

Proceeding Paper

Bridging the Data Gap in ML-Based NIDS: An Automated Honeynet Platform for Generating Real-World Malware Traffic Datasets [†]

Gabriel Ulloa Cano ^{1,*}, Gabriel Sánchez Pérez ¹, José Portillo-Portillo ¹, Linda Karina Toscano Medina ¹, Aldo Hernández Suárez ¹, Jesús Olivares Mercado ¹, Héctor Manuel Pérez Meana ¹, Luis Javier García Villalba ² and Pablo Velarde Alvarado ³

¹ Instituto Politécnico Nacional (National Polytechnic Institute), ESIME Culhuacán, Mexico City 04440, Mexico; gasanchezp@ipn.mx (G.S.P.); jportillo@ipn.mx (J.P.-P.); ltoscano@ipn.mx (L.K.T.M.); alhernandezsu@ipn.mx (A.H.S.); jolivares@ipn.mx (J.O.M.); hmperezm@ipn.mx (H.M.P.M.)

² Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Faculty of Computer Science and Engineering, Complutense University of Madrid (UCM), 28040 Madrid, Spain; javierv@ucm.es

³ Unidad Académica de Ciencias Básicas e Ingenierías, Universidad Autónoma de Nayarit, Tepic 63000, Mexico; pvelarde@uan.edu.mx

* Correspondence: gulloac2100@alumno.ipn.mx

[†] Presented at the First Summer School on Artificial Intelligence in Cybersecurity, Cancun, Mexico, 3–7 November 2025.

Abstract

The effectiveness of Machine Learning (ML)-based Network Intrusion Detection Systems (NIDS) is critically hampered by the scarcity of realistic and up-to-date malware traffic datasets. To address this gap, we present an automated platform for generating real-world malware traffic datasets. Our solution leverages a production-environment honeynet (T-Pot), deployed within a university network and segmented via a secure WireGuard VPN, to capture live attacks using high-interaction honeypots (Dionaea, Cowrie, ADBhoney). A fully automated pipeline handles traffic capture, transfer, filtering based on honeypot logs, and malware analysis (VirusTotal, VxAPI). The output is the IPN-UAN-23 dataset—a curated, labeled corpus of malicious network traffic. This platform functions as a vital automated security tool, providing the continuous stream of actionable intelligence required to develop and refine robust ML-based NIDS within a DevSecOps lifecycle.

Keywords: malware traffic generation; automated security tool; high-interaction honeypot; ML-based NIDS; network security dataset; DevSecOps



Academic Editor: Ana Lucila Sandoval Orozco

Published: 13 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The efficacy of Machine Learning-based Network Intrusion Detection Systems (NIDS) is critically limited by their reliance on outdated or synthetic datasets, which fail to represent the evolving threat landscape and lead to poor performance against novel attacks [1]. This data gap is particularly problematic within the DevSecOps paradigm, which requires automated, continuous security testing but lacks tools for generating realistic threat intelligence.

To bridge this gap, this work presents an automated platform for generating real-world malware traffic datasets. Functioning as an automated security tool, our system leverages a high-interaction honeynet to capture live malware campaigns and their full network behavior within a production-like environment. The platform automates the entire

process from traffic capture to preliminary labeling, providing holistic, behavioral data for refining NIDS algorithms. By supplying fresh data for continuous retraining, it offers a sustainable solution to data obsolescence, enabling more adaptive defenses aligned with DevSecOps principles.

Recent developments (2022–2025) in ML-based NIDS have emphasized the need for adaptive datasets that reflect evolving attack patterns [2]. Collaborative platforms for dataset generation have emerged as promising solutions to address data scarcity while maintaining ethical standards in security research [3].

2. Materials and Methods

The core of our contribution is an automated platform deployed across two geographical locations—a Data Center and a Laboratory—interconnected via the university’s WAN and a secure WireGuard VPN [4]. The infrastructure is virtualized on Proxmox VE clusters [5].

2.1. Automated Workflow and Data Processing

The platform’s operation is a fully automated, cyclic process orchestrated by custom scripts (Figure 1):

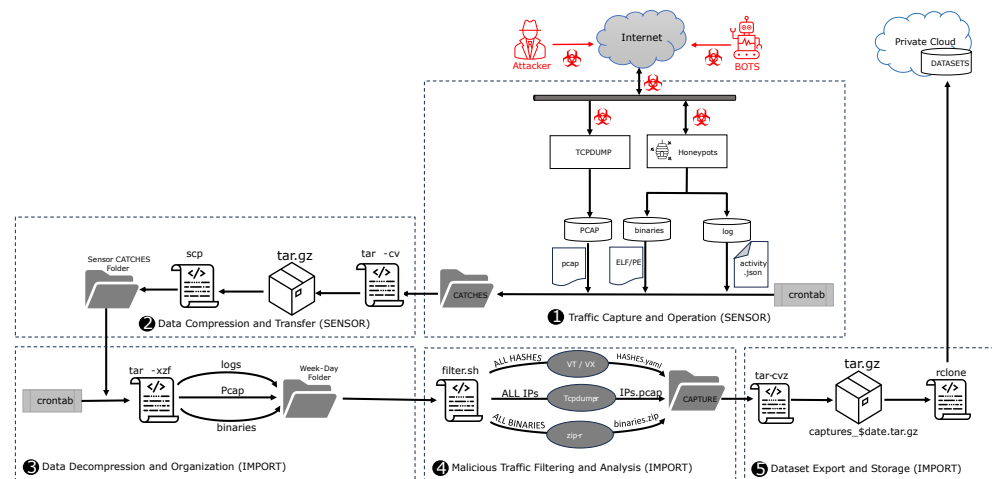


Figure 1. Flow diagram of the fully automated data processing pipeline.

1. **Traffic Capture (SENSOR):** Upon reboot, the node automatically launches the honeypots and initiates `tcpdump`, capturing all incoming traffic for a set period (e.g., 24 h).
2. **Data Compression and Transfer (SENSOR):** Pre-reboot, a script stops the capture, consolidates the data (PCAPs, logs, binaries) into a folder, compresses it, and transfers the archive to the IMPORT node via SCP over the VPN.
3. **Data Decompression and Organization (IMPORT):** The received archive is decompressed into a date-stamped directory for processing.
4. **Malicious Traffic Filtering and Analysis (IMPORT):** A key processing step. Using honeypot logs containing malicious IPs, a script filters the large PCAP to extract only flows related to these IPs, creating smaller, curated PCAPs. Concurrently, malware binary hashes are queried against VirusTotal and VxAPI for threat intelligence labeling.
5. **Dataset Export and Storage (IMPORT):** The filtered PCAPs, enriched logs, and binaries are compressed into a final dataset archive and uploaded to the private cloud storage. The SENSOR node reboots, cleansing its state for the next cycle.

2.2. Implementation

Orchestration and data processing scripts were developed in Bash 5.0.3 and Python 3.7.3 on Debian 10 (Buster), leveraging tools like `tcpdump`, `scp`, `jq`, and the VirusTotal/VxAPI

CLIs for a fully automated pipeline with minimal manual intervention. The complete source code and deployment scripts are available in the public GitHub repositories [6,7].

3. Results

The implemented automated platform operated continuously, successfully generating the IPN-UAN-23 dataset. System operation was robust, with interruptions only for scheduled maintenance.

3.1. Dataset Overview and Threat Analysis

The platform generated 12 high-quality PCAP traces. Key quantitative metrics are summarized in Table 1. The dataset comprises approximately 15.8 million packets and 516,000 netflows, with a total volume of 5.9 GB. The average capture duration was 23 h, demonstrating system stability for long-term operation.

Table 1. Key characteristics of the generated IPN-UAN-23 malware traffic dataset.

Capture ID	Duration (hrs)	Packets (k)	Size (MB)	Primary Threats
11-29	9.00	584	292	WannaCry, Downloaders Mirai, Miners Mirai variants
11-30	22.98	1123	438	
12-01	22.98	1135	458	
...	...	...	...	...
12-14	22.98	1459	563	Aidra, Gafgyt
Total/Avg.	23.0	15,800	5900	

Qualitative analysis revealed a significant prevalence of trojan malware, particularly from the Mirai botnet family and its variants (e.g., Gafgyt, Bashlite). The automated pipeline extracted 130 unique malicious objects. Crucially, threat intelligence enrichment (VirusTotal, VxAPI) identified 29 potential new variants of trojans and downloaders not widely recognized in existing databases at the time of capture.

A key finding was the identification of a previously unreported Mirai botnet variant. This discovery was automatically reported via integrated APIs, contributing actionable intelligence to the security community and underscoring the platform’s value in identifying evolving threats near the zero-day attack vector.

3.2. External Validation Using TrafficColab Platform

To demonstrate the practical utility of the generated datasets, the IPN-UAN-23 corpus was independently validated using TrafficColab [8], a web-based collaborative platform for network traffic analysis. Five supervised classifiers were evaluated across four different scenarios derived from the TrafficColab_IPN_UAN_2025 dataset: Weighted Logistic Regression (W-LR), Weighted Decision Tree (W-DT), Logistic Regression with SMOTE (LR-SMOTE), Support Vector Machines with One-Sided Selection (SVM-OSS), and Extreme Gradient Boosting (XGB).

As shown in Figure 2, all classifiers achieved consistently high performance, with macro-F1 scores and Matthews Correlation Coefficient (MCC) values exceeding 0.90 across all scenarios. The XGB and W-DT models demonstrated superior performance, achieving the highest MCC scores, which indicates their enhanced capability to handle class imbalance in intrusion detection tasks.

The experimental results confirm that datasets generated by our automated platform enable training of highly effective intrusion detection systems. The consistently high MCC values (≥ 0.95 for top performers) demonstrate robust performance in realistic, imbalanced class scenarios, outperforming benchmarks on traditional public datasets. The superiority

of XGB and W-DT highlights the importance of selecting appropriate algorithms for real-world NIDS deployment where class imbalance is prevalent.

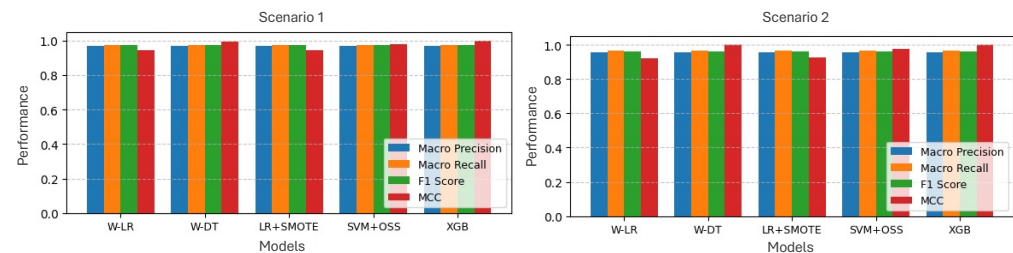


Figure 2. Performance comparison of ML classifiers across different scenarios on the TrafficColab_IPN_UAN_2025 dataset. XGB and W-DT show consistent superiority in MCC metrics.

4. Discussion

The results confirm that the automated platform successfully bridges the critical data gap for ML-based NIDS, providing a continuous stream of realistic, labeled malware traffic.

The quantitative metrics (Table 1) demonstrate the system's ability to generate substantial volumes of diverse real-world data with minimal intervention. The capture of widespread threats like Mirai and ransomware confirms the honeynet's effectiveness. Most significantly, the identification of 29 novel malware variants proves the platform functions as an early-warning system, generating actionable intelligence for adapting NIDS to emerging threats—a core DevSecOps requirement.

Unlike static datasets (e.g., NSL-KDD, CIC-IDS) that offer snapshots of past threats [9,10], the value of the IPN-UAN-23 dataset lies in its timeliness, realism, and continuous generation capability. Our platform provides a sustainable model for maintaining an evolving data resource, moving beyond manual curation.

The experimental validation confirms the practical utility of the platform-generated datasets for real-world NIDS deployment. The TrafficColab framework establishes ethical guidelines and compliance standards for responsible data generation and distribution, addressing privacy and regulatory concerns inherent to honeynet deployment.

This work directly addresses the need for automated security tools in the SDLC. The platform can act as a dedicated stage in a CI/CD pipeline for security models, where periodically pulling the latest dataset to retrain NIDS ensures detection capabilities evolve with the threat landscape, transforming security into a continuous feedback loop.

A current limitation is the focus on network traffic analysis. Future work will involve applying deep learning models to this curated dataset to develop next-generation NIDS. Expanding the diversity of emulated services on the honeynet to attract a broader attack range is also planned.

5. Conclusions

This paper presented the design and implementation of an automated platform for generating realistic malware traffic datasets. The system leverages a production honeynet and a fully automated pipeline to capture, filter, enrich, and store network traffic from live malware campaigns. The operational results, culminating in the IPN-UAN-23 dataset, demonstrate the platform's effectiveness. It not only captures high-volume traffic and identifies 29 novel malware variants, but also enables training of high-performance ML models—with XGB and W-DT classifiers achieving MCC scores ≥ 0.95 in external validation. By automating the entire process, this work offers a sustainable solution to the data obsolescence problem and serves as a foundational tool for enabling robust, adaptive, and continuous security within a modern DevSecOps framework.

Author Contributions: Conceptualization, G.U.C., G.S.P. and P.V.A.; methodology, G.U.C. and A.H.S.; software, G.U.C. and J.P.-P.; validation, G.U.C., L.J.G.V. and P.V.A.; formal analysis, G.U.C.; investigation, G.U.C. and G.S.P.; resources, P.V.A.; data curation, J.O.M. and H.M.P.M.; writing—original draft preparation, G.U.C.; writing—review and editing, G.U.C., G.S.P. and A.H.S.; visualization, L.K.T.M.; supervision, P.V.A.; project administration, P.V.A.; funding acquisition, P.V.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable for studies not involving humans or animals.

Informed Consent Statement: Not applicable for studies not involving humans.

Data Availability Statement: The primary, curated malware interaction datasets generated during this study are available in the Zenodo repository under restricted access due to the presence of active malware and sensitive network data. The persistent identifier is: <https://doi.org/10.5281/zenodo.18512027>. Access can be requested via the Zenodo platform and is subject to a formal data use agreement. The raw PCAP captures are retained privately under institutional protocols. To enable full reproducibility, the complete open-source software platform for generating such datasets is publicly available at: <https://github.com/Gabriel-Ulloa/SENSOR> and <https://github.com/Gabriel-Ulloa/IMPORT>, all accessed on 25 October 2025.

Acknowledgments: The authors would like to thank the organizers of the First Summer School on Artificial Intelligence in Cybersecurity for providing the academic framework that inspired this research.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NIDS	Network Intrusion Detection System
ML	Machine Learning
DevSecOps	Development, Security, and Operations
PCAP	Packet Capture
VPN	Virtual Private Network

References

1. Ji, I.H.; Lee, J.H.; Kang, M.J.; Park, W.J.; Jeon, S.H.; Seo, J.T. Artificial Intelligence-Based Anomaly Detection Technology over Encrypted Traffic: A Systematic Literature Review. *Sensors* **2024**, *24*, 898. [CrossRef] [PubMed]
2. Ulloa-Cano, G. Diseño e Implementación de una Plataforma para la Generación de Datasets de Malware en tráfico Real Mediante una HoneyNet. [Master's Thesis, National Polytechnic Institute]. Repositorio Digital Institucional (RDI). 2023. Available online: <https://www.repositoriodigital.ipn.mx/> (accessed on 14 November 2024).
3. Velarde-Alvarado, P.; Gonzalez, H.; Martínez-Peláez, R.; Mena, L.J.; Ochoa-Brust, A.; Moreno-García, E.; Félix, V.G.; Ostos, R. A Novel Framework for Generating Personalized Network Datasets for NIDS Based on Traffic Aggregation. *Sensors* **2022**, *22*, 1847 [CrossRef] [PubMed]
4. pfSense® Software Configuration Recipes—WireGuard Site-to-Site VPN Configuration Example | pfSense Documentation. Available online: <https://docs.netgate.com/pfsense/en/latest/recipes/wireguard-s2s.html> (accessed on 29 November 2023).
5. Proxmox VE Administration Guide. Available online: <https://pve.proxmox.com/pve-docs/pve-admin-guide.html> (accessed on 28 November 2023).
6. Gabriel-Ulloa/SENSOR: A Honeypot System Capturing Suspicious Network Traffic in a T-Pot Installation. Available online: <https://github.com/Gabriel-Ulloa/SENSOR> (accessed on 28 October 2025).
7. Gabriel-Ulloa/IMPORT: Processes and Filters Malicious Traffic Data Captured by SENSOR. Available online: <https://github.com/Gabriel-Ulloa/IMPORT> (accessed on 28 October 2025).
8. Rodríguez-Adame, O.G. Generación de Data Sets Etiquetados para Diseño de Sistemas de Detección de Intrusos en Red. [Master's Thesis, National Polytechnic Institute]. Repositorio Digital Institucional (RDI). 2025. Available online: <https://www.repositoriodigital.ipn.mx/> (accessed on 25 October 2025).

9. NSL-KDD | Datasets | Research | Canadian Institute for Cybersecurity | UNB. Available online: <https://www.unb.ca/cic/datasets/nsl.html> (accessed on 5 December 2023).
10. Intrusion Detection Evaluation Dataset (CIC-IDS2017). Available online: <https://www.unb.ca/cic/datasets/ids-2017.html> (accessed on 5 December 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.