normal and anomalous log events. The results showed that while Random Forest achieved the highest accuracy for structured features, LSTM excelled at capturing temporal patterns. The hybrid RF + LSTM model successfully leveraged the complementary capabilities of both approaches, delivering balanced performance and improved recall, making it well-suited for real-world Security Information and Event Management (SIEM) environments where anomalies often exhibit both static and sequential characteristics.

A key novelty of this research lies in integrating LLMs, specifically GPT-3.5-Turbo, GPT-4, GPT-4-Turbo, and GPT-4o, to provide semantic interpretations and actionable recommendations for detected anomalies. GPT-4o emerged as the most effective model, offering a strong balance between interpretive quality and inference latency, thereby enabling near real-time decision support for Security Operations Centers (SOCs). This operational layer bridges the traditional gap between anomaly detection and actionable response, allowing security teams to transition rapidly from detection to remediation.

By addressing limitations in prior studies such as the lack of unified ML–DL–LLM architectures and the absence of real-world Wazuh log evaluations this work contributes both theoretically and practically to the field of cybersecurity. The findings suggest that hybrid learning architectures, augmented with advanced language models, can significantly improve the effectiveness and efficiency of SIEM-based security monitoring. Future research could extend this framework by exploring additional hybrid architectures, incorporating unsupervised learning for zero-day threat detection, and expanding LLM integration for multi-lingual SOC environments.

References

[1] Verizon, “2024 Data Breach Investigations Report,” 2024. [Online]. Available: https://www.verizon.com/business/resources/Tad3/reports/2024-dbir-data-breach-investigations-report.pdf

[2] M. F. Asyrofi and I. G. D. Nugraha, “Cybersecurity of Work from Anywhere Model for Government : A Systematic Literature Review,” *Int. J. Electr. Comput. Biomed. Eng.*, vol. 3, no. 1, May 2025, doi: 10.62146/ijecbe.v3i1.113.

[3] Wazuh Community Analytics, “Wazuh Documentation,” 2024. https://documentation.wazuh.com/current/user-manual/ruleset/index.html (accessed Aug. 04, 2025).

[4] B. Wibowo, A. Nurrohman, and L. Hafiz, “Deep Learning in Wazuh Intrusion Detection System to Identify Advanced Persistent Threat (APT) Attacks,” *Int. J. Sci. Educ. Cult. Stud.*, vol. 4, no. 1, pp. 1–10, Jan. 2025, doi: 10.58291/ijsecs.v4i1.311.

Received: August 02, 2025

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[5] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 1, Jan. 2021, doi: 10.1002/ett.4150.

[6] A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Knowledge-Based Syst.*, vol. 189, p. 105124, Feb. 2020, doi: 10.1016/j.knosys.2019.105124.

[7] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," *J. Inf. Secur. Appl.*, vol. 50, p. 102419, Feb. 2020, doi: 10.1016/j.jisa.2019.102419.

[8] G. Kocher and G. Kumar, "Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges," *Soft Comput.*, vol. 25, no. 15, pp. 9731–9763, Aug. 2021, doi: 10.1007/s00500-021-05893-0.

[9] N. Thapa, Z. Liu, D. B. KC, B. Gokaraju, and K. Roy, "Comparison of Machine Learning and Deep Learning Models for Network Intrusion Detection Systems," *Futur. Internet*, vol. 12, no. 10, p. 167, Sep. 2020, doi: 10.3390/fi12100167.

[10] C. Zhang, D. Jia, L. Wang, W. Wang, F. Liu, and A. Yang, "Comparative research on network intrusion detection methods based on machine learning," *Comput. Secur.*, vol. 121, p. 102861, Oct. 2022, doi: 10.1016/j.cose.2022.102861.

[11] R. Harwahyu, F. H. Erasmus Ndolu, and M. V. Overbeek, "Three layer hybrid learning to improve intrusion detection system performance," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 2, p. 1691, Apr. 2024, doi: 10.11591/ijece.v14i2.pp1691-1699.

[12] A. Hashmi, O. M. Barukab, and A. Hamza Osman, "A hybrid feature weighted attention based deep learning approach for an intrusion detection system using the random forest algorithm," *PLoS One*, vol. 19, no. 5, p. e0302294, May 2024, doi: 10.1371/journal.pone.0302294.

[13] X. Xue and L. Chen, "Intrusion Detection Based on LSTM and Random Forests," 2022, pp. 23–30. doi: 10.1007/978-3-030-89698-0_3.

[14] A. Venturi, C. Zanasi, M. Marchetti, and M. Colajanni, "Robustness Evaluation of Network Intrusion Detection Systems based on Sequential Machine Learning," in *2022 IEEE 21st International Symposium on Network Computing and Applications (NCA)*, Dec. 2022, pp. 235–242. doi: 10.1109/NCA57778.2022.10013643.

[15] D. Sudyana *et al.*, "Improving Generalization of ML-Based IDS With Lifecycle-Based Dataset, Auto-Learning Features, and Deep Learning," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, pp. 645–662, 2024, doi: 10.1109/TMLCN.2024.3402158.

[16] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A Deep Learning Approach to Network Intrusion Detection," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 2, no. 1, pp. 41–50, Feb. 2018, doi: 10.1109/TETCI.2017.2772792.

[17] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S.

Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," *IEEE Access*, vol. 7, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.

[18] M. R. A. Suhendi, Alfarizi, A. A. Sukmandhani, and Y. D. Prabowo, "Network Anomaly Detection Analysis using Artillery Honeypot and Wazuh SIEM," in *2023 IEEE 9th International Conference on Computing, Engineering and Design (ICCED)*, Nov. 2023, pp. 1–6. doi: 10.1109/ICCED60214.2023.10425009.

[19] N. C. Lasantha, R. Abeysekara, and M. Maduranga, "Defending Cloud Web Applications using Machine Learning-Driven Triple Validation of IP Reputation by Integrating Security Operation Center," *Glob. J. Comput. Sci. Technol.*, pp. 1–14, Oct. 2024, doi: 10.34257/GJCSTEVOL24IS1PG1.

[20] E. Karlsen, X. Luo, N. Zincir-Heywood, and M. Heywood, "Large language models and unsupervised feature learning: implications for log analysis," *Ann. Telecommun.*, vol. 79, no. 11–12, pp. 711–729, Dec. 2024, doi: 10.1007/s12243-024-01028-2.

[21] S. Han *et al.*, "Log-Based Anomaly Detection With Robust Feature Extraction and Online Learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 2300–2311, 2021, doi: 10.1109/TIFS.2021.3053371.

Received: August 02, 2025