Journal of Information Technology, Cybersecurity, and Artificial Intelligence

(Research)

Enhancing Cloud Security through Intrusion Detection: A Comprehensive Study Using the ISOT-CID Dataset

Safana Alzide

Information Technology Programs
Florida State College at Jacksonville
Jacksonville, FL, USA

Journal of Information Technology, Cybersecurity, and Artificial Intelligence, 2024, 1(1), 39–46

Publication history: received October 30, 2024; revised November 1, 2024; accepted November 3, 2024

Article DOI: https://doi.org/10.70715/jitcai.2024.v1.i1.005

Abstract

Cloud computing has revolutionized data management, offering unparalleled scalability, flexibility, and efficiency. However, its open and multi-tenant nature introduces significant security vulnerabilities, making it an attractive target for cyber threats. Intrusion Detection Systems (IDS) tailored for cloud environments are essential in mitigating these risks. Despite various IDS models, benchmarking datasets representing realistic cloud environments is a substantial limitation. This study utilizes the ISOT Cloud Intrusion Detection Benchmark Dataset (ISOT-CID), a publicly available dataset featuring a range of network and application layer attacks collected from a real production cloud environment. The research explores the dataset's structure, analyzes attack patterns, and evaluates IDS models' performance to provide actionable insights for enhancing cloud security. This work contributes to the field by presenting a systematic analysis of ISOT-CID, identifying effective IDS models, and proposing improvements for future cloud intrusion detection research.

Keywords: Cloud security; Intrusion detection system (IDS); ISOT-CID dataset; Cyber threat detection; Machine learning in cloud security.

1. Introduction

1.1. Background

Cloud computing has transformed modern information technology by enabling seamless data storage, processing, and accessibility across distributed environments. Cloud services allow organizations to scale operations, optimize resources, and reduce costs. However, this convenience comes with significant security challenges. Cloud infrastructure inherently relies on shared resources, third-party services, and complex configurations, which expand the attack surface and expose vulnerabilities not typically encountered in traditional IT infrastructures.

Cloud security faces unique threats beyond the standard attacks, including unauthorized hypervisor access, inter-tenant data leakage, and specific network-based assaults designed to exploit virtualization [1]. Given this landscape, Intrusion Detection Systems (IDS) have emerged as critical components for identifying suspicious activities and preventing unauthorized access in cloud environments. Despite various developments, IDS for cloud systems faces notable challenges in detecting cloud-specific threats effectively.

1.2. Importance of IDS in Cloud Security

IDS are fundamental for cloud security because they provide a proactive defense by identifying and flagging anomalous behaviors before they escalate into serious breaches. However, IDS's effectiveness in cloud environments depends on

* Corresponding author: Safana Alzide

having access to comprehensive, realistic datasets that simulate both benign activities and complex attack patterns typical in cloud scenarios. Public datasets such as KDD Cup 1999 and NSL-KDD are widely used in IDS research. Still, they lack the fidelity and realism necessary to address cloud-specific attack vectors, thus hindering the development of effective IDS tailored to cloud infrastructures [2].

1.3. Overview of ISOT-CID Dataset

The ISOT-CID dataset addresses this need by providing a benchmark for cloud IDS research. Collected over multiple phases from an operational cloud environment, ISOT-CID includes more than 2.5 terabytes of data, featuring a diverse mix of benign and malicious traffic [3]. The dataset is organized into two primary phases, each representing different timeframes and attack scenarios encompassing network-level and application-layer threats. ISOT-CID's real-world relevance and comprehensive data structure make it a valuable resource for IDS researchers seeking to develop and validate intrusion detection models tailored for cloud environments.

1.4. Research Objectives and Contribution

This study aims to:

1. Analyze the structure and attack scenarios within the ISOT-CID dataset.
2. Identify patterns and characteristics of the attack data, providing insights into common threat vectors in cloud environments.
3. Evaluate the effectiveness of various IDS models on the dataset, assessing their strengths and weaknesses in detecting cloud-specific threats.
4. Through this research, we contribute a detailed examination of ISOT-CID, insights into effective IDS methodologies for cloud environments, and recommendations for enhancing IDS models based on cloud-specific attack vectors.

2. Literature Review

2.1. Cloud Intrusion Detection Systems

Intrusion Detection Systems (IDS) in cloud environments must handle diverse and dynamic data flows generated by multiple tenants and users. Traditional IDS models, often developed for on-premises IT infrastructures, struggle with scalability and performance in cloud settings. Research has explored multiple IDS approaches, from rule-based systems to machine learning-driven models. For instance, rule-based IDS rely on predefined signatures of known threats to detect malicious activities, but their effectiveness is limited in novel or zero-day attack scenarios. In contrast, machine learning (ML) and deep learning (DL) models are increasingly preferred for their ability to detect previously unknown attack patterns.

Machine learning-based IDS have shown promising results in detecting anomalies in cloud data streams. Techniques such as Support Vector Machines (SVM), Decision Trees, and Neural Networks can adapt to evolving attack patterns, improving detection rates [4]. However, these models are often computationally intensive and require a robust training dataset. ISOT-CID, with its high-fidelity data representing real cloud traffic and varied attack scenarios, provides an ideal benchmark for evaluating these advanced models in a cloud context.

2.2. Challenges in Cloud Intrusion Detection

Cloud-specific challenges for IDS include handling the high volume and velocity of data generated across virtualized infrastructure. Analyzing such massive data in real-time necessitates efficient data processing techniques. Additionally, distinguishing between legitimate multi-tenant activities and malicious actions remains a key challenge. Due to this complexity, IDS in cloud environments often struggle with high false-positive rates, leading to inefficient alerting and resource allocation.

Another significant hurdle is the dynamic nature of cloud networks. Virtual machines (VMs) can be spun up or down on demand, leading to constantly changing IP addresses and configurations. This dynamism complicates maintaining accurate threat signatures and models, necessitating IDS to be capable of adapting to changes in network topology.

ISOT-CID's diverse data collection, encompassing various VM activities and configurations, enables the study and development of IDS that can account for these challenges [3].

2.3. Existing Benchmark Datasets

The lack of realistic, cloud-specific IDS datasets has been a limiting factor in cloud security research. Popular datasets like KDD Cup 1999 and NSL-KDD are widely used in IDS studies but do not represent the unique characteristics of cloud environments, particularly regarding virtualization and multi-tenancy [5]. UNSW-NB15 is a newer dataset with a broader range of attack types, but it lacks the complexity and volume required for cloud-specific research.

In this context, ISOT-CID offers a valuable alternative. It contains comprehensive logs, network traffic, and metadata from a real cloud infrastructure [1]. The dataset's inclusion of multiple types of attacks and benign activities enables a more robust evaluation of IDS models designed for cloud environments.

3. ISOT-CID Dataset Overview

3.1. Dataset Structure

The ISOT-CID dataset is organized into two primary phases, each capturing data across various cloud infrastructure layers, including hypervisors, virtual machines, and network interfaces. Phase One of the dataset, collected in December 2016, consists of four days of attack data interspersed with benign traffic, providing a mix of network and application-layer attacks. Phase Two, collected in February 2018, extends this coverage to five days, focusing on sophisticated application-layer threats such as SQL injection and cross-site scripting (XSS) [1].

Table 1 ISOT-CID Dataset Components

Node	VMs hosted	Capture label
A	5	poseidon0050
B	4	poseidon0049
C	1	-

Data within the ISOT-CID dataset is categorized into logs, network packet captures, memory dumps, and system events, which are stored in a structured file format. The dataset includes over 2.5 terabytes of data, enabling the examination of both packet-level and flow-level intrusion detection models. This extensive structure allows the analysis of distinct traffic types, including external, internal, and local flows within the cloud network.

Table 2 Hypervisor and VM Configuration

Node	ID	Hostname	Operating System
C	1	isotvm-1	Centos
A	2	isotvm-2	Centos
	3	hpisot-dj	Debian
	4	ohp-win12	Windows Server 12
	5	isotvm-in1	Ubuntu
	6	hpisot-centos7	Centos
B	7	hpisot-ubuntu	Ubuntu
	8	ohp-ubuntu	Ubuntu
	9	hpisot-winserve	Windows Server 12
	10	2hpisot-centos7	Centos

3.2. Collection Environment

The ISOT-CID environment encompasses three hypervisor nodes and ten virtual machines, each with unique configurations and operating systems such as CentOS, Debian, and Windows Server. These VMs are connected across different IP ranges, simulating a realistic cloud network with internal and external traffic flows. Including multiple

hypervisors, each supporting distinct sets of VMs facilitates the examination of inter-tenant isolation mechanisms and hypervisor-specific security challenges.

Data was collected from network interfaces configured promiscuously, ensuring comprehensive traffic capture across the cloud infrastructure. The dataset also includes detailed labeling information, with specific IP addresses flagged as benign or malicious. This labeling enables IDS researchers to train models on accurately classified data, enhancing detection rates and reducing false positives [3].

3.3. Data Labeling and Sampling

Data labeling in ISOT-CID is based on IP addresses, with packets originating from or directed to compromised hosts labeled as malicious. Each phase of the dataset provides CSV files listing labeled packet headers, making it straightforward to separate benign from malicious traffic. Sampling techniques were used in certain scenarios, such as network traffic data in Phase One, where data was collected in bursts of random duration. This approach balances storage requirements with dataset fidelity, allowing researchers to analyze realistic traffic flows without overwhelming computational resources [1].

---

4. Attack Scenarios and Data Analysis

The ISOT-CID dataset provides a unique collection of application and network layer attacks, each carefully curated and documented to reflect real-world conditions within a cloud environment. The dataset is divided into two distinct phases, each representing different timelines and types of attack scenarios. This structure allows for a comprehensive exploration of diverse attack types, aiding IDS researchers in effectively evaluating models capable of detecting various intrusion techniques.

4.1. Phase One Attack Scenarios

Phase One of the ISOT-CID dataset, collected in December 2016, features attacks targeting the internal network and specific application vulnerabilities. These attacks, spanning four days, include unauthorized login attempts, SSH brute force attacks, and network probes conducted to compromise VM systems. Attackers in this phase utilized different IP sources to execute Remote-to-Local (R2L) attacks and Denial of Service (DoS) attempts, creating a realistic simulation of multi-vector cloud attacks [2].

For instance, the first day of attacks included repeated SSH password-guessing attempts on VM 8 (IP: 172.16.1.24), followed by successful unauthorized login attempts, allowing the attacker to initiate internal network scans. The data labeling provided in ISOT-CID enables researchers to isolate benign from malicious traffic, with attack packets tagged for easy reference [2]. This labeling provides insights into traffic patterns and packet frequency, supporting robust IDS model training.

4.2. Phase Two Attack Scenarios

Phase Two, collected in February 2018, introduces more advanced application-layer attacks, including SQL Injection, Cross-Site Scripting (XSS), Directory Traversal, and Denial of Service (DoS). This phase spans five days and incorporates broader attack strategies to exploit web application vulnerabilities. Additionally, phase two includes crypto mining attempts, where malicious actors install and execute crypto miner software on compromised virtual machines [2].

This phase introduces distinct attack sequences, such as multiple attackers executing SQL injections and directory traversal to gain access to sensitive files and database credentials. By offering detailed logs of each attack sequence, ISOT-CID facilitates in-depth analysis of packet flows and aids in evaluating IDS models targeting cloud-specific attack patterns. Moreover, the structured labeling of benign and malicious traffic across both phases ensures that researchers can effectively benchmark and assess IDS model performance.

4.3. Temporal Patterns and Attack Intensities

An essential feature of the ISOT-CID dataset is its timestamped records, allowing for temporal analysis of attack patterns. For example, during phase one, the frequency and intensity of SSH brute-force attacks peak on specific days, reflecting a deliberate escalation of attack intensity. Similarly, in phase two, DoS attacks exhibit sustained periods of high packet flow, challenging IDS models to differentiate between benign traffic surges and malicious floods [2].

Temporal pattern analysis within ISOT-CID supports the development of adaptive IDS models that can respond to changing attack intensities. For instance, machine learning-based models can leverage these patterns to classify abnormal traffic based on temporal features, enhancing detection accuracy in real-time cloud environments.

5. Methodology

The methodology for analyzing the ISOT-CID dataset and developing IDS models encompasses several key phases: data preprocessing, feature extraction, and modeling. This section outlines the approach to preparing and utilizing the dataset effectively, ensuring that IDS models can be trained and evaluated comprehensively.

5.1. Data Preprocessing

Data preprocessing is critical in handling the ISOT-CID dataset, given its size and complexity. The dataset includes over 2.5 terabytes of raw data, necessitating efficient data handling methods. Initial steps involved data filtering to isolate relevant traffic based on IP addresses and timestamps, aligning with the labeled packets for benign and malicious activities. The ISOT-CID documentation provides CSV files containing packet-level labels, facilitating the separation of attack and normal traffic for model training [2].

Sampling techniques were applied to reduce data volume for specific experiments, particularly in phase one, where network traffic was collected in random-duration bursts. These techniques ensured that IDS models could be trained on a representative subset without overwhelming computational resources. Moreover, memory dumps and system logs were parsed to extract critical features, such as system calls, resource utilization metrics, and timestamps, adding contextual data to enhance model accuracy.

5.2. Feature Extraction

Effective IDS relies on carefully selecting features from network traffic and system logs. The ISOT-CID dataset provides many features, including packet headers, timestamps, source and destination IPs, protocol types, and payload sizes. In particular, network layer features such as packet rates, port numbers, and protocol distribution are essential for detecting DoS and DDoS attacks. For application-layer attacks, features derived from HTTP and SQL traffic, including query structure and URL patterns, were identified as critical [1].

In this study, both basic and advanced feature extraction methods were employed. Basic methods involved standard network features, while advanced extraction used parsing techniques to identify specific patterns, such as SQL injection strings and directory traversal paths. These extracted features were normalized and prepared for input into machine learning models.

5.3. Modeling Approaches

Three primary modeling approaches were evaluated on the ISOT-CID dataset: statistical, machine learning (ML), and deep learning (DL) methods.

Statistical Methods: Traditional statistical models, such as anomaly detection using Gaussian distributions, were first employed to establish a baseline. These models analyze traffic by comparing observed data distributions against historical norms, flagging outliers as potential intrusions [4].

Machine Learning Models: ML models, including Support Vector Machines (SVM), Decision Trees, and Random Forests, were applied to classify traffic based on extracted features. The ability of these models to handle non-linear relationships made them suitable for detecting complex attack patterns. Each model was trained on labeled subsets of ISOT-CID data, and hyperparameters were optimized to enhance detection rates.

Deep Learning Models: Finally, advanced DL models, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, were used to analyze time-series data within the dataset. CNNs effectively captured spatial relationships within packet flows, while LSTMs leveraged temporal dependencies, making them highly effective for detecting DDoS and sequence-based attacks in the ISOT-CID dataset [5].

6. Experimental Setup and Results

6.1. Experimental Design

Experiments were conducted by splitting the ISOT-CID dataset into training and testing sets, with phase one used primarily for training and phase two for validation. This approach allowed models to be tested on different attack scenarios and timelines, ensuring robustness across multiple cloud attack patterns. Evaluation metrics included accuracy, precision, recall, and F1-score, commonly used in IDS research to assess model effectiveness [1].

Evaluation Metrics

- Accuracy: Measures the proportion of correctly classified instances over the total instances.
- Precision: Indicates the proportion of true positives among all positive classifications.
- Recall: Measures the proportion of true positives among all actual positives.
- F1-score: The harmonic mean of precision and recall, providing a balanced assessment for imbalanced datasets.

Results and Observations

The performance of each IDS model varied according to attack type and data phase.

Statistical Models: These models successfully detected basic attacks, such as port scans and DoS attempts. However, they showed limitations in accurately detecting sophisticated attacks, particularly those at the application layer.

Machine Learning Models: Decision Trees and Random Forests exhibited high precision and recall rates for network-level attacks. SVM achieved strong results in classifying benign vs. malicious packets, though its performance declined for multi-vector attacks present in phase two.

Deep Learning Models: CNNs and LSTMs provided the best overall results, with accuracy rates surpassing 90% for most attack types. The temporal pattern recognition capability of LSTMs was particularly effective for sequence-based attacks, such as brute-force SSH login attempts [4].

6.2. Model Comparison and Observations on Attack Detection

The deep learning models consistently outperformed statistical and traditional ML methods, especially in handling the complex, multi-layered attacks characteristic of the ISOT-CID dataset. While traditional ML models performed strongly in network traffic classification, they struggled with application-layer attacks. Conversely, CNN and LSTM models leveraged the large-scale structure of the dataset to detect subtle patterns in packet flows and HTTP requests, identifying SQL injection and XSS attacks with high accuracy.

7. Discussion

This study reveals critical insights into the challenges and effectiveness of IDS models when deployed in cloud environments. The ISOT-CID dataset is essential for understanding the types of threats commonly encountered in cloud infrastructure and evaluating IDS models under realistic cloud conditions.

7.1. Insights on Cloud IDS Effectiveness

The evaluation of various IDS models demonstrated that traditional statistical and machine learning (ML) methods are adequate for detecting simpler attack types, such as DoS and brute-force attacks. However, they often fail to handle the complex, multi-layered threats in modern cloud environments. This limitation is particularly evident when dealing with application-layer attacks such as SQL injection and cross-site scripting (XSS), which require a more nuanced understanding of request patterns and payload structures [1][2].

Deep learning (DL) models, specifically Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, exhibited superior performance in capturing both spatial and temporal patterns in the dataset. These models effectively identified complex attacks by analyzing patterns over time and across multiple packet flows, highlighting the advantages of DL techniques for cloud IDS applications. The temporal depth provided by LSTMs proved particularly valuable in detecting sequential attacks, such as credential stuffing and brute-force logins, which follow a distinct temporal sequence [4].

7.2. Challenges and Limitations

Several challenges emerged from this study, primarily related to data volume, model scalability, and computational intensity. The ISOT-CID dataset's substantial size, over 2.5 terabytes, presented storage and processing challenges that required data sampling and dimensionality reduction strategies. While this approach preserved a representative subset of the data, it limited the ability to perform certain real-time intrusion detection tasks, underscoring the need for scalable solutions in cloud IDS research [3].

Another challenge was the high rate of false positives in certain ML models. Distinguishing between benign and malicious multi-tenant activities proved difficult, especially when benign behaviors shared characteristics with attack traffic. This issue highlights the need for more refined feature engineering and adaptive models capable of learning from dynamically evolving cloud traffic.

7.3. Practical Implications

The findings from this study have important implications for cloud security. IDS models utilizing deep learning frameworks show promise for real-time application in large-scale cloud environments due to their ability to learn complex patterns and adapt to new attack types. However, their computational requirements demand efficient deployment strategies, such as leveraging cloud-native resources like containerized deployments and hardware accelerators. Additionally, the insights gained from ISOT-CID underscore the importance of benchmark datasets that capture realistic traffic in cloud scenarios, providing a foundation for IDS development and facilitating more robust security practices in cloud infrastructure [1].

---

8. Future Work

This study opens up several avenues for future research in cloud IDS and using benchmark datasets like ISOT-CID for intrusion detection model evaluation.

8.1. Enhanced Feature Engineering

One area of improvement lies in the feature engineering process. Future studies could explore more advanced feature extraction methods, such as deep packet inspection and natural language processing (NLP), for analyzing application-layer attacks. For instance, NLP techniques could parse SQL queries to detect injection patterns more accurately. Furthermore, features capturing user behavior and session data could provide additional context, enhancing the model's ability to differentiate between benign and malicious multi-tenant activities [2].

8.2. Hybrid and Ensemble Modeling

While deep learning models demonstrated superior performance in this study, hybrid and ensemble modeling approaches could be developed to optimize IDS performance further. By combining statistical, machine learning, and deep learning techniques, hybrid models can leverage the strengths of each approach to provide more accurate and robust intrusion detection. For example, a hybrid IDS could use rule-based detection for known signatures while relying on machine learning to flag anomalies, improving efficiency and accuracy [4].

8.3. Real-Time IDS and Model Optimization

Implementing real-time IDS systems in cloud environments is critical for future work. Real-time IDS requires models that can process high volumes of data quickly and efficiently. Research into model optimization, such as using federated learning to update IDS models across distributed cloud nodes, could improve detection speed and accuracy. Additionally, exploring lightweight DL architectures or edge-based processing for cloud IDS may enhance real-time capabilities without overwhelming computational resources [5].

8.4. Expanding the ISOT-CID Dataset

Future iterations of ISOT-CID could include additional types of attacks, such as ransomware, insider threats, and Advanced Persistent Threats (APTs), which have become increasingly prevalent in cloud environments. Expanding the dataset to incorporate these attack types would further enhance its utility for IDS research. Moreover, creating a labeled dataset with finer-grained annotations, such as specific protocol breakdowns and application-level features, could improve the granularity and applicability of IDS model training.

9. Conclusion

In this study, we comprehensively analyzed the ISOT Cloud Intrusion Detection Benchmark Dataset (ISOT-CID), exploring its structure, attack scenarios, and utility for developing and testing cloud-specific Intrusion Detection Systems (IDS). The ISOT-CID dataset, with its rich mix of network and application-layer attacks, proved invaluable in evaluating IDS models tailored to cloud environments. Through this research, we assessed various IDS models, including statistical, machine learning, and deep learning approaches, highlighting their strengths and limitations in cloud settings.

Our findings underscore the advantages of deep learning models, particularly Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, in detecting complex and temporally sequenced attacks. However, we also identified limitations in computational scalability and false-positive rates, emphasizing the need for hybrid and real-time IDS solutions tailored to the dynamic nature of cloud networks.

This study contributes to cloud security research by providing insights into effective IDS methodologies and the application of the ISOT-CID dataset for realistic intrusion detection model evaluation. Future work could build upon these findings by enhancing feature extraction, optimizing real-time processing capabilities, and expanding ISOT-CID to include a broader range of cloud-specific threats. Ultimately, advancing IDS research with robust benchmarks like ISOT-CID will play a crucial role in strengthening the defenses of cloud infrastructures against evolving cyber threats.

10. References

[1] A. Aldribi, I. Traore, and B. Moa, "Data Sources and Datasets for Cloud Intrusion Detection Modeling and Evaluation," in Cloud Computing for Optimization: Foundations, Applications, and Challenges, Studies in Big Data, vol. 39, Springer, 2018, pp. 333-366.

[2] A. Aldribi, I. Traore, P. G. Quinan, and O. Nwamuo, "Documentation for the ISOT Cloud Intrusion Detection Dataset," Technical Report #ECE-2020-10-10, University of Victoria, ECE Department, 2020.

[3] A. Aldribi, I. Traore, B. Moa, and O. Nwamuo, "Hypervisor-Based Cloud Intrusion Detection through Online Multivariate Statistical Change Tracking," Computers & Security, vol. 87, 2019, doi: https://doi.org/10.1016/j.cose.2019.101646.

[4] U. Tupakula and V. Varadharajan, "A Practical Approach to Implement IDS for Cloud Computing," in IEEE Transactions on Services Computing, vol. 5, no. 1, pp. 188-199, 2012, doi: 10.1109/TSC.2010.53.

[5] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, 2009, pp. 1-6, doi: 10.1109/CISDA.2009.5356528.