Journal of Information Systems Engineering and Management
2025, 10(11s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Proposed Improving Protection of Cloud Computing Environments Based on Machine Learning Techniques

Osamah M. Abduljabbar[1], Omar Dhafer Madeeh[2*], Safa Mohammed Mushib[3]
[1,2]*Electronic Computer Center, University of Fallujah, Anbar, Iraq*
[3]*Department of Computer, Collage of Engineering. Al-Nahrain University, Jadriyah, Baghdad, Iraq*
***Corresponding Author**: almihamdi@uofallujah.edu.iq*

ARTICLE INFO

Received: 12 Nov 2024

Revised: 27 Dec 2024

Accepted: 14 Jan 2025

ABSTRACT

Cloud computing contains a huge amount of data, which makes it a common target for cyberattacks to access confidential data using various illegal methods. One of these attacks is the Structure Query Language injection attack SQLIA, It is categorised as one of the most prevalent threats to obtaining, modifying, or destroying data by the Open Web Application Security Project (OWASP). Therefore, it has become necessary to create a model to detect attacks on data on the cloud to protect it, in order to increase trust between individuals and institutions and not make this data available to people who are not authorized to access it. To solve these problems, this study presents a proposal to improve the protection of the cloud computing environment through two contributions.The first contribution is developing a machine learning model, known as a logistic regression framework, that serves as a mediator between the client and the server. Its goal is to ascertain the kind of requests that are received from the customer layer and whether or not they include hazardous or typical payloads. The second contribution is illustrating how dangerous it is for the cloud computing infrastructure and for consumers and organisations to rely on false forecasts regarding the confidentiality, integrity, and real-time availability of data.. The results obtained from applying the proposed model showed a very high accuracy of 99.82, and showed low rates of false negatives and positives. In addition, the time it takes to determine the type of request sent is 0.1514 seconds.

Keywords: Cloud computing, SQLIA, Machine Learning, Logistic Regression, False Positive

INTRODUCTION

Cloud computing virtualization technology offers effective resources to end-users. Cloud computing is characterized by its manageability, scalability, and availability. Cloud computing confers several advantages, including economic viability, accessibility of on-demand services, convenience, ubiquity, multi-tenancy, adaptability, and dependability [1], . Cloud computing offers a range of service delivery models and development patterns, including Platform as a Service (PaaS), Software as a Service (SaaS), and Infrastructure as a Service (IaaS). Additionally, cloud computing encompasses various deployment models, such as Public Cloud, Private Cloud, Hybrid Cloud, Community Cloud, and Virtual Private Cloud [1],[2]. Although the cloud has many benefits, there are also many security risks. Among these dangers, (SQLIA) has lately gained more attention since it enables attackers to overcome authentication, access confidential information, edit data, or destroy databases [3].

SQLIA is a technique that can be employed to exploit database applications that are driven by web-based interfaces. These attacks may appear in various forms, contingent upon the attacker's objectives. The primary cause of (SQLIA) is inadequate validation of user input. According to OWASP Project statistics, it is among the top most dangerous attacks susceptible to database-driven web applications [4]. Thus, the remaining sections of this work will be organised as follows: Previous research about this study will be discussed in the second section. An overview of some of the ideas utilised in this work is given in the third section. The Framework of this paper and the obtained results will be clarified in the fourth and fifth sections, respectively. In the six sections, the most critical conclusions that have been reached will be mentioned.

RELATED WORK

Numerous articles about the function of cloud computing in the security industry have been published recently. But as of yet, no comprehensive answer to this issue has emerged. For instance, Muhammad Azizi M. et al. [5] suggested a method to show how to start API depletion attacks and spying on cloud API authentication services. The AD3 algorithm is suggested for use in assault detection in this study. This study, however, is hampered by a thorough practical analysis.

Nicole V. Nanane and others [6] detected several cloud threats using machine learning algorithms Support Vector Machine (SVM); nevertheless, this study does not describe the actual cloud environment utilized in the work, nor does it go into detail about the results produced and their accuracy percentage.

Yan Hou and Kuisheng Wang [7] the suggested kind of SQL detection technique embeds input cleaning and dynamic assessment into the cloud environment. There are three steps in the procedure. The method's initial step is to analyze the SQL keywords, and then it creates a rule tree by analyzing the syntax rules in the SQL statement. Finally, it uses a model established by SQL syntax regulation to investigate ternary trees to detect attacks; however, the accuracy is not observed.

Dharitri Tripathy and et al [8] Proposed to use machine learning methods for application-level SQL injection detection. With a detection rate of more than 98%, the algorithms used can discriminate between legitimate and malicious payloads. Although this method was successful, it did not illustrate the temporal complexity o f obtaining the results.

B. Shunmugapriya and Dr. B. Paramasivan [9] offered Twofish encryption algorithm to encrypt the data that the owner uploaded. The approach worked well to reduce the SQLIA but was unable to identify the assaults.

Solomon Ogbomon Uwagbole and et al [10] Suggested to create a web application that anticipates dictionary word lists as vector variables to display massive amounts of learning data. The prior method for teaching machine learning (ML) to anticipate and prevent SQLIA yields accurate results, however the precision found in my article was 99.82.

BACKGROUND

Information Security Requirements

Information confidentiality, Integrity, and availability are at the foundation of Information Security Requirements and have found widespread application in a wide range of academic disciplines [11]. Cloud technologies are extensively employed in developing IT infrastructure for various entities such as businesses, academic institutions, governments, and individuals, as they offer practical data processing and storage options. Despite numerous benefits, several limitations exist, particularly in the domains of security, dependability, and efficiency of computing and communication [12] ,[13].

Confidentiality relates to regulations and limitations that restrict entry to specific categories of data and safeguard user information as confidential, exclusive, and inaccessible even to the cloud vendor. The fundamental concept of information security involves the maintenance of a comprehensive and unimpaired data framework, which is sustained by the principle of Integrity. The data stored in cloud computing systems must remain unaltered by any party other than the rightful owner. The term)Availability (denotes the uninterrupted accessibility of a service to the user.

Machine Learning Techniques

The utilization of the Logistic Regression (LR) approach has been widely featured in numerous areas. The LR approach is applied when the aim is to categorize data elements into classes. LR typically involves a binary target variable, wherein the data is classified as either 1 or 0, representing positive or negative detection status [14],[15]. Our logistic regression technique seeks to represent the connection between the target factor and the variables that predict it by locating the best fit that is detection possible.

Data preparation

To address the problem of the ML models not understanding the dataset when it is text, preprocessing is used to convert text vectors into numerical data. As an illustration, the terms denote categorical attributes within the documents, and a singular vector will be assigned to each phrase. The technique in question is commonly referred to as vectorization. The methods of CountVectorizer and TF-IDFVectorizer are frequently employed for text vectorization. These vectorization methods are utilized for generating vectorized representations of textual information. The TF-IDFVectorizer differs from the countVectorizer in that it captures the weighted probability of every token concerning the total number of times occurs in a document [16], [17].

CountVectorizer is a popular method for obtaining numerical text data and generating class attributes. Frequent terms in the training text are the only ones considered. The text is converted into a word repetition matrix by applying CountVectorizer using the matrix fit function. This matrix is then used to determine how many times each word appears. [4].

Algorithm 1: Preprocessing dataset
Input: dataset before starting preparation. **Output: set of vocabularies**
Begin:
Stage 1: Based on CountVectorizer, alter text into a list of vocabularies. Stage 2: Remove frequently utilized expressions. Stage 3: Eliminate the terms with the lowest frequency of usage. Stage 4: Remove all stopwords. Stage 5: Transform all vocabularies to lowercase letters. Stage 6: Organize the dictionary in ascending order. The presence of a term is denoted by the numerical value of 1 within the text, while its absence is represented by the numerical value of 0. Stage 7: Iterate through steps 1 to 6 to transform the textual dataset into numerical representations.
End

Training and Testing Method

The study utilized the holdout method, whereby 80% of the dataset was allocated for learning and the remaining 20% for the testing phase [18].

Performance evaluation

The basic values used to evaluate the machine learning algorithm's performance are FP, FN, TP), and (TN), which represent the confusion matrix's basic values by which the Accuracy, Precision, and Recall equations are calculated. and F1-score[19].

True positives refer to cases where expected values are accurate, specifically when the expected category is positive and, in fact, the current positive.

True negatives refers to instances where negative values are accurately predicted. Specifically, this means that the expected class value is no, and the actual class value is also no.

The term **false positives** refer to a situation in which the predicted classification is positive while the correct category is negative.

A **false negative** is a term used to describe a situation where the predicted classification is negative while the actual classification is positive.

The purpose of using the previous variables is to measure the efficiency of the proposed model by applying a set of metrics as shown below:

Accuracy: The ratio of accurately forecast observations to total observations . In mathematics, an equation is defined formally as [20], [21], [22]:

$$\boldsymbol{accuracy} = \frac{(\boldsymbol{TP} + \boldsymbol{TN})}{(\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN})}$$

Precision: It is the ratio of true positives to the set of true positives and false positives. Formally, the formula is defined as::

$$\mathbf{Precision} = \frac{\boldsymbol{TP}}{(\mathbf{TP} + \mathbf{FP})}$$

Recall: It is the ratio of true positives to the set of true positives and false negative. Formally, the formula is defined as. Formally, the equation is defined as:

$$\mathbf{Recall} = \frac{\boldsymbol{TP}}{(\mathbf{TP} + \mathbf{FN})}$$

F1-score: The F1-score is calculated as the harmonic mean of precision and recall. The equation is formally defined as:

$$\mathbf{F1\ score} = \frac{\mathbf{2} * (\mathbf{recall} * \mathbf{precision})}{(\mathbf{recall} + \mathbf{precision})}$$ [23] ,[24], [25]

FRAMEWORK

Building a framework to detect SQLIA against the cloud computing environment requires many stages, as the framework in this study involves several steps:

1. Requests to be submitted to the environment of cloud computing are entered through the first layer, which is representative of the user layer.
2. The second layer, which represents the safety layer, is composed of several stages, as will be described below:

    - The first stage: The first stage is to collect data on the study problem
    - The second stage: conducting the process of preparing the dataset to organize it in a way that is compatible with the machine learning algorithms.
    - The third stage: dividing the dataset into a group to learn the approach ML and another group to test the approach.
    - The fourth stage: applying the logistic regression model to the training data set to adjust and configure the model to classify new requests.
    - Fifth stage: using the test dataset to test the model.
    - The sixth phase involves assessing the model with a confusion matrix and a range of effectiveness efficiency metrics.

This layer contains a logistic regression model that classifies sent requests as to whether they have harmful payloads.

3. The third layer: involves the cloud computing environment layer.

The diagram below describes the Framework of the proposed model:

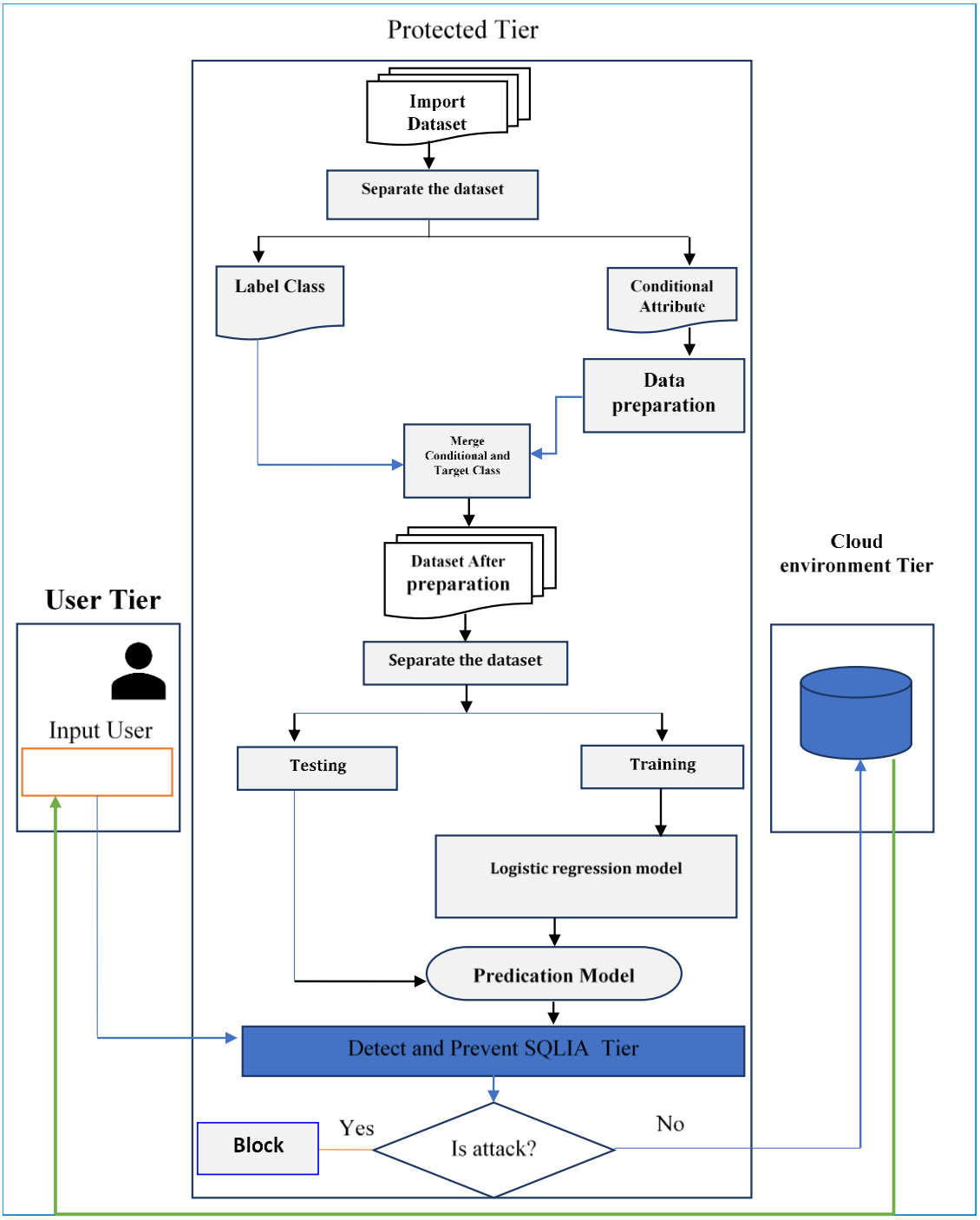

Figure 1. Block Diagram for Proposed Framework

RESULT

This section presents the results using 3780 cases containing normal and harmful payloads. The table below shows the results of a set of measures used to determine the efficiency and Accuracy of the model in identifying and classifying submitted queries.

Table 1. Outcomes of the LR approach

Datasets: SQLIA Volume dataset: 18900 instance Learning stage: (15120, 10678) Testing stage: (3780, 10678)				
Accuracy	**Precision**	**Recall**	**F1-score**	**Time Complexity**
98.78	99.82	98.60	99.21	0.1514
TP	**FP**	**TN**	**FN**	
2892	5	842	41	

In this part, a comparison will be made between the previous study clarified in the second section of this study and the results obtained from the application of this model. Where most of the previous research did not mention the accuracy of the classification and the time it takes to discover and classify the type of payload sent, as well as the method of protection. Is the protection layer adjacent to the user layer or the server layer.

CONCLUSION

As a result of the increasing uses of cloud computing, it has become necessary to provide the necessary protection and identify attacks against it at the required time to maintain the principles of basic information security (CIA), in addition to preserving customer and institutional data. This study suggested a model to identify the type of request sent by a user to address malicious payloads containing SQLIA that threaten users' and organizations' data in a cloud computing environment. This is done by building a model that uses a dataset containing samples containing both harmful and benign payloads. This model works as an intermediate protection layer between the user layer and the cloud computing environment layer. The most important contributions made by this model are:

- Create a model that acts as a layer that separates the data layer and the user layer, to enhance security and prevent unauthorized individuals from accessing the data.
- When machine learning models are applied, they produce a set of values (FP, FN, TP, TN), where the accuracy of these values affects the confidentiality, integrity and availability of the data. Therefore, when the value is FN, the model has classified the malicious payloads as normal, allowing a malicious user to access enterprise and user data, violating data confidentiality, integrity, and availability.
- If the result is FP, it concludes that the model classified normal payloads as malicious, resulting in a data availability violation and not allowing authorized people to access their data.
- But if the values are TP and TN, it is concluded that the model has accurately classified the requests. Therefore, when building a SQLIA detection model, false positives and negatives should be reduced to as few as possible because of their impact on basic information security principles.

REFRENCES

[1] P. J. Sun, “Security and privacy protection in cloud computing: Discussions and challenges,” J. Netw. Comput. Appl., vol. 160, p. 102642, 2020, doi: 10.1016/j.jnca.2020.102642.

[2] S. M. Khudaier and B. A. Mahmood, “A Review of Assured Data Deletion Security Techniques in Cloud Storage,” Iraqi J. Sci., vol. 64, no. 5, pp. 2492–2511, 2023, doi: 10.24996/ijs.2023.64.5.33.

[3] T. Y. Wu, C. M. Chen, X. Sun, S. Liu, and J. C. W. Lin, “A Countermeasure to SQL Injection Attack for Cloud Environment,” Wirel. Pers. Commun., vol. 96, no. 4, pp. 5279–5293, 2017, doi: 10.1007/s11277-016-3741-7.

A. H. Farhan and R. F. Hasan, “Detection SQL Injection Attacks Against Web Application by Using K-Nearest Neighbors with Principal Component Analysis,” in Proceedings of Data Analytics and Management: ICDAM 2022, Springer, pp. 631–642,2023.

[4] M. A. M. Ariffin, M. F. Ibrahim, and Z. Kasiran, “API vulnerabilities in cloud computing platform: Attack and

detection," Int. J. Eng. Trends Technol., no. 1, pp. 8–14, 2020, doi: 10.14445/22315381/CATI1P202.
[5] Dhivya R and Dharshana R, "Security Attacks Detection in Cloud using Machine Learning Algorithms," Int. Res. J. Eng. Technol., vol. 223, no. 01, pp. 309–314, 2008, [Online]. Available: www.irjet.net
[6] K. W. I and V. Hou, "Detection Method of SQL injection Attack in Cloud," pp. 487–493, 2018.
[7] D. Tripathy, R. Gohil, and T. Halabi, "Detecting SQL Injection Attacks in Cloud SaaS using Machine Learning," Proc. - 2020 IEEE 6th Intl Conf. Big Data Secur. Cloud, BigDataSecurity 2020, 2020 IEEE Intl Conf. High Perform. Smart Comput. HPSC 2020 2020 IEEE Intl Conf. Intell. Data Secur. IDS 2020, pp. 145–150, 2020, doi: 10.1109/BigDataSecurity-HPSC-IDS49724.2020.00035.
[8] B. Shunmugapriya and Dr. B. Paramasivan, "Protection Against SQL Injection Attack in Cloud Computing," Int. J. Eng. Res., vol. V9, no. 02, pp. 502–510, 2020, doi: 10.17577/ijertv9is020273.
[9] S. O. Uwagbole, W. J. Buchanan, and L. Fan, "Applied Machine Learning predictive analytics to SQL Injection Attack detection and prevention," Proc. IM 2017 - 2017 IFIP/IEEE Int. Symp. Integr. Netw. Serv. Manag., pp. 1087–1090, 2017, doi: 10.23919/INM.2017.7987433.
[10] Tchernykh, U. Schwiegelsohn, E. ghazali Talbi, and M. Babenko, "Towards understanding uncertainty in cloud computing with risks of confidentiality, integrity, and availability," J. Comput. Sci., vol. 36, 2019, doi: 10.1016/j.jocs.2016.11.011.
[11] N. Z. Khidzir, K. Azhar, and M. Daud, "Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016)," Reg. Conf. Sci. Technol. Soc. Sci. (RCSTSS 2016), no. Rcstss, pp. 229–237, 2018, doi: 10.1007/978-981-13-0074-5.
[12] O. S. F. Shareef, R. F. Hasan, and A. H. Farhan, "Analyzing SQL payloads using logistic regression in a big data environment," J. Intell. Syst., 2023, [Online]. Available: https://doi.org/10.1515/jisys-2023-0063
[13] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," Informatics Med. Unlocked, vol. 17, no. March, p. 100179, 2019, doi: 10.1016/j.imu.2019.100179.
[14] N. Y. Anad AlSaleem, "Network Traffic Prediction Based on Time Series Modeling," Iraqi J. Sci., vol. 64, no. 8, pp. 4160–4168, 2023, doi: 10.24996/ijs.2023.64.8.36.
[15] H. El Rifai, L. Al Qadi, and A. Elnagar, "Arabic text classification: the need for multi-labeling systems," Neural Comput. Appl., vol. 34, no. 2, pp. 1135–1159, 2022, doi: 10.1007/s00521-021-06390-z.
[16] R. F. Hasan, O. S. F. Shareef, and A. H. Farhan, "Analysis of the False Prediction of the Logistic Regression Algorithm in SQL Payload Classification and its Impact on the Principles of Information Security (CIA)," Iraqi J. Comput. Sci. Math., vol. 4, no. 4, pp. 191–203, 2023, doi: 10.52866/ijcsm.2023.04.04.015.
[17] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," ICT Express, vol. 8, no. 2, pp. 183–188, 2022, doi: 10.1016/j.icte.2021.05.001.
[18] O. D. Madeeh and S. By, "Customer Basket Prediction for Stock Market using Data Mining Techniques," 2020.
[19] H. Farhan and R. F. Hasan, "Using random forest with principal component analysis to detect SQLIA," in AIP Conference Proceedings, 2023.
[20] O. D. Madeeh and H. S. Abdullah, "An Efficient Prediction Model based on Machine Learning Techniques for Prediction of the Stock Market," J. Phys. Conf. Ser., vol. 1804, no. 1, 2021, doi: 10.1088/1742-6596/1804/1/012008.
[21] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," Augmented Human Research, vol. 5, no. 1. 2020. doi: 10.1007/s41133-020-00032-0.
[22] F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," Int. J. Inf. Technol., vol. 13, no. 4, pp. 1503–1511, 2021, doi: 10.1007/s41870-020-00430-y.
[23] A. H. Farhan and R. F. Hasan, "Detection SQL injection attacks against web application by using support vector machine with principal component analysis," in AIP Conference Proceedings, 2024.
[24] A. R. F. H. A. H. Farhan, O. S. F. Shareef, "The Effect of False Predictions of Machine Learning on the Security of the Big Data Environment," Iraqi J. Sci., vol. 66, no. 1.