

Security attack classification and performance analysis of machine learning algorithms over CSIC HTTP dataset

Dipesh Vaya *

Rakesh Kumar Saxena †

Department of Computer Science & Engineering
Poornima University
Jaipur 303905
Rajasthan
India

Abstract

This research study analyzes the performance of machine learning-based algorithms in terms of efficiency in classifying security attacks in cloud computing systems. The proposed research consists of two phases: identifying and wrangling data as per the required format and performing analysis of different machine learning algorithms based on security attack detection and classification using the CSIC HTTP dataset. It contains a set of static and dynamic attacks created using connected objects. Security attacks are detected and analyzed on machine learning algorithms, like k-nearest Neighbor, random forest, and support vector machine. This research helps to find out the set of illegal requests that deteriorate the privacy of shared data. The performance of algorithms is explored through a confusion matrix. During analysis KNN has shown good balance between precision and recall, which is significant in the classification task over Random Forest and support Vector Machine models for different-sized test data.

Subject Classification: Primary 68T07, Secondary 68T05.

Keywords: Security attack classification, Detection, Machine learning, Performance analysis

1. Introduction

In today's digital world, the criticality of cloud computing is a significant part of the research community. With the rise in technology, security and privacy concerns are also taking equal focus. With the

* E-mail: dipesh.vaya88@gmail.com (Corresponding Author)

† E-mail: saxenark06@gmail.com

increased security challenges, the need for efficient security measures has become paramount [1, 2]. Organizations are now dependent on cloud-based infrastructures for their day-to-day operations. Ensuring data security, integrity, privacy, and confidentiality is a nightmare for organizations. This research explores the domain of web-based security attack detection, focusing on the efficiency of machine learning-based algorithms. Edge computing is one of the present fields that serve as utility models. Relied on their premises, the users presented at the edges can access or share the data. On the flip side, the demand for computational assessment increases in today's life regarding data rates and quality of services and ensures availability [3-6]. In this context, the current study unfolds in two distinct phases, each essential to reaching the study's primary objective.

In the first phase, identifying and preparing publicly available datasets was the primary concern for reliable and accurate analysis. In the case of this research, data preprocessing is a crucial initial step to confirm that the machine learning algorithms are applied to a clean and standardized dataset. A thorough analysis and comparison of various machine learning based techniques for security attack classification and detection is part of this study's second phase. The CSIC HTTP data set is taken for analyzing the machine learning algorithms due to a diverse range of static security attacks. The CSIC HTTP dataset is primarily designed for research and experimentation in web security and detecting web-based attacks. It contains actual network traffic data related to HTTP requests and responses. The dataset is notable for including benign (regular) traffic and malicious traffic. The machine learning algorithms, like the support vector machine, random forest, and k-nearest Neighbor analyzed in this phase, play a pivotal role in identifying and classifying these security attacks. This research is significant in uncovering illicit requests in cloud computing systems that compromise the security of shared data. The study provides valuable analysis for strengthening the integrity of cloud-based services and protecting sensitive information by precisely detecting and categorizing security assaults. The effectiveness of the machine learning algorithms is assessed using a confusion matrix, which provides information on the classification results like accuracy, precision, recall, and F1-score. This research paper is divided into many parts. In literature review multiple techniques, and some of them, primarily based on machine learning, are discussed for various scenarios. Research Methodology describes the security attack detection models using CSIC HTTP datasets.

2. Literature Review

For the classification of security attacks, researchers have contributed multiple techniques, and some of them, primarily based on machine learning, are discussed below.

The authors incorporated a Support Vector Machine (SVM) to classify routing attacks and intrusion detection for IoT. SVM achieved up to 100% and 81% accuracy [7]. Ontology-based machine learning models were analyzed using SPARQL to detect Anomaly and buffer-overflow attacks with the help of protege tool [8]. Different open-source languages that dealt with the ontology's development were available in literature. The behavior of the nodes was examined by encountering all error messages. The system reduced the low false positive and alarm rate. To detect low-rate DDoS attacks using WSNs and Internet of Things (IoT) networks, Hilbert-Huang Transform (HHT) is used to monitor DDoS [9]. To detect DDoS attacks using 6LoWPAN, With the help of standard protocol IETF, an adaptation layer was developed on low-power devices. Then, the IDS model was integrated with the EU FP7 project ebbits. Penetration testing was performed over all types of IoT attacks in [10]. Healthcare data is sensitive in nature. Thus, high-end security must be provided for healthcare applications. Most cloud security applications use Elliptic Curve Cryptography (ECC), which brings issues like delay and latency on high data volumes. An extensive survey is done to develop an understanding of cloud IoT-based security using machine learning algorithms [11]. The study has helped other researchers learn about IoT sensors, machine learning, and cloud technologies in healthcare sectors. In [12], technology was embedded with situation-aware sensors and devices to develop an IDS for detecting complex event processing (CEP). It was modulated in Esper software. The suggested IDS have four functions: data filtering, event modeling, event analysis, and security response. Compared to traditional IDS, this IDS lowered the time complexity of data scale, CPU utilization, and Memory consumption. In [13], Generative Adversarial Networks (GAN) and Differential Privacy (DP) are incorporated in hybrid to secure crucial and sensitive data in IIoT operations. The proposed method preserves data privacy with minute loss in accuracy and also provides protection to confidential sensitive data in terms of privacy. Authors in [14] use Local Differential Privacy (LDP) to address privacy concerns in Blockchain-based IoT. LDP mechanism perturbs input numerical data with digital bits, adaptive privacy budget allocation, high data utility, and better performance in different scenarios.

To preserve data privacy in the Split Learning (SL) model for IoT applications, authors proposed the Binarized SL (B-SL) model [15]. B-SL model reduces computation time and memory usage, reduces privacy leakage, and can integrate differential privacy for enhanced privacy preservation. In [16], an adversarial learning-based framework is utilized to preserve privacy in graph embedding methods for social networks. The proposed framework reduces the adversary's prediction accuracy on sensitive links, conserves graph topology and node attributes, and improves privacy preservation. Authors in [17] worked for vehicular networks by proposing reputation management system for MCS using lightweight privacy preservation. The PPRM scheme preserves location, identity, sensing data, reputation value privacy, and efficient and accurate reputation value in the updated algorithm. A comprehensive analysis of privacy-preservation techniques for IoT devices, cryptography techniques for authentication, and discussion of privacy interference is carried out in [18]. Authors utilized blockchain and federated learning to preserve privacy and misbehavior detection in lightweight IoMT devices [19]. The proposed method is efficient regarding high recall rate, secure misbehavior detection in IoMT devices, and privacy preservation with blockchain integration. In [20], authors proposed a scalable approach for big data multidimensional anonymization in cloud computing based on MapReduce, a tree indexing structure. The proposed method improved scalability for multidimensional anonymization, recursive computation, and applicability to differential privacy. For Privacy preservation in smart cities, authors in [21] used differential privacy. In [22], authors researched privacy preservation in IoMT using federated learning and advanced techniques. However, it may require sophisticated models and methods, posing potential challenges in deployment and implementation in real-life systems.

Multiple research papers by various researchers are studied to understand possible threats to cloud-based systems. Authors in [23] raised issues related to trust and threats to cloud computing. Authors in [24, 25] pointed out privacy, trust, and security issues in mobile and cloud computing. Authors also conducted similar studies in [26–30]. Various researchers' research on intrusion detection using multiple machine learning algorithms was also studied while compiling the related literature. In [31], the author studied Multi-Layer Perception (MLP) and found it appropriate for intrusion detection in IoT systems. Many researchers have conducted more studies [32–35] to identify the suitability and correctness

of machine learning-based algorithms for enhancing security and intrusion detection and implications of artificial intelligence as a technology.

3. Research Methodology

This section presents the workflow of the research study as depicted in figure 1. This part presents the security attack detection model using CSIC HTTP datasets. This dataset contains three sorts of attacks: static, dynamic, and unintentional illegal requests. The dataset is parted into two sets, viz, training and testing. The training dataset is preprocessed, and the required attack features are selected. These features are then trained with the different machine learning models. The initial step in the research methodology involves data collection and preprocessing. The CSIC HTTP dataset, containing various types of security attacks and legitimate traffic, was acquired for training and testing the security attack detection models. The dataset is carefully preprocessed to remove the noise, duplicates, or irrelevant information that can impact the performance of the machine learning models. Features relevant to attack types, network traffic patterns, and user behavior are extracted and normalized to ensure consistency and accuracy in subsequent analyses. Recall, Precision, Accuracy, and F-scores will be calculated to observe the efficiency of the methodology used.

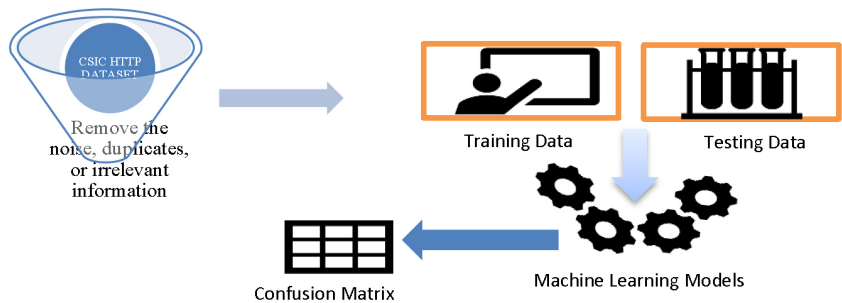


Figure 1
Methodology Used

4. Results and Analysis

For analysis, all the columns are transformed with distinct values so that data can be fed to analysis models. Table 1 shows the importance of F1-Score, Precision, Accuracy, and Recall when the test data size is 10%,

Table 1
Analysis of different security attack classification models

S.N.	Name of Model	Size of Test Data	F1 Score	Precision	Accuracy	Recall
1	KNN	0.1	0.821394	0.820892	0.822123	0.823284
		0.2	0.877664	0.877865	0.81707	0.880096
		0.3	0.815596	0.815154	0.816497	0.81879
2	RF	0.1	0.473465	0.788948	0.600889	0.560049
		0.2	0.652663	0.857901	0.606617	0.707842
		0.3	0.483313	0.794303	0.611863	0.563568
3	SVM	0.1	0.407551	0.499063	0.539188	0.499738
		0.2	0.267761	0.355164	0.552127	0.337496
		0.3	0.390474	0.524712	0.554773	0.503486

20%, and 30%, and the rest is taken as training data. From the following figure 2(a), it can be observed that in all three cases (for testing data sizes 10%, 20%, and 30%), KNN showed the highest accuracy in classifying the attacks over Random Forest and SVM models. But KNN is more accurate when the testing data size is 10%.

Figure 2(b) shows that in all three cases (for testing data sizes 10%, 20%, and 30%), KNN showed the highest F1-Score in classifying the attacks over Random Forest and SVM models. However, KNN has the highest F1 score when the testing data size is 20%. From figure 2(c), it can be observed that in all three cases (for testing data sizes 10%, 20%, and 30%), the precision of KNN is the highest in classifying the attacks over Random Forest, and SVM models. However, KNN has the highest precision when the testing data size is 20%.

Here, the precision value of Random Forest is also competing with KNN. Figure 2(d) shows that in all three cases (for testing data sizes 10%, 20%, and 30%), the recall capability of KNN is the highest in classifying the attacks over Random Forest, and Support Vector Machine models. KNN has the highest precision when the testing data size is 20%.

5. Conclusion and Future Scope

The research study has shown the performance of KNN over Random Forest and SVM models for different-sized test data. KNN achieved an F1-score of 0.821394, offering a good balance between precision and recall,

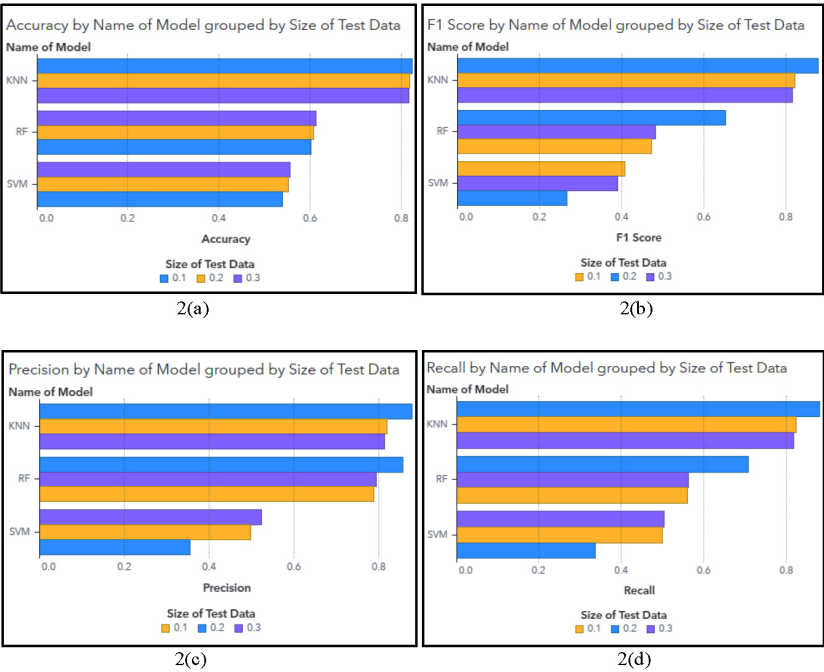


Figure 2

Comparison based on Accuracy 2(a), F1 Score values 2(b), Precision (2c), and Recall 2(d) by Name of Model grouped by Size of Test Data

which is significant in the classification task, especially when dealing with imbalanced datasets. The precision of 0.820892 means that when the KNN model predicts an attack, it is correct about 82.08% of the time. High precision is required when false positives are costly or unwanted. An accuracy of 0.822123 means that the KNN model is accurate in its predictions approximately 82.21% of the time. A recall score of 0.823284 indicates that the model can capture 82.33% of the actual attacks. High recall is necessary when missing a positive attack is costly. The KNN model performed well in classifying attacks on the CSIC HTTP dataset, with a good balance between precision and recall and relatively high accuracy. Random Forest and SVM models were less efficient in accuracy and recall when compared with KNN in all three cases (for 10%, 20%, and 30% of test data size). KNN was more efficient when test data size was taken as 20% with an F1 Score of 0.877664, Precision score of 0.877865, 0.81707 of Accuracy, and Recall score of 0.880096.

This study can be extended further by examining the performance of deep learning algorithms or by implementing hybrid algorithms for classifying malware attacks in the cloud environment.

References

- [1] D. Vaya and S. Khandelwal, "A fast and hardware-efficient visual cryptography scheme for images," *Advances in Intelligent Systems and Computing*, vol. 379, pp. 133–142 (2016).
- [2] D. Vaya and T. Hadpawat, "Enhanced fingerprint authentication using blockchain", *Blockchain 3.0 for Sustainable Development*, pp. 123–130 (2021).
- [3] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile", *Edge Computing: The Communication Perspective* (2017).
- [4] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges", *IEEE Internet Things*, vol. 3, no. 5, pp. 637–646 (2016).
- [5] D. Vaya and T. Hadpawat, "Internet of Everything (IoE): A New Era of IoT", *Lecture Notes in Electrical Engineering, Springer*, vol. 570 (2020).
- [6] W. Shi and S. Dustdar, "The Promise of Edge Computing", *Computer (Long Beach Calif)*, vol. 49, no. 5, pp. 78–81 (2016).
- [7] C. Ioannou and V. Vassiliou, "Classifying Security Attacks in IoT Networks Using Supervised Learning", *ICDCOSS: IEEE*, pp. 652–658 (2019).
- [8] M. Ansarinia, S. A. Asghari, A. Souzani, and A. Ghaznavi, "Ontology-based modeling of DDoS attacks for attack plan detection", *IST: IEEE*, pp. 993–998 (2012).
- [9] H. Chen, C. Meng, Z. Shan, Z. Fu, and B. K. Bhargava, "A Novel Low-Rate Denial of Service Attack Detection Approach in ZigBee Wireless Sensor Network by Combining Hilbert-Huang Transformation and Trust Evaluation", *IEEE Access*, vol. 7, pp. 32853–32866 (2019).
- [10] P. Kasinathan, C. Pastrone, M. A. Spirito, and M. Vinkovits, "Denial-of-Service detection in 6LoWPAN based Internet of Things", *IC-WiMob: IEEE*, pp. 600–607 (2013).
- [11] P. Ghosal, D. Das, and I. Das, "Extensive survey on cloud-based IoT-healthcare and security using machine learning", *Proceedings ICRICN*, pp. 1–5 (2018).
- [12] C. Jun and C. Chi, "Design of Complex Event-Processing IDS in Internet of Things", *ICMTMA: IEEE*, pp. 226–229 (2014).

- [13] Y. S. Hindistan and E. F. Yetkin, "A Hybrid Approach with GAN & DP for Privacy Preservation of IIoT Data", *IEEE Access*, vol. 11, pp. 5837–49 (2023).
- [14] K. Zhang, J. Tian, H. Xiao, Y. Zhao, W. Zhao, and J. Chen, "A Numerical Splitting and Adaptive Privacy Budget-Allocation-Based LDP Mechanism for Privacy Preservation in Blockchain-Powered IoT", *IEEE Internet Things*, vol. 10, no. 8, pp. 6733–6741 (2023).
- [15] N. D. Pham, A. Abuadbba, Y. Gao, K. T. Phan, and N. Chilamkurti, "Binarizing Split Learning for Data Privacy Enhancement and Computation Reduction", *IEEE Transactions*, vol. 18, pp. 3088–3100 (2023).
- [16] K. Zhang, Z. Tian, Z. Cai, and D. Seo, "Link-privacy preserving graph embedding data publication with adversarial learning", *Tsinghua Sci Technol*, vol. 27, no. 2, pp. 244–256 (2022).
- [17] Y. Cheng, J. Ma, Z. Liu, Y. Wu, K. Wei, and C. Dong, "A Lightweight Privacy Preservation Scheme with Efficient Reputation Management for Mobile Crowdsensing, in Vehicular Networks", *IEEE Transactions* (2022).
- [18] D. Torre, A. Chennamaneni, and A. Rodriguez, "Privacy-Preservation Techniques for IoT Devices: A Systematic Mapping Study", *IEEE Access*, vol. 11, pp. 16323–16345 (2023).
- [19] S. Rahmadika, P. V. Astillo, G. Choudhary, D. G. Duguma, V. Sharma, and I. You, "Blockchain-Based Privacy Preservation Scheme for Misbehavior Detection in Lightweight IoMT Devices", *IEEE Biomed Health Inform*, vol. 27, no. 2, pp. 710–721 (2023).
- [20] X. Zhang, "MRMondrian: Scalable Multidimensional Anonymisation for Big Data Privacy Preservation", *IEEE Big Data*, vol. 8, no. 1, pp. 125–139 (2022).
- [21] M. Gheisari, "An Agile Privacy-Preservation Solution for IoT-Based Smart City Using Different Distributions", *IEEE Open Journal of Vehicular Technology*, vol. 4, pp. 356–362 (2023).
- [22] M. Ali, F. Naeem, M. Tariq, and G. Kaddoum, "Federated Learning for Privacy Preservation in Smart Healthcare Systems: A Comprehensive Survey", *IEEE J Biomed Health Inform*, vol. 27, no. 2, pp. 778–789 (2023).
- [23] M. T. Khorshed, A. B. M. Shawkat Ali, and S. A. Wasimi, "Trust issues that create threats for cyberattacks in cloud computing", *International Conference on Parallel and Distributed Systems - ICPADS*, pp. 900–905 (2011).

- [24] K. Yang, K. Zhang, J. Ren, and X. S. Shen, "Security and privacy in mobile crowdsourcing networks: Challenges and opportunities", *IEEE Communications*, vol. 53, no. 8, pp. 75–81 (2015).
- [25] M. Alouane and H. El Bakkali, "Security, privacy and trust in cloud computing: A comparative study", *ICCCTA:CloudTech 2015* (2015).
- [26] F. Medjek, D. Tandjaoui, I. Romdhani, and N. Djedjig, "A trust-based intrusion detection system for mobile RPL based networks", *IEEE Green Computing and Communications, IEEE Cyber, Physical and Social Computing, IEEE Smart Data*, pp. 735–742 (2018).
- [27] V. M., "An Advanced Survey on Cloud Computing and Alleviation Techniques", *Int J Comput Appl*, vol. 181, no. 24, pp. 19–22 (2018).
- [28] P. M. Khilar, V. Chaudhari, and R. R. Swain, "Trust-Based Access Control in Cloud Computing Using Machine Learning". *Springer* (2019).
- [29] J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data Security and Privacy-Preserving in Edge Computing Paradigm: Survey and Open Issues", *IEEE Access*, vol. 6, pp. 18209–18237 (2018).
- [30] M. Zolotukhin and T. Hämäläinen, "On Artificial Intelligent Malware Tolerant Networking for IoT," *2018 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, pp. 1-6 (2018).
- [31] I. Gupta, R. Gupta, A. K. Singh, and R. Buyya, "MLPAM: A Machine Learning and Probabilistic Analysis Based Model for Preserving Security and Privacy in Cloud Environment", *IEEE Syst J*, pp. 1–12 (2020).
- [32] D. Bhamare, T. Salman, M. Samaka, A. Erbad, and R. Jain, "Feasibility of Supervised Machine Learning for Cloud Security", *ICISS*, pp. 31–35 (2017).
- [33] M. Shashi, A. Saxena, D. Vaya, N. Pillai, SS. Chavan and C. Balarama Krishna, "Enhancing Trust and Privacy in Edge and Cloud Computing Through Blockchain Technology", *IEEE IC3I*, pp. 1812-1816 (2023).
- [34] S. Gupta, M. Shahid, A. Goyal, R. K. Saxena and K. Saluja, "Black Hole Detection and Prevention Using Digital Signature and SEP in MANET", *ICETET-SIP-22*, pp. 1-5 (2022).
- [35] D. Moud and R. K. Saxena, "Development of signature recognition system using VGG16," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 26, no. 3, pp. 807–813 (2023).