UDC 004.896:004.852.4

CONCEPTUAL FRAMEWORK AND THEORETICAL CHALLENGES IN UNSUPERVISED ANOMALY DETECTION FOR NETWORK TRAFFIC DATA

X. WANG, A.M. PRUDNIK

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus

Received April 4, 2025

Abstract. This paper outlines the initial design considerations for a system aimed at processing and analyzing network traffic data to detect anomalies using machine learning. The study explores anticipated challenges in data preprocessing, scalability, and algorithm selection, emphasizing the potential of unsupervised learning methods to identify unusual patterns in network traffic. The proposed approach serves as a foundation for future development of anomaly detection systems.

Keywords: network traffic processing, anomaly detection, machine learning, system design challenges

Introduction

Rising network traffic volume and complexity, driven by connected devices and digital services, challenge security systems [1]. Modern networks process high-dimensional, variable data streams, complicating anomaly detection [2]. Signature-based methods excel at known threats but fail against novel attacks [3]. This research proposes a machine learning-based framework for adaptive network traffic analysis, targeting the reliance of traditional methods on predefined signatures and static rules [4]. It leverages machine learning to detect subtle deviations without extensive labeled data, despite technical and operational hurdles [5]. The framework focuses on three challenges: algorithm selection, large-scale preprocessing, and real-time scalability. Unlike implementation-focused studies, it emphasizes theoretical and architectural foundations for robust anomaly detection, contributing to data-driven network security advancements.

Proposed System Architecture

This section outlines a modular architecture for the anomaly detection system, addressing challenges from the introduction. It includes three components: data input/preprocessing, processing layer, and output generation, designed for scalability and adaptability.

The data acquisition framework uses synthetic traffic data for initial validation, aligning with practices in [6]. It incorporates features like temporal patterns and protocol distributions [7]. Operational deployment faces data quality issues, with excessive traffic volume [8]. Adaptive preprocessing pipelines normalize formats, preserving key features. Flexible interfaces support batch and streaming data, meeting large-scale monitoring needs [9].

The processing layer employs a pipeline for analysis. Feature extraction uses dimensionality reduction and *z*-score normalization to unify diverse metrics. The detection module supports multiple algorithms, addressing varied attack vectors and ensuring adaptability, with flexibility for future methods.

Output Generation and Alert Management

The output module uses a multi-tiered severity classification framework for network anomalies. It generates normalized scores (0–1) based on deviations from baseline traffic, following visualization

practices. Scores align with findings that consistent metrics aid decision-making [10]. Anomalies scoring above 0.8 are critical, prioritized per research showing graduated alerts reduce cognitive load [11]. Protocol-agnostic data structures ensure integration with analysis frameworks, retaining threat context.

The system employs a stateless architecture, storing results in a format compatible with SIEM (Security Information and Event Management) systems, supporting interoperability. Alert generation uses correlation and aggregation to cut redundancy while preserving temporal links. Figure 1 depicts the architecture, showing the relationship between input processing, analysis, and output components.

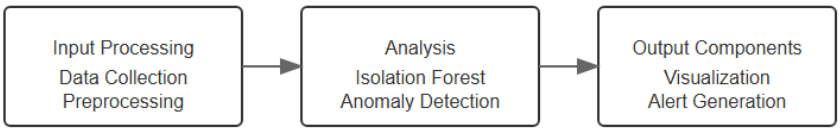

Figure 1. High-level system architecture

Machine Learning Approach

Machine learning outperforms signature-based methods in network anomaly detection, excelling at zero-day and evolving threats. Unsupervised learning addresses the lack of labeled data in operational networks, where attack vectors evolve rapidly. It detects anomalies by deviations from normal behavior, offering scalability for novel threats.

The isolation forest algorithm enhances efficiency in high-dimensional traffic data [12]. It isolates sparse anomalies with $O(n \log n)$ complexity, enabling real-time processing [13]. However, dynamic traffic complicates model adaptation and threshold tuning, and opaque outputs hinder interpretability for security responses. Our framework uses a hybrid approach with visualization. Figure 2 shows normal vs. anomalous patterns in feature space, aiding decision-making. The isolation forest decision boundaries highlight regions where anomalies deviate from normal traffic distributions. High-volume attacks cluster in the upper right, while low and slow attacks appear in the lower left, demonstrating distinct traffic behaviors. The shades of blue in the background represent varying anomaly scores, with darker regions indicating areas more likely to be classified as normal, providing a visual guide for decision-making.

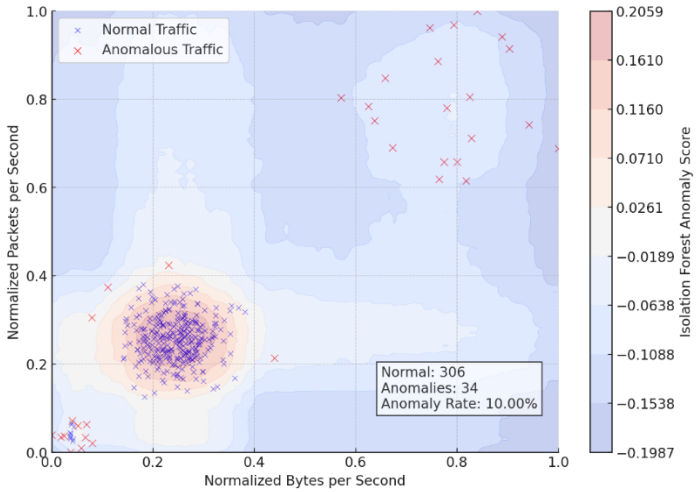

Figure 2. Conceptual visualization of normal versus anomalous traffic patterns

Challenges and Limitations

Implementing unsupervised anomaly detection in network traffic faces challenges beyond algorithms, affecting accuracy and performance. Data quality is a hurdle in operational networks, producing heterogeneous traffic with inconsistent sampling and missing values, cutting detection accuracy. Robust preprocessing is needed to manage irregularities and retain anomaly indicators. Feature selection is critical in high-dimensional traffic, with over 40 features identified [6]. Optimal sets vary by attack type, requiring adaptive mechanisms. Scalability limits real-time analysis, with traffic

over 100 Gbps straining resources [14]. Even optimized systems falter during spikes, needing efficient algorithms and resource management. Parameter optimization in isolation algorithms impacts accuracy by 15 % [15], complicated by dynamic traffic and no labeled data. Interpretability lags, with "black box" outputs increasing alert investigation time by 73 % [16]. Adaptive tuning and explainable frameworks are essential. Figure 3 shows these challenges' overlap.

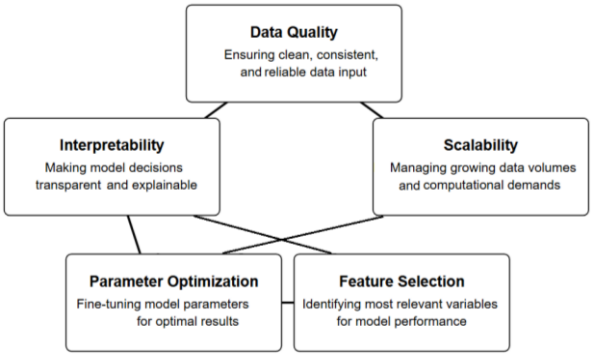

Figure 3. Interconnected relationship between key system design challenges

Conclusion

This research outlines a theoretical framework for network traffic anomaly detection, using unsupervised learning in a modular design. It tackles limitations in current security methods via algorithmic adaptability and flexible architecture. Key challenges identified include data quality, feature optimization, scalability, parameter tuning, and interpretability, offering avenues to advance network security analytics. Future work will involve empirical validation with operational data and algorithm refinement based on performance metrics. This foundation supports robust anomaly detection systems for emerging threats. Ongoing study of these challenges will enhance understanding of unsupervised learning in network security.

References

1. Sommer R., Paxson V. // IEEE Symposium on Security and Privacy. 2010. P. 305–316.
2. García-Teodoro P., Díaz-Verdejo J., Maciá-Fernández G., Vázquez E. // Computers & Security. 2009. Vol. 28. P. 18–28.
3. Buczak A. L., Guven E. // IEEE Communications Surveys & Tutorials. 2016. Vol. 18. No. 2. P. 1153–1176.
4. Ahmed M., Mahmood A. N., Hu J. // Journal of Network and Computer Applications. 2016. Vol. 60. No. C. P. 19–31.
5. Bhuyan M. H., Bhattacharyya D. K., Kalita J. K. // IEEE Communications Surveys & Tutorials. 2014. Vol. 16. P. 303–336.
6. Ring M., Wunderlich S., Grüdl D., Landes D., Hotho A. // Proc. of the 16th European Conf. on Cyber Warfare and Security. 2017. P. 361–369.
7. Sharafaldin I., Lashkari A. H., Ghorbani A. A. // Proc. of the 4th International Conf. on Information Systems Security and Privacy. 2018. P. 108–116.
8. Bhuyan M. H., Bhattacharyya D. K., Kalita J. K. // Network Traffic Anomaly Detection and Prevention: Concepts, Techniques, and Tools. Springer International Publishing, 2017.
9. Cordero C. G., Vasilomanolakis E., Milanov N., Koch C., Hausheer D., Mühlhäuser M. // IEEE Conf. on Communications and Network Security (CNS). 2015. P. 739–740.
10. D'Amico A., Whitley K. // VizSEC 2007: Proc. of the Workshop on Visualization for Computer Security. 2008. P. 19–37.
11. Perdisci R., Ariu D., Fogla P., Giacinto G., Lee W. // Computer Networks. 2009. Vol. 53. No. 6. P. 864–881.
12. Liu F. T., Ting K. M., Zhou Z. H. // ACM Transactions on Knowledge Discovery from Data. 2012. Vol. 6. No. 1. P. 1–39.
13. Ding Z., Fei M. // IFAC Proceedings Volumes. 2013. Vol. 46. No. 20. P. 12–17.
14. Perdisci R., Lee W., Feamster N. // 7th USENIX Symposium on Networked Systems Design and Implementation. 2010. Vol. 12. P. 26.
15. Wang H., Bah M. J., Hammad M. // IEEE Access. 2019. Vol. 7. P. 107964-108000.
16. Ribeiro M. T., Singh S., Guestrin C. // Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135–1144.