

ML to Detect Zero-Day Attacks: An Experiment

Achilleas Spanos
aspanos@uniwa.gr
University of West Attica
Egaleo, Greece

Stavros Derziotis
sderziotis@uniwa.gr
University of West Attica
Egaleo, Greece

Ioanna Kantzavelou
ikantz@uniwa.gr
University of West Attica
Egaleo, Greece

Abstract

While many ML-based intrusion detection studies report impressive in-dataset performance, few examine whether these models remain effective to generalize to unseen attack patterns. To provide insights on this matter, a series of cross-dataset generalization experiments is designed to evaluate models' behavior under fixed conditions that simulate zero-day attacks. These experiments assess the model's ability to recognize structural similarities among distinct attack types, under a practical and reproducible setting, using known features. Nonetheless, we caution against interpreting these results as fully reflective of zero-day attack detection performance.

CCS Concepts

• Security and privacy → Intrusion detection systems.

Keywords

Intrusion Detection Systems, ML, Zero-Day Attacks, Generalization

ACM Reference Format:

Achilleas Spanos, Stavros Derziotis, and Ioanna Kantzavelou. 2018. ML to Detect Zero-Day Attacks: An Experiment. In *Proceedings of (GEC 2026)*. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Detecting zero-day attacks with ML-based approaches requires further research and study to prove their effectiveness. Although optimal detection is not a feasible objective, it is interesting and valuable to explore whether these proposals could identify even a small portion of novel attacks. In this paper, we present an experimental research work with preliminary results that investigates the problem of detecting zero-day attacks when ML-based models are applied. These experiments assess the model's ability to identify structural similarities among distinct attack types in a practical, reproducible setting using known features. The findings reveal significant gaps in generalization across attacks, suggesting interesting future research directions.

2 Experimental Design

Due to the balance between recency and broad adoption in the research community, the CICIDS2017 [1] dataset has been selected.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
GEC 2026, Patra, GR

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

The dataset's dual-level structure includes both raw packet captures and flow characteristics. This enables fair comparisons against emerging techniques and supports multimodal intrusion detection methodologies, combining packet payload with flow-level information. CICIDS2017 is organized into daily capture files, each representing diverse attack scenarios. To examine the models' generalization capabilities on unseen attacks, the daily-file split structure of the CICIDS2017 is leveraged, treating each daily file as a standalone dataset. Since the Monday file of the CICIDS2017 includes only benign data, it is excluded from our experimental study. Similarly, the Thursday afternoon file of the CICIDS2017 dataset enumerates only 36 malicious samples, it is also excluded from the experiment. This setup leverages temporal variation, while also keeping the network environment invariant, allowing the generalization test to unseen attacks, without domain shifts. Although some studies focus exclusively on certain attack types during training or testing (leave-one-attack-out), the presented experimental setup is more broadly representative and generic. To the best of our knowledge, this day-based cross-dataset generalization evaluation has not been explored in the literature, offering a high-level model's generalization evaluation without introducing domain shifts.

Two tree-based supervised ML classifiers are employed: RF, due to its consistently robust performance reported in recent literature [2], and XGB, due to its robustness in handling complex data. To address class imbalance, two sampling techniques have been utilized, namely, SMOTE to synthetically generate minority class instances and avoid original data information loss, and Random Under-Sampling to solely rely on predictions on original observed data. Finally, two scaling techniques are tested: Min-Max scaling and Z-score standardization. Only slight variations in performance were observed between these two normalization techniques. In total, each model, for each day-dataset, was evaluated across all possible configuration combinations. A random state of zero was applied across all possible components to support the reproducibility of the presented experimental component.

Since the nature of the experiment is to highlight the importance of generalization and emphasize behavioral differences, rather than optimized generalization models, they are trained under default settings without finetuning. This allows for an unbiased, consistent, and fair comparison of classifiers' general behavior. Nevertheless, it is acknowledged that this may favor architectures robust to hyperparameters.

To account for class imbalance in the test sets, Random Under-Sampling is applied during evaluation, preserving only the original data. This ensures accuracy remains a meaningful metric. Because of the large number of models and configurations presented, reports include only accuracy and recall as the main performance metrics. Additional metrics are not included in order to avoid burdening with the results of a small-scale experiment.

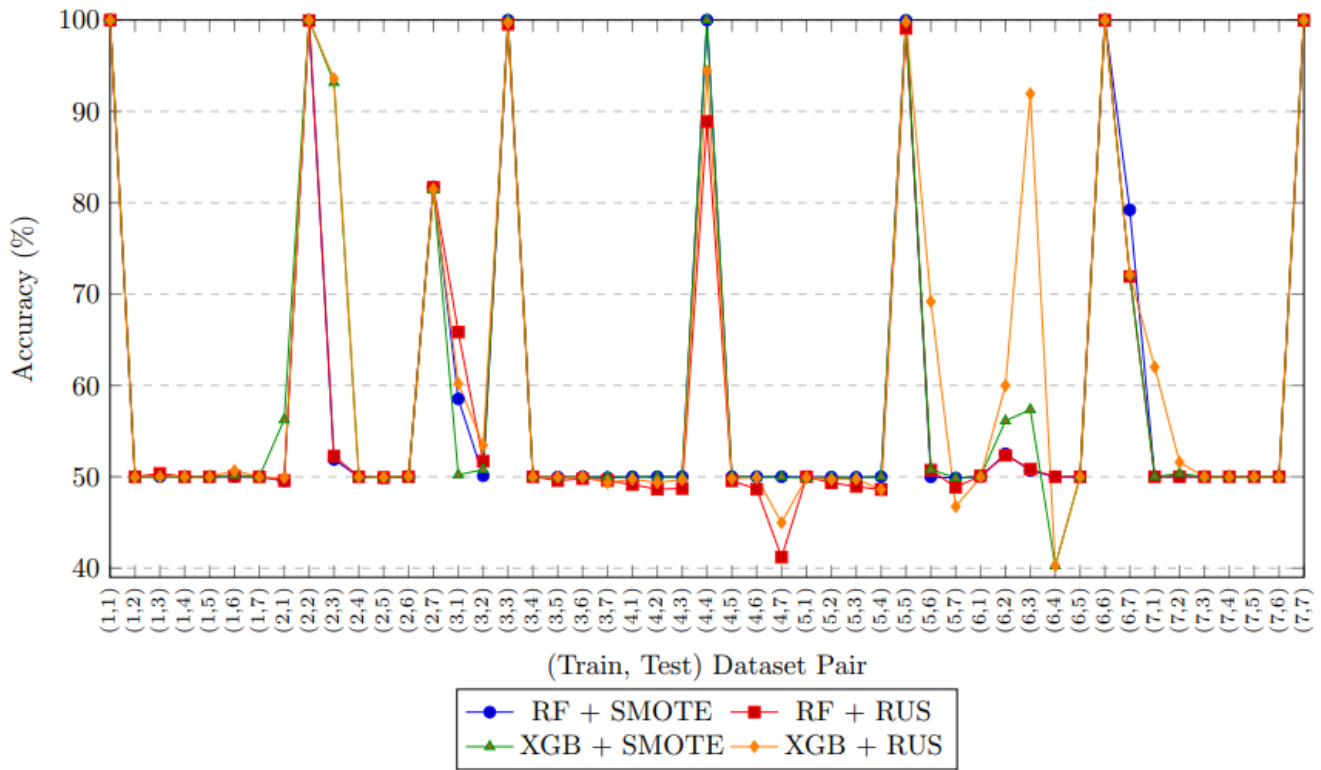


Figure 1: Machine Learning IDS zero-day attack comparison

3 Experimental Results

In order to test the generalization across a variety of network attack scenarios, an end-to-end evaluation framework is constructed. The framework constitutes intra-day and cross-day experiments, which incorporate various sampling methods and feature preprocessing techniques. Each model was trained on a specific day-dataset and tested on others to evaluate its performance while exposed to unseen attack classes. Unseen attacks are defined as those not present in the training distribution, thus representing realistic, previously unencountered threats in a production environment. For the train and intra-day test, a 75-25 split is applied. Figure 1 illustrates the accuracy among four configurations of RF and XGB combined with SMOTE and Random Under-Sampling.

3.1 The illusion of model efficacy

The objective of the experimental component is not to perform optimal detection, but to explore whether the models could identify even a fraction of unseen attacks. Detecting even a small portion of a novel attack is practically valuable, demonstrating a meaningful potential to generalize models.

3.2 Cross-Attack Generalization

In real-world scenarios, the ability of a model to detect previously unseen attacks is paramount. Despite the limitations on cross-day generalization, several specific test cases revealed promising potential. Classifier trained on DoS day-dataset achieved 93.6% accuracy

when tested on Web Attack traffic, and a reasonable 81.48% accuracy when detecting DDoS, since both DoS and DDoS share structural similarities. WebAttacks trained classifier, detected Brute Force with 60.21% accuracy, a plausible generalization since WebAttacks include Brute Force attacks. The bot traffic trained classifier led to 69.19% accuracy when identifying PortScan. Furthermore, the PortScan trained classifier, generalized on WebAttacks (91.93%), DoS (60.23%), and DDoS(72.12%). Finally, the DDoStrained model identified 24.01% of Brute Force attacks, achieving a 62.01% accuracy.

4 Discussion and Future Directions

Our findings align with concept drift and within-domain attack generalization challenges. Even under the consistent structure of CICIDS2017, the introduction of a new traffic degrades model performance. While high same-day accuracy is reported, cross-day evaluation reveals substantial performance drops, highlighting critical gaps in generalization across attacks, reflecting distributional fragility.

References

- [1] Iman Sharafaldin, Arash Habibi Lashkari, Ali A Ghorbani, et al. 2018. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp* 1 (2018), 108–116.
- [2] Mubarak Albarka Umar, Zhanfang Chen, Khaled Shuaib, and Yan Liu. 2025. Effects of feature selection and normalization on network intrusion detection. *Data Science and Management* 8, 1 (2025), 23–39.