Anomaly Detection in System Logs Through Contrastive Self-Supervised Learning Integrated with the Wazuh SIEM Platform

Aakash Singh[1], Sujit Shrestha[2]

[1]Faculty Member, Divya Gyan College, Tribhuvan University

[2]Faculty Member, Divya Gyan College, Tribhuvan University

Article History

Received: 14/12/2025

Revised: 7/1/2026

Accepted: 10/01/2026

Published: 15/01/2026

Corresponding Author

Name: Aakash Singh

Email: aakashsingh@divyagyan.edu.np

ORCiD: https://orcid.org/0009-0006-2271-7354

Abstract

Anomalies in system logs nowadays are very hard and difficult to identify due to their nature and originations from accounts of legitimate users. Traditional security systems seem to be very struggling to detect because threats depend on explicit attack signatures and the complex behavioral patterns of insider persons. This study gives us a framework which integrates contrastive self-supervised learning with the Security Information and Event Management (SIEM) platform to improve the detection of anomalies in system logs. The proposed system is using a data preprocessing pipeline, contrastive learning engine, and also integration interface which is capable of analyzing logs without hampering operational works. To evaluate the performance this study evaluated unsupervised and supervised algorithms. The results gain a high accuracy and an F1-score in favor of Random Forest algorithm. This research shows if we combine temporal activity patterns with organizational context in open source SIEM platform we can find improved threat detection capabilities. The research focuses on modern SIEM platforms for better detection of anomalies in real time environments, showing better results which are based on different evaluation techniques.

Keywords: ***Anomaly Detection, Contrastive Learning, Random Forest, Isolation Forest, Wazuh SIEM, NSL-KDD.***

Introduction

Security Information and Event Management (SIEM) systems are crucial for modern cybersecurity operations. As organizations are nowadays expanding their digital infrastructure by adopting cloud services as well as distributed systems and remote work models. The volume and complexity of system logs have increased in a devastating way. SIEM platforms were developed to address the different challenges given by centralizing logs collection by enabling real-time monitoring and correlating different events across the systems and also generating notifications when suspicious activity is found (Muhammad et al., 2023; Khayat et al., 2025). Traditional SIEM platforms are largely driven by rule-based logic and signature matching methods. These mechanisms perform good when identifying known attack patterns [2]. The real problem arises when attackers come from predictable behaviors. Modern threat actors continuously modify their techniques and exploit already detected unknown vulnerabilities. In cases like that, static detection rules are often not sufficient (Schindler, 2018).

Enterprise networks like applications and also servers including different node from user's perspectives frequently produces vast streams of data. Analysts who work as a Security Operations Centers must deal with thousands and sometimes lakhs of alerts per day (Chamkar et al., 2025). This will create burden for analyst.

Among open-source SIEM solutions the Wazuh system has gained good adoption nowadays due to its flexibility and extensible feature set. It supports intrusion detection with file integrity monitoring as well as vulnerability assessment and compliance auditing (Muhammad et al., 2023). It integrates better with other tools in the ecosystem of security and also it provides organizations with cost-effective real time monitoring features. If we compare traditional SIEM systems with Wazuh it primarily depends on predefined rules of detection and database of signature. Also, these rules can be customized but they still need manual maintenance and continuous updates (Khayat et al., 2025). This dependency on static rules hinders Wazuh's effectiveness against attacks like zero-day exploits and hinders behavioral anomalies which may not violate patterns of existing detections techniques (Schindler, 2018).

The recent progress in machine learning have introduced different alternatives. Self-supervised learning has come up as a powerful mechanism for extracting meaningful patterns from datas that are unlabeled (Liu et al., 2021). Contrastive self-supervised learning is seen training models by contrasting between similar and dissimilar data pairs without the need of manual labels and also it has gained noticeable success among field like computer vision and also natural language processing including cybersecurity (Hojjati & Armanfard, 2022; Liu et al., 2021).

Log anomaly detection is found best suited for the approaches like this. In real world environments there is scarcity of log data and also there is imbalanced datas and they are expensive to obtain (Aziz & Munir, 2024; Grover, 2018). Self-supervised contrastive learning gives us a way to learn normal behavioral representations straight from log(raw) sequences. Also, the deviations from these representations can be represented as anomalies (Le & Zhang, 2021).

This study is seen exploring on how the contrastive self-supervised learning can also be integrated with the SIEM platform that is found enhancing anomaly detection capacities.

Statement of the Problem

Despite the heavy use the existing SIEM systems is found struggling to detect threat and face many limitations. One of the most popular challenge which is seen is the dependence on predefined detection rules (Muhammad et al., 2023). Also, they are seen struggling to detect novel attack techniques or versatile behavior of attacking. Advanced persistent threats (APTs) and zero-day exploits seems removing this gap by mimicking activity of genuine or legitimate users (Schindler, 2018; Dumitrasc, 2023).

Alert overload is another critical issue. Modern IT infrastructures obtains enormous amount of log volumes and traditional SIEM systems continuously flag normal activity as anomalies or suspicious (Chamkar et al., 2025). The massive false positives increase cost of operations and allows real incidents to be overlooked (Khayat et al., 2025).

Maintaining detection rules is seems to have demanding the good expertise and continuous updates. As threats evolve the security analysts must do revision of rule which sets to remain

effective [14]. This process is very time consuming and also it introduces the risk of configuration errors and also in organizations where there are limited cybersecurity resources (Ahmad, 2025).

Although we have machine learning that are proposed as a good solution there are many existing approaches that depend on supervised techniques which also require training data that are labeled (Aziz & Munir, 2024). In practice, labeled data that are high quality and good datasets of anomalies are rarely available (Grover, 2018; Le & Zhang, 2021). So exactly this scarcity in data stops the scalability and applicability of solutions that are supervised.

So, there is a clear need for detection frameworks which are capable of learning from datas that are unlabeled and also integrating with existing SIEM platforms, and which reduces both false positives and rule maintenance which are manual (Liu et al., 2021; Khayat et al., 2025).

This research has following questions:

1. How contrastive self-supervised learning can be effectively integrated into the Wazuh SIEM platform that can enhance accuracy of log anomaly detection?
2. What architectural and implementation strategies is found enabling integration without hampering the existing workflows?
3. How does the proposed approach will compare rule-based detection and supervised learning methods in terms of accuracy and also false positive rate including computational efficiency?
4. What training strategies, hyperparameter configurations will optimize the performance in enterprise log formats?

The objectives of this study are as follows:

- To design and evaluate a contrastive self-supervised anomaly detection framework by integrating with Wazuh.
- To benchmark its performance against rule-based detection and supervised learning models with different metrics like precision, recall, F1 score and utilization of resource

Literature Review

The integration of machine learning with SIEM systems nowadays is an active area of research. Traditional log anomaly detection mainly focuses on statistical methods and rule-based systems. Schindler (2018) had showed the disadvantages of commercial SIEMs which handles very complex attack patterns.

Le and Zhang (2021) proposed NeuralLog, a BERT-based approach which is found eliminating the need for log parsing.

Liu et al. (2021) presented CoLA, which is a contrastive framework for detection of anomalies on attributed networks. Zheng et al. (2021) elaborated the field by adding contrastive and generative learning which is used for graph anomaly detection. Hojjati and Armanfard (2022) showed that contrastive learning principles are used in anomaly detection which can be transferred to log-based analysis.

Despite these factors, limited research has been done on the practical integration of advanced techniques of machine learning with SIEM platforms. Many studies are only theoretical which showed a kind of gap in understanding actual deployment in real-world environments (Grover, 2018).

Research Methodology

This study shows a mixed methods techniques which combines experimental research with practical implementation. The system used has three parts: a part that gets the data ready a part that helps the system learn on its own and a part that connects with the Wazuh SIEM platform. The part that gets the data ready takes care of a things. It looks at the logs that come in makes sure they are all in the format and pulls out the important information. The logs from different places are taken by it and it makes them useful by focusing on the good things and content. This part is also good at figuring out how to read kinds of logs that you might find in a big company. The part that helps the system learn on its own is showing the part of the system which they are proposing. The system they are talking about uses this learning part to make it work. The learning part is really important for the system to be good at its job.

The system is made up of the data preparation part the learning part and the connection part, with the Wazuh SIEM platform and the learning part is what makes the system special. This component implements a novel contrastive learning architecture specifically designed for sequential log data. The model learns meaningful representations by contrasting normal operational patterns against potential anomalies through carefully designed positive and negative pair generation strategies. The integration interface provides good connectivity with the SIEM platform which enables the real time anomaly detection and alert generation without hampering existing operational workflows.

System Architecture

The architecture of this study consists of three primary components:

1. **Pipelining of Data Preprocessing:** It handles the ingestion of log, normalization, and feature extraction using technique like adaptive parsing.
2. **Contrastive Self-Supervised Learning Engine:** It uses a architecture of contrastive learning which learns by contrasting normal operational patterns against potential anomalies.
3. **Integration Interface:** This connects with the SIEM platform via APIs to enable detection in real-time.

Algorithms Implemented

To evaluate the performance and the differences between supervised and unsupervised methods following algorithm was implemented by the study:

- **Random Forest:** An ensemble learning method using bagging which is used to build multiple decision trees.
- **Isolation Forest:** An unsupervised algorithm which isolates anomalies rather than showing normal data points.
- **InfoNCE:** A loss function is used to train models to find differences between similar and dissimilar data pairs.

Data Preparation

The study used the NSL-KDD dataset which is a balanced version of the KDD dataset. It is widely used for intrusion detection systems. Data preprocessing involved parsing, normalization, feature extraction, and temporal sequencing.

Integration Strategy

An Ubuntu node with SIEM agent was configured to gain resource usage and log data to the Wazuh server.

Finding and Result

The system was tested in Ubuntu with the SIEM server as a real-world scenario. Various anomalies were simulated which include failed login attempts and also abnormal utilization of Linux resource.

Operational Results

The integration allowed for real-time monitoring on the dashboard of SIEM platform.

- **Failure in Logins:** The anomaly detector showed irregular login patterns.
- **Utilization of Resource:** Anomalies in CPU and also in memory usage were found and visualized on the dashboard.
- **Threat Detection:** The system captured alerts for vulnerability and also malware detection within the last 24-hour.

Performance Evaluation

These models were evaluated which are based on Accuracy, Precision, Recall, and F1-Score.

Table 1: Model Comparison based on performances

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.997	0.970	0.980	0.970
Isolation Forest	0.806	0.555	0.147	0.249

Discussion

The Random Forest model showed very good performance with high accuracy and a high F1-score. This showed that when labeled data is available, supervised methods are highly effective.

The Isolation Forest showed significantly lower performance mainly in Recall. Whereas the accuracy in initial stage was seen acceptable but it greatly reflects the dominance of normal transactions in the dataset. Also, the low recall indicates that the unsupervised model missed a number of real anomalies, and the low precision indicates a high rate of false positives.

A scatter plot analysis showed us that the algorithm could identify outliers (red points) far from dense clusters of normal logs (blue points) where the unsupervised approach struggled with the data imbalance which leads to high flagging of normal logs as threats which is also called false positives.

Conclusion

This study generates anomaly framework identification that is based on contrastive self-supervised learning which is integrated with the SIEM platform. The evaluation results shows that supervised methods like Random Forest significantly outperform unsupervised methods like Isolation Forest in detecting anomalies. The integration with Wazuh proves that when we combine temporal activity patterns with organizational context it gives better detection performance and offers a scalable solution for enterprise environments. This approach reduces the dependency on static rules and manual analysis.

Traditional models which seem to be dependent on labeled data or static rules, this approach is flexible, less resource required and also better suited for real-world deployment mainly in the organizations or offices where there are limited cybersecurity capabilities.

In dynamic network environments where attack patterns are constantly changing, the suggested framework exhibits strong adaptability. Without requiring a lot of manual labeling, the system can learn meaningful representations of typical behavior straight from raw log sequences by utilizing contrastive self-supervised learning. This greatly lessens the need for frequent rule updates and expert-driven rule engineering, which are frequent problems in traditional SIEM deployments.

Improvements in detection accuracy, precision, and recall are also highlighted by the experimental analysis, which also shows a decrease in false positive rates, which usually

overwhelm Security Operations Center (SOC) analysts. Improved operational efficiency and quicker incident response times are two benefits of fewer false alarms. Furthermore, compatibility with current enterprise infrastructures is guaranteed by the modular integration with Wazuh, which makes deployment feasible and economical. Overall, the framework provides a balanced solution with the combination of intelligent anomaly detection and real time monitoring which shows the way for adaptive mechanisms for cyber security defense in modern organizations.

References

Ahmad, W. (2025). Unveiling Anomalies: Leveraging Machine Learning for Internal User Behaviour Analysis–Top 10 Use Cases. *International Journal of Innovative Technology and Interdisciplinary Sciences, 8*(1), 1789-1805.

Aziz, A., & Munir, K. (2024). Anomaly Detection in Logs using Deep Learning. *IEEE Access, 12*, 176129-176145. https://doi.org/10.1109/ACCESS.2024.3512847

Chamkar, S. A., Zaydi, M., Maleh, Y., & Gherabi, N. (2025). ML-Driven Log Analysis for Real-Time Cyber Threat Detection in Security Operation Centers. *Preprints*. https://doi.org/10.20944/preprints202412.0123.v1

Dumitrasc, V. (2023). *Anomaly Detection Through User Behaviour Analysis* [Master's thesis, Universitat Politècnica de Catalunya].

Grover, A. (2018). *Anomaly detection for application log data* [Master's thesis, San José State University].

Hojjati, H., & Armanfard, N. (2022). Self-supervised acoustic anomaly detection via contrastive learning. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3371-3375). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9746207

Khayat, M., Barka, E., Serhani, M. A., Sallabi, F., Elmedany, M., & Bentahar, H. (2025). Empowering Security Operation Center with Artificial Intelligence and Machine Learning–A Systematic Literature Review. *IEEE Access, 13*, 12045-12067. https://doi.org/10.1109/ACCESS.2025.3523901

Le, V. H., & Zhang, H. (2021). Log-based anomaly detection without log parsing. *Proceedings of the 36th IEEE/ACM International Conference on Automated Software Engineering* (pp. 492-504). https://doi.org/10.1109/ASE51524.2021.9678773

Liu, Y., Li, Z., Pan, S., Gong, C., Zhou, C., & Karypis, G. (2021). Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE Transactions on Neural Networks and Learning Systems, 33*(8), 8270-8282. https://doi.org/10.1109/TNNLS.2021.3068344

Muhammad, A. R., Sukarno, P., & Wardana, A. A. (2023). Integrated security information and event management (SIEM) with intrusion detection system (IDS) for live analysis based on machine learning. *Procedia Computer Science, 217*, 1528-1537. https://doi.org/10.1016/j.procs.2022.12.352

Schindler, T. (2018). Anomaly detection in log data using graph databases and machine learning to defend advanced persistent threats. *arXiv preprint arXiv:1802.00259*.