Automated Labeled Dataset Generation for Training Healthcare Network Intrusion Detection Systems

Miro Moffett [1] **Viktor Schlegel** [1]

[1] *Imperial Global Singapore*. Correspondence to: Miro Moffett mgm20@ic.ac.uk.

1. Background

Healthcare systems face increasing cybersecurity threats due to the proliferation of connected medical devices and complex interoperability requirements [1]. Securing medical networks prove particularly difficult due to critical healthcare infrastructure often being comprised of legacy devices. Overall, this has led to the adoption of network level solutions as the primary method of security [2, 3].

Machine learning-based Network Intrusion Detection Systems (NIDS) and Network Intrusion Prevention Systems (NIPS) require large volumes of diverse, accurately labeled training data to achieve robust performance. However, obtaining realistic labeled network traffic from healthcare environments presents significant challenges. Healthcare networks contain specialized protocols (DICOM, HL7), diverse medical device communications, and complex multi-departmental workflows that are poorly represented in existing public datasets [4].

Furthermore, privacy regulations and operational constraints severely limit the availability of real hospital network traffic for ML research. Existing healthcare security datasets often lack protocol diversity, realistic temporal patterns, and comprehensive attack scenario coverage, limiting the generalisability of trained models. Current approaches to dataset generation through manual capture or synthetic generation fail to scale and struggle to maintain clinical authenticity across diverse hospital workflows [5].

2. Objectives

This work presents Chicomoztac, an automated pipeline for generating labeled training datasets for healthcare NIDS/NIPS systems. The primary objectives are to: (1) automate the generation of diverse, realistic healthcare network traffic across multiple hospital departments; (2) provide automatic ground-truth labeling for both benign and malicious traffic patterns; (3) ensure clinical authenticity by deriving traffic patterns from real patient medical records; and (4) enable scalable generation of training datasets with configurable attack scenarios for supervised learning.

3. Methodology

Chicomoztac employs a three-stage automated data generation pipeline. The first stage processes real clinical data from the MIMIC-IV FHIR database [6] (1,817,172 clinical events), converting FHIR resources into structured patient workflows. This ensures that generated network traffic reflects authentic clinical patterns, temporal relationships, and departmental interactions observed in real hospital operations.

The second stage transforms patient workflows into protocol-specific network traffic across four simulated hospital wings: ICU (vital signs monitoring, smart infusion pumps), radiology (DICOM imaging workflows), administration (HL7 ADT messages, EHR queries), and pharmacy (automated dispensing cabinets, medication orders via HL7 ORM/RDS/RAS messages). The distributed simulation architecture generates traffic at realistic volumes (up to 1 Gb/s throughput) with proper timing characteristics derived from actual patient encounter sequences.

The third stage provides automatic ground-truth labeling for generated traffic. Benign traffic is labeled by departmental source, protocol type, and clinical event category. Attack traffic is generated by injecting known attack patterns (network reconnaissance, protocol manipulation, unauthorized access, data exfiltration) at configurable points in patient workflows, providing precise labels for supervised learning. The Kubernetes-based infrastructure enables parallel generation of diverse scenarios for comprehensive training set coverage.

4. Results and Discussion

The automated pipeline successfully generates labeled training data covering 42% of clinical events from MIMIC-IV (772,298 of 1,817,172 events). Table 1 presents dataset composition by clinical event type, demonstrating diversity across multiple healthcare workflows.

Table 1: Generated dataset composition by clinical event type

Event Type	Clinical Events	Generated Flows	Coverage
Observations (vitals/labs)	1,627,080	585,749	36%
Medication administration	113,070	90,456	80%
Medication orders	35,104	33,349	95%
Specimens	24,916	24,916	100%
Conditions (diagnoses)	10,102	10,102	100%
Procedures (imaging)	6,900	920	13%
Total	**1,817,172**	**772,298**	**42%**

Protocol validation confirmed that generated traffic matches clinical specifications: DICOM C-FIND/C-STORE/C-GET operations follow imaging workflow patterns, HL7 v2.x messages (ORM, RDS, RAS, ORU) adhere to standard formats, and temporal patterns reflect realistic departmental interactions. The pharmacy wing implementation significantly enhanced dataset diversity, contributing 147,846 additional labeled flows (8.1% coverage gain) representing medication dispensing workflows, automated cabinet operations, and infusion pump communications.

Attack scenario injection testing validated four common attack patterns from benchmark datasets [7], demonstrating the platform's capability to generate labeled malicious traffic with known ground-truth for supervised learning. Each attack can be precisely timed within patient workflows, enabling generation of training data where attacks occur during specific clinical operations (e.g., during medication dispensing, imaging procedures, or vital sign monitoring).

5. Conclusions

This work presents an automated pipeline for generating labeled healthcare network traffic datasets suitable for training ML-based NIDS/NIPS systems. By leveraging real clinical records from MIMIC-IV, the approach ensures that generated traffic reflects authentic hospital workflows while providing automatic ground-truth labels for supervised learning. The automated conversion from FHIR clinical data to labeled network flows addresses critical data scarcity challenges in healthcare security ML research.

The current pipeline generates 772,298 labeled flows across diverse clinical scenarios, providing training data spanning multiple hospital departments, protocols (DICOM, HL7 v2.x, proprietary medical device protocols), and both benign and attack traffic patterns. The automated, scalable approach enables rapid generation of large training datasets with configurable characteristics, supporting diverse ML model architectures and training requirements.

Future enhancements will expand dataset coverage through improved event classification patterns (projected 181,717–272,576 additional flows) and implementation of EHR clinical documentation workflows (projected 454,293–545,152 additional flows), potentially reaching 80–85% coverage (1,453,738–1,544,596 labeled flows). These expansions will further improve training data diversity for healthcare NIDS/NIPS systems, enabling more robust ML model development for medical IoT security.

Acknowledgments

References

[1] J. Cawthra, B. Hodges, J. Kuruvilla, K. Littlefield, B. Niemeyer, C. Peloquin, S. Wang, R. Williams, and K. Zheng. Securing picture archiving and communication system (pacs) cybersecurity for the healthcare sector. Technical report, 2020.

[2] F. Hussain, S. G. Abbas, G. A. Shah, I. M. Pires, U. U. Fayyaz, F. Shahzad, N. M. Garcia, and E. Zdravevski. A framework for malicious traffic detection in iot healthcare environment. *Sensors*, 21(9):3025, 2021.

[3] George Hatzivasilis, Othonas Soultatos, Sotiris Ioannidis, Christos Verikoukis, Giorgos Demetriou, and Christos Tsatsoulis. Review of security and privacy for the internet of medical things (iomt). In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 457–464, 2019.

[4] Maxime Lanvin, Pierre-François Gimenez, Yu Han, Frédéric Majorczyk, Ludovic Mé, and Éric Totel. Errors in the cicids2017 dataset and the significant differences in detection performances it makes. In *Lecture Notes in Computer Science*, pages 18–33, 2023.

[5] Mireya Lucia Hernandez-Jaimes, Alfonso Martinez-Cruz, Kelsey Alejandra Ramírez-Gutiérrez, and Claudia Feregrino-Uribe. Artificial intelligence for iomt security: A review of intrusion detection systems, attacks, datasets and cloud–fog–edge architectures. *Internet of Things*, 23:100887, 2023.

[6] A. M. Bennett, H. Ulrich, P. van Damme, J. Wiedekopf, and A. E. W. Johnson. Mimic-iv on fhir: converting a decade of in-patient data into an exchangeable, interoperable format. *Journal of the American Medical Informatics Association*, 30(4):718–725, 2023.

[7] S. Dadkhah, E. C. P. Neto, R. Ferreira, R. C. Molokwu, S. Sadeghi, and A. A. Ghorbani. Ciciomt2024: A benchmark dataset for multi-protocol security assessment in iomt. *Internet of Things*, 28:101351, 2024.