Machine Learning for Network Anomaly and Failure Detection

CUNY School of Professional Studies

Michael Hernandez

IS 499 Information Systems Capstone

Professor John Bouma

September 27, 2025

Contents

1 Introduction

This paper examines machine learning techniques for detecting and localizing network anomalies and failures in large-scale environments, using data from BGP routing updates, SNMP metrics, and syslog messages.

Traditional network monitoring relies on threshold-based alerts from SNMP and syslog, often producing many false positives and offering little context for locating failures (Wang, 2020; Manna & Alkasassbeh, 2019). Recent research has demonstrated that machine learning approaches applied to SNMP-MIB datasets can significantly improve anomaly detection accuracy and operational efficiency, with Random Forest classifiers achieving up to 100% accuracy in identifying network failures (Manna & Alkasassbeh, 2019). This project implements a dual-pipeline machine learning architecture with specialized algorithms for each data source: Matrix Profile for BGP time-series analysis (Scott et al., 2024) and Isolation Forest for SNMP hardware anomaly detection (Liu et al., 2008).

The system integrates two parallel detection pipelines for comprehensive network monitoring. The BGP pipeline uses Matrix Profile to detect routing anomalies in update streams, while the SNMP pipeline employs Isolation Forest to identify hardware and environmental failures in multi-dimensional feature spaces. Syslog messages provide temporal correlation signals, and multi-modal fusion reduces false positives by requiring confirmation across multiple data sources before generating alerts.

2 Topic Description

2.1 In-depth Description of the Chosen Topic

This project implements a dual-pipeline machine learning architecture for network anomaly detection and failure localization. The system employs specialized algorithms optimized for different data characteristics: Matrix Profile for BGP time-series analysis and Isolation Forest for SNMP high-dimensional feature analysis. This architecture enables comprehensive monitoring by processing BGP routing updates, SNMP hardware metrics, and syslog event messages through appropriate detection algorithms, then fusing signals for confirmed anomaly identification.

Border Gateway Protocol (BGP) is a standardized exterior gateway protocol that provides connectivity and fault tolerance for network devices on the Internet by exchanging routing and reachability information among autonomous systems. As a path-vector protocol, BGP maintains tables of IP network prefixes and makes routing decisions based on network policies configured by administrators. BGP is fundamental to Internet operation, dynamically updating routing information to adapt to network changes such as link failures or topology modifications. In the context of this project, BGP updates serve as critical indicators of network state changes, with anomalous BGP behavior often signaling underlying network issues (Scott et al., 2024).

The dual-pipeline architecture reflects the distinct characteristics of network telemetry data (Feltin et al., 2023). The BGP pipeline processes time-series data consisting of routing update sequences, where temporal patterns and sudden changes indicate network instability (Scott et al., 2024). The SNMP pipeline processes multi-dimensional feature vectors extracted from hardware metrics, where outliers in the feature space indicate component failures or environmental issues (Manna & Alkasassbeh, 2019). This separation allows each pipeline to use algorithms optimized for its specific data type, while the fusion layer combines their outputs for comprehensive failure detection (Mohammed et al., 2021). Real-time processing requirements demand efficient algorithms: Matrix Profile operates with O(n log n) complexity for time-series analysis (Scott et al., 2024), while

Isolation Forest provides O(n log n) outlier detection in high-dimensional spaces (Liu et al., 2008). Syslog messages provide temporal correlation signals that confirm anomalies detected in either pipeline.

The system uses Matrix Profile for time-series anomaly detection, Isolation Forest for unsupervised pattern recognition, and multi-modal fusion to combine signals from various sources. Matrix Profile is a data structure and algorithm that computes pairwise distances between all subsequences within a time series, enabling efficient identification of patterns (motifs) and anomalies (discords). By analyzing subsequences that deviate significantly from normal patterns, Matrix Profile provides unsupervised anomaly detection without requiring labeled training data (Scott et al., 2024). This approach is particularly effective for detecting subtle BGP routing anomalies that may indicate network failures or attacks. Isolation Forest complements Matrix Profile by providing anomaly detection in high-dimensional feature spaces through an ensemble of random decision trees that isolate anomalous data points. The algorithm operates on the principle that anomalies are easier to isolate than normal points, requiring fewer splits in the decision trees and thus having shorter path lengths from root to leaf (Liu et al., 2008). This makes Isolation Forest computationally efficient for detecting outliers in multi-dimensional feature vectors extracted from SNMP metrics without assuming specific data distributions. For syslog message analysis, deep learning approaches such as DeepSyslog demonstrate high effectiveness in detecting sequential anomalies through LSTM-based pattern recognition combined with sentence embeddings, achieving precision and recall rates exceeding 97% on benchmark datasets (Zhou et al., 2022). Device role mapping provides topology awareness, categorizing network elements by function for context-aware anomaly localization.

Figure 1 illustrates the dual-pipeline ML architecture showing how BGP and SNMP data flow through specialized detection algorithms before multi-modal fusion.

2.2 Why This Topic Was Chosen

This topic was chosen due to fundamental challenges in network operations where traditional monitoring systems often produce many false positives yet miss critical anomalies (Skazin, 2021). Modern network architectures, such as large-scale BGP-routed environments with anycast services and overlay networks, surpass the capabilities of threshold-based alerting systems, requiring more sophisticated approaches to anomaly detection and failure localization.

Network operations centers managing enterprise-scale infrastructures face several critical challenges. Alert fatigue from excessive false positives leads to decreased operator responsiveness, while the lack of contextual information makes it difficult to distinguish routine network variations from genuine anomalies requiring immediate attention. Manual correlation of events across thousands of devices becomes impractical, resulting in extended mean time to resolution (MTTR) for incidents. Machine learning approaches offer the potential to address these challenges through automated pattern recognition and anomaly detection (Mohammed et al., 2021).

Consider a representative use case: a large enterprise network with multiple data centers connected via BGP-routed fabric experiences an intermittent link failure that causes route flapping. Traditional monitoring generates hundreds of alerts from affected routers, but provides no indication of the failure's root cause or scope. Network engineers must manually correlate BGP update logs, SNMP interface counters, and syslog messages across dozens of devices to identify the failing link and assess impact on services. This manual investigation process can take extended periods during critical outages, contributing to increased mean time to resolution (MTTR). An ML-based system using Matrix Profile analysis can automatically detect the anomalous pattern in BGP update frequency, correlate it with SNMP interface error counters, and use topology awareness to localize the failure to the specific failing link, significantly reducing investigation time and improv-

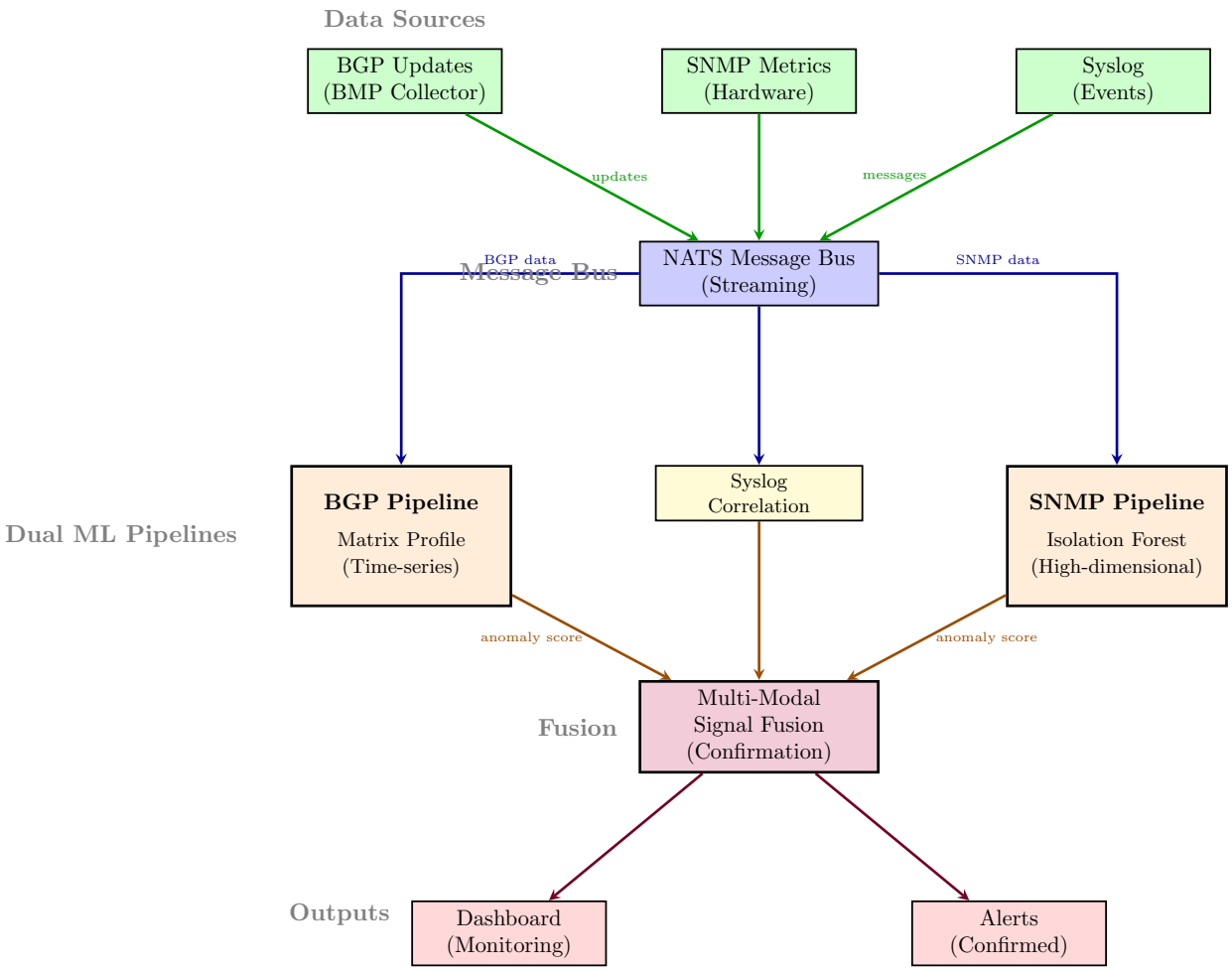

Figure 1: Dual-Pipeline Architecture: BGP updates processed by Matrix Profile (time-series), SNMP metrics by Isolation Forest (high-dimensional), with syslog correlation. Multi-modal fusion confirms anomalies across sources before alerting.

ing operational efficiency (Mohammed et al., 2021).

The selection of Matrix Profile and Isolation Forest as core algorithms reflects specific technical requirements. Matrix Profile provides efficient subsequence similarity analysis for time-series data without requiring labeled training examples, making it suitable for detecting novel anomaly patterns in BGP update streams. The algorithm computes a distance profile that identifies subsequences with unusual characteristics relative to the rest of the time series, with computational complexity of O(n log n) enabling real-time processing. Isolation Forest complements this approach by providing unsupervised anomaly detection in high-dimensional feature spaces extracted from SNMP metrics. By constructing random decision trees that isolate anomalous points with fewer splits than normal points, Isolation Forest efficiently identifies outliers in multi-modal feature vectors without assuming specific distribution patterns. For syslog analysis, while deep learning methods such as DeepSyslog have demonstrated superior performance on log anomaly detection benchmarks with precision and recall both exceeding 97% through LSTM-based sequential pattern recognition (Zhou et al., 2022), the current implementation utilizes pattern-based correlation to maintain system simplicity and real-time processing requirements. This approach allows the syslog component to provide temporal

correlation signals for multi-modal fusion while keeping computational overhead minimal. Together, these techniques enable robust unsupervised anomaly detection across diverse data sources, with the architecture designed to accommodate future integration of LSTM-based syslog analysis for enhanced sequential pattern detection.

3 Problem Description

3.1 Detailed Content Around the Problems Being Solved

The problem I am trying to solve is the inadequacy of traditional network monitoring systems in detecting and localizing network anomalies and failures in large-scale, complex environments. This project addresses alert fatigue from false positives, delayed failure detection, and insufficient context for assessing failure scope and impact. Traditional SNMP threshold alerts and syslog pattern matching produce excessive benign alerts and miss subtle anomalies, while conventional log analysis lacks the sophistication needed for modern networks (Allagi & Rachh, 2019). Research has shown that among SNMP-MIB groups, the Interface and IP groups are most affected by various failure types and anomalies, while ICMP, TCP, and UDP groups are less impacted (Manna & Alkasassbeh, 2019), highlighting the need for targeted monitoring strategies that can quickly identify scope and severity.

These problems arise from the complexity of modern network architectures and the limits of traditional monitoring. Large networks have thousands of devices with various failure modes, such as hardware issues, environmental factors, and routing anomalies. Current systems lack topology awareness and multi-modal data correlation, making it hard to separate normal variations from true anomalies.

These issues vary by network environment but are common in large-scale BGP-routed networks with anycast services and overlay technologies. Enterprise networks requiring dedicated network operations centers (NOCs) with multiple engineers often span thousands of devices across campus and data center environments, making manual correlation across devices and services impractical (Skazin, 2021).

Solving these problems is urgent as they affect network reliability and efficiency. Detection delays extend resolution times, alert fatigue risks missed critical issues, and lacking automated correlation means failures are often found only after escalating to service-impacting levels.

3.2 Current Network Monitoring Limitations

Current network monitoring systems exhibit fundamental limitations in their design and operational effectiveness. Traditional approaches based on SNMP threshold monitoring and syslog pattern matching can detect hard failures such as interface down events or device unreachability, but they produce excessive alerts for benign events while simultaneously missing subtle anomalies that may indicate developing problems. Research has shown that conventional log analysis methods lack the sophistication needed for modern network environments, with traditional rule-based systems struggling to adapt to the dynamic nature of large-scale networks (Allagi & Rachh, 2019). As a result, operations teams experience alert fatigue from false positives while critical anomalies go undetected until they escalate to service-impacting failures (Mohammed et al., 2021).

The insufficiency of current monitoring systems manifests in several specific ways. First, threshold-based alerting requires manual configuration of acceptable ranges for each monitored metric, but optimal thresholds vary based on device role, time of day, and network load patterns. Static thresholds either trigger excessive false positives during normal traffic variations or fail to

detect anomalies that remain below configured thresholds. Second, current systems lack correlation capabilities across different data sources. A gradual increase in BGP update frequency combined with rising SNMP interface error counters and specific syslog error patterns may collectively indicate an impending link failure, but traditional monitoring evaluates each metric independently without cross-modal correlation. Third, existing systems provide no topology awareness or understanding of failure propagation patterns. When a core aggregation switch experiences issues, downstream devices also generate alerts, but traditional monitoring cannot distinguish the root cause from cascading effects (Skazin, 2021).

The scale and complexity of modern networks intensify these monitoring challenges. Large BGP-routed environments often include thousands of devices, anycast services, and global VXLAN overlays, introducing multiple failure modes that demand varied detection and response strategies. Studies of SNMP-MIB data have revealed that among all monitored groups, the Interface and IP groups are most affected by various failure types and anomalies, while ICMP, TCP, and UDP groups show less sensitivity to network issues (Manna & Alkasassbeh, 2019). This differential sensitivity requires intelligent monitoring that can focus on relevant metrics rather than treating all data sources equally.

Research demonstrates that machine learning approaches can significantly improve upon traditional methods. Classifiers such as Random Forest and Decision Trees applied to SNMP-MIB datasets have achieved high accuracy in identifying network failures, suggesting that supervised and unsupervised learning can extract patterns from network telemetry that static thresholds cannot capture (Manna & Alkasassbeh, 2019). Furthermore, the integration of multiple data modalities through feature selection and correlation enables more comprehensive anomaly detection than any single data source alone (Feltin et al., 2023). These findings motivate the development of ML-based monitoring systems that can adapt to network dynamics, correlate multi-modal signals, and provide topology-aware failure localization.

4 Analysis

4.1 Laboratory Environment and Testing Infrastructure

The system is validated using a Containerlab-based network environment that provides realistic BGP sessions and network topology for testing the dual-pipeline anomaly detection architecture.

Figure 2 shows the laboratory environment used for testing and validation, consisting of containerized FRR routers running authentic BGP sessions, integrated with data collectors and the dual-pipeline ML detection system.

[SECTION PLACEHOLDER - Additional analysis content to be completed]

This section will include analysis of how the topic was chosen, how problems were identified, how solutions fulfill operational and strategic goals, implementation timeline, and current implementation status.

5 Research

[SECTION PLACEHOLDER - TO BE COMPLETED]

This section will include coursework background and supporting references that provide the research foundation for the project.

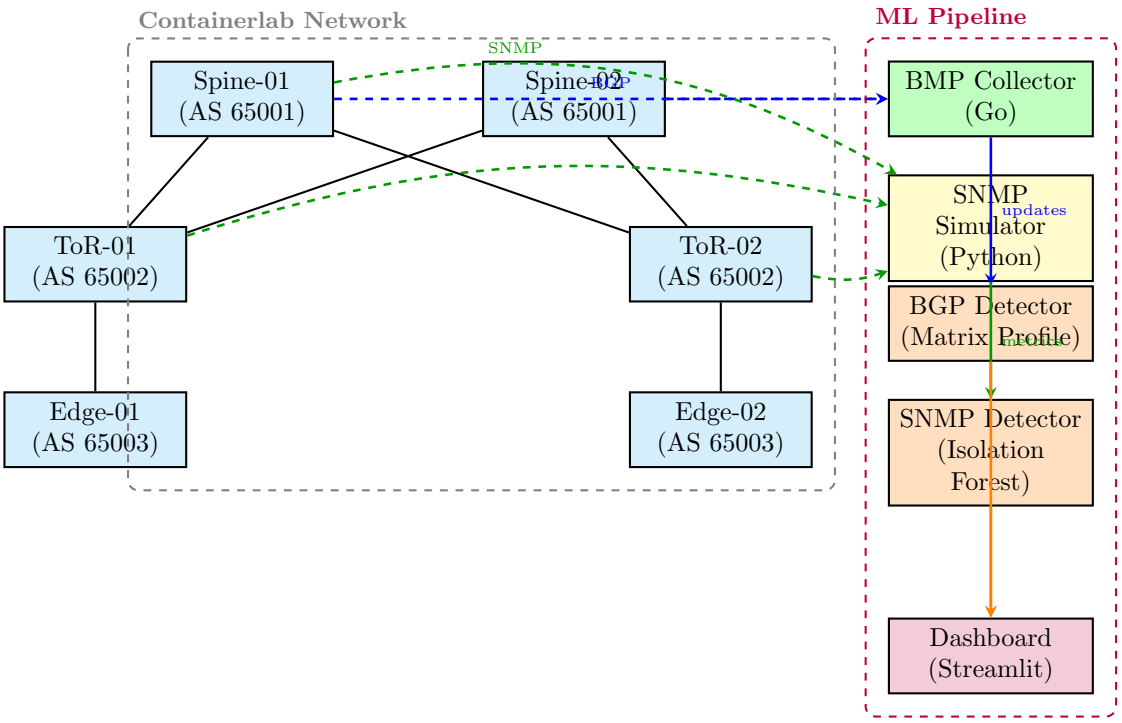

Figure 2: Laboratory Test Environment: Containerlab network with 6 FRR routers (2 spine, 2 ToR, 2 edge) running real BGP sessions. BMP collector captures routing updates, SNMP simulator generates hardware metrics. Both feed the dual-pipeline ML system for validation testing.

6 References

References

[1] Cheng, M., Li, Q., Lv, J., Liu, W., & Wang, J. (2021). Multi-Scale LSTM Model for BGP Anomaly Classification. *IEEE Transactions on Services Computing*, 14(3), 765–778. Available at: https://doi.org/10.1109/TSC.2018.2824809

[2] Mohammed, S. A., Mohammed, A. R., Côté, D., & Shirmohammadi, S. (2021). A machine-learning-based action recommender for Network Operation Centers. *IEEE Transactions on Network and Service Management*, 18(3), 2702–2713. Available at: https://doi.org/10.1109/TNSM.2021.3095463

[3] Scott, B., Johnstone, M. N., Szewczyk, P., & Richardson, S. (2024). Matrix Profile data mining for BGP anomaly detection. *Computer Networks*, 242, 110257.

[4] Tan, Y., Huang, W., You, Y., Su, S., & Lu, H. (2024). Recognizing BGP Communities Based on Graph Neural Network. *IEEE Network*, 38(6), 232–238. Available at: https://doi.org/10.1109/MNET.2024.3414113

[5] Allagi, S., & Rachh, R. (2019). Analysis of Network log data using Machine Learning. *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, 1–3. Available at: https://doi.org/10.1109/I2CT45611.2019.9033528

[6] Skazin, A. (2021). Detection of network anomalies in log files. *IOP Conference Series: Materials Science and Engineering*, 1069(1), 012021. Available at: https://doi.org/10.1088/1757-899X/1069/1/012021

[7] Feltin, T., Cordero Fuertes, J. A., Brockners, F., & Clausen, T. H. (2023). Understanding Semantics in Feature Selection for Fault Diagnosis in Network Telemetry Data. *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, 1–9. Available at: https://doi.org/10.1109/NOMS56928.2023.10154455

[8] Wang, H. (2020). Improvement and implementation of Wireless Network Topology System based on SNMP protocol for router equipment. *Computer Communications*, 151, 10–18. Available at: https://doi.org/10.1016/j.comcom.2020.01.001

[9] Manna, A., & Alkasassbeh, M. (2019). Detecting network anomalies using machine learning and SNMP-MIB dataset with IP group. *arXiv preprint arXiv:1906.00863*. Available at: https://arxiv.org/abs/1906.00863

[10] Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. Available at: https://doi.org/10.1109/ICDM.2008.17

[11] Zhou, J., Qian, Y., Zou, Q., Liu, P., & Xiang, J. (2022). DeepSyslog: Deep Anomaly Detection on Syslog Using Sentence Embedding and Metadata. *IEEE Transactions on Information Forensics and Security*, 17, 3051–3066. Available at: https://doi.org/10.1109/TIFS.2022.3198188