Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Implementation of Isolation Forest-Based Machine Learning in Batch Anomaly Detection on Zeek Log Data (Case Study: Langkat Regency Communication and Information Agency)

Muhammad Al Kahfi[1]*, Relita Buaton [2], I Gusti Prahmana [3]

[1,2,3]*STIMK Kaputama,Indonesia*

muhammadalkahfi1900@gmail.com[1*], bbcbuaton@gmail.com[2], igustiprahmana4@gmail.com[3]

Abstract

The high escalation of cyber threats against government institutions requires an adaptive and intelligent digital security system. The Langkat Regency Communication and Information Agency faces challenges in analyzing large volumes of network log data to effectively detect suspicious activity. This study aims to implement the Isolation Forest machine learning algorithm to detect anomalies in batches on Zeek log data, and classify detected anomalies into threat levels to facilitate security audits. Using the CRISP-DM framework, this study analyzed 12.1 million lines of Zeek conn.log data from December 2024 through the stages of data preparation, unsupervised modeling with Isolation Forest, and manual threshold determination for classification. The effectiveness of the model is evaluated using Precision, Recall, and F1-Score metrics against proxy labels, and the results are enriched with rule-based labeling to determine threat levels. The results of the study show that the model successfully identified 15.34% of connections as anomalies, with the dominant pattern categorized as a "High" threat detected in DNS and unknown services, indicating potential malicious activity. Quantitative evaluation yielded a precision of 0.41 and a recall of 0.08, highlighting the model's ability to detect more subtle anomalies beyond simple rules. Thus, the implementation of Isolation Forest proved effective in identifying diverse network anomaly patterns, where its combination with rule-based labeling provides functional threat context for cybersecurity teams.

Keywords: *Anomaly Detection; Cybersecurity; Isolation Forest; Network Logs; Zeek*

1. Introduction

Cybersecurity is a crucial global issue, in line with the rapid development of information and communication technology [3, 4]. Cyber threats such as malware, DDoS, and illegal access continue to increase, targeting various sectors, including government institutions that manage vital data and critical service systems. In Indonesia, the National Cyber and Crypto Agency (BSSN) noted that in 2022 there were 370.02 million cyber attacks, an increase of 38.72% from the previous year, highlighting the urgency of improving digital security systems.

Local government agencies, such as the Langkat Regency Communication and Information Agency (Kominfo), manage large-scale network infrastructure to support public services. However, this complexity also increases the risk of cyber attacks . Analysis of network log data is a critical step in detecting suspicious activity, but the enormous volume of log data makes manual review ineffective and prone to errors. Machine learning offers a practical solution, and the Isolation Forest algorithm is particularly effective for this task [7]. This algorithm works by isolating data based on anomaly levels without requiring labeled data, making it suitable for identifying anomalous patterns quickly and efficiently in large-scale datasets [2, 5, 7].

This research aims to implement the Isolation Forest algorithm to identify anomaly [1] patterns in Zeek log data at the Langkat Regency Kominfo, evaluate its effectiveness using Precision, Recall, and F1-score, and classify the detected anomalies into threat levels using a rule-based approach to assist the cybersecurity team.

2. Research Method

This study adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which consists of six structured phases to ensure a reproducible and goal-oriented workflow [8].

Fig. 1 : CRISP-DM

2.1. Business and Data Understanding

The primary objective is to detect batch anomalies in Zeek's conn.log data from December 2024 to enhance the network security of the Langkat Regency Kominfo. The dataset consists of 12.1 million rows of network connection logs generated by the Zeek system. An initial exploration was conducted to understand the distribution of attributes and identify potential data quality issues.

2.2. Data Preparation

The raw data underwent several preparation steps:

1. Data Cleaning: Missing values (represented by '-') were handled; the service column's missing values were filled with 'unknown', and key numerical columns were filled with 0. The ts (timestamp) column was converted to a datetime format.
2. Feature Engineering: New features were created to enrich the model's input, including time-based features (hour_of_day, day_of_week), a ratio feature (bytes_per_packet), and a binary proxy label (is_failed_conn) to identify failed connections for evaluation purposes.
3. Encoding and Scaling: Categorical features like proto and service were transformed using Frequency Encoding, which is effective for anomaly detection as rare categories receive lower numerical values. All numerical features were then scaled using RobustScaler, which is resistant to outliers.

2.3. Modeling

The Isolation Forest algorithm from the Scikit-learn library was chosen for its unsupervised nature and efficiency with large datasets [5]. The model was trained with 100 estimators (n_estimators=100) to learn the data structure without labels. After training, the model calculated a raw anomaly score for each connection, where a more negative score indicates a higher likelihood of being an anomaly. A manual classification threshold was set at -0.01 based on an analysis of the score distribution to distinguish between "Normal" and "Anomalous" connections [6].

2.4. Evaluation and Deployment

The model's performance was quantitatively measured against the created proxy label (is_failed_conn) using a Confusion Matrix, Precision, Recall, and F1-score . To add context to the findings, a Rule-Based Labeling system was applied to classify detected anomalies into "High," "Medium," and "Low" threat levels . The final results are presented in a structured Jupyter Notebook to facilitate auditing and investigation by the security team.

3. Results and Discussion

3.1. Anomaly Detection Results

Using the -0.01 threshold, the model identified 1,858,411 connections (15.34%) as anomalies, while the remaining 84.66% were classified as normal. The distribution of anomaly scores showed a clear separation between normal data (concentrated in positive scores) and anomalous data (spread across negative scores).

Analysis of the services associated with anomalies revealed that the "unknown" and "dns" services had the highest number of detections. This is a significant finding, as "unknown" services often indicate the use of non-standard ports to evade detection, while DNS anomalies can be linked to malicious activities like DNS tunneling.

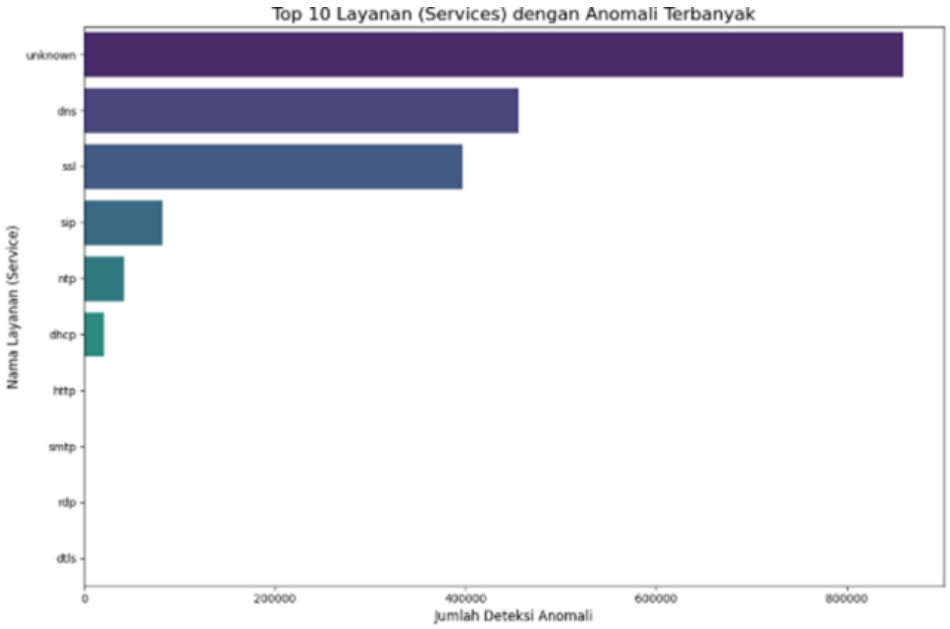

Fig. 2 : Top 10 Services with the Most Anomalies

3.2. Model Performance Evaluation

When evaluated against the proxy label (failed connections), the model achieved the following for the anomaly class:

1. Precision: 0.41
2. Recall: 0.08

The low recall suggests that the model did not identify most of the simple "failed" connections. However, this is expected from an unsupervised method like Isolation Forest, whose strength lies in detecting more subtle and complex anomalies that a simple rule (like is_failed_conn) would miss. The precision of 0.41 indicates that when the model flags an anomaly, there is a 41% chance it corresponds to a failed connection, which is a reasonably strong signal.

Table 1 : Classification information table

	precision	recall	f1-score	support
Normal	0.15	0.59	0.24	2642633
Anomali (Proxy)	0.41	0.08	0.14	9473926

3.3. Threat Level Classification

The rule-based system classified the detected anomalies into threat levels, providing actionable context for the security team:

1. High Threat (1,315,333 instances): Anomalies in 'unknown' or 'dns' services.
2. Medium Threat (168 instances): Failed connection attempts on critical services like 'http' or 'ssh'.
3. Low Threat (542,910 instances): Other detected anomalies.

This classification successfully addresses the research goal of not only detecting but also contextualizing anomalies, allowing for prioritized investigation.

4. Conclusion

This research successfully demonstrated the implementation of the Isolation Forest algorithm for detecting anomalies in Zeek network log data. The model effectively identified dominant anomaly patterns, particularly in "unknown" and "dns" services, which indicate potential sophisticated threats. Although the quantitative evaluation against a simple proxy label showed low recall, the model proved its value by detecting a broader range of subtle anomalies that rule-based systems would overlook. The combination of unsupervised detection with a rule-based labeling system provides a functional and flexible framework that allows security teams to classify threats and prioritize their response efforts effectively.

References

[1] Airlangga, G. (2023). UNSUPERVISED MACHINE LEARNING FOR SEISMIC ANOMALY DETECTION: ISOLATION FOREST ALGORITHM APPLICATION TO INDONESIAN EARTHQUAKE DATA. *4*(3), 1827–1836. https://doi.org/10.46306/lb.v4i3

[2] Akoh Atadoga, Enoch Oluwademilade Sodiya, Uchenna Joseph Umoga, & Olukunle Oladipupo Amoo. (2024). A comprehensive review of machine learning's role in enhancing network security and threat detection. *World Journal of Advanced Research and Reviews, 21*(2), 877–886. https://doi.org/10.30574/wjarr.2024.21.2.0501

[3] Chua, W., Pajas, A. L. D., Castro, C. S., Panganiban, S. P., Pasuquin, A. J., Purganan, M. J., Malupeng, R., Pingad, D. J., Orolfo, J. P., Lua, H. H., & Velasco, L. C. (2024). Web Traffic Anomaly Detection Using Isolation Forest. *Informatics, 11*(4). https://doi.org/10.3390/informatics11040083

[4] Djidjev, C. (2024). *siForest: Detecting Network Anomalies with Set-Structured Isolation Forest*. http://arxiv.org/abs/2412.06015

[5] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data, 6*(1). https://doi.org/10.1145/2133360.2133363

[6] Moomtaheen, F., Bagui, S. S., Bagui, S. C., & Mink, D. (2024). Extended Isolation Forest for Intrusion Detection in Zeek Data. *Information (Switzerland), 15*(7). https://doi.org/10.3390/info15070404

[7] Ripan, R. C., Sarker, I. H., Anwar, M. M., Furhad, Md. H., Rahat, F., Hoque, M. M., & Sarfraz, M. (2020). *An Isolation Forest Learning Based Outlier Detection Approach for Effectively Classifying Cyber Anomalies*. http://arxiv.org/abs/2101.03141

[8] Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science, 181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199