A REVIEW ON MACHINE LEARNING CLASSIFICATION-BASED ALGORITHM FOR DETECTION OF DATA LEAKAGES IN A CLOUD COMPUTING ENVIRONMENT

Abba Mohammed Yayaji[1], Badamasi Imam Ya'u[2], Fatima Umar Zambuk[3], Mustapha Abdulrahman Lawal[4]

[1] Research Scholar, Department of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi, Nigeria

[2] Research Supervisor, Department of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi, Nigeria

[3] Research Supervisor, Department of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi, Nigeria

[4] Research Reviewer, Department of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi, Nigeria

ABSTRACT

Cloud computing has introduced new security challenges, particularly in preventing data leakages. To address this issue, we propose an improved machine learning classification-based algorithm for detecting data leakages in cloud computing environments. Our algorithm leverages machine learning techniques to identify anomalous patterns and behaviors that may indicate potential leakages. We utilize a comprehensive feature selection process and consider factors such as file access patterns, user behavior, and network traffic to enhance detection capabilities. A large dataset was used to train the algorithm and high-tech machine-learning algorithms was employed. Experimental results exhibit higher performance compared to conventional methods in terms of accuracy and efficiency. The real-time detection capabilities of the algorithm enable organizations to prevent and mitigate data breaches, safeguarding their sensitive information. Future research can focus on refining and integrating the algorithm into existing cloud security frameworks for enhanced data protection.

Keyword: - *Deep Learning, Machine Learning, Cloud Computing and Data Leakage*

1. INTRODUCTION

Data leakage is a significant problem in the modern business environment because it must be safeguarded against unauthorized access. Data leakage refers to the unintended or deliberate exposure of confidential business data to unauthorized parties. It is crucial to prevent unauthorized users from exploiting critical data, including intellectual property rights, patents, functionality, and other pertinent information. This vital organizational information has been shared with stakeholders outside the corporation on a regular basis. Consequently, identifying the individual or entity responsible for the data breach can be challenging [6].

An intrusion detection system is a computer or network monitoring system that detects intruders. An intrusion prevention system (IPS) is a security mechanism that employs a variety of security technologies to limit harmful network traffic while preventing infiltration in real-time. The majority of non-manual defect detection systems employ fault detection approaches that are model-based, signal-based, or knowledge-based. Data gathering, feature extraction, and data classification are three critical components of these intelligent systems [7].

To effectively address these difficulties, cognitive data processing abilities are required. Machine learning has been used in this context to handle challenging issues across several disciplines. This study focuses on using machine learning approaches to enhance the capabilities of a standard DLP system for proactive security against insider threats.

2. LITERATURE REVIEW

Data leakage refers to the unauthorized transfer of sensitive or confidential information from a computer or data center to an external or untrusted source. Several data losses have occurred throughout the previous several decades, wreaking havoc on businesses, and these losses have increased in recent years. It is also necessary to disclose that private users are also victims of data loss, and it is difficult to determine the level of data loss that has occurred (Gupta & Singh, 2022).

2.1 Causes of Data Leakage

One of the most challenging issues in reducing data leakage is the multitude of factors that can lead to data loss within an organization. Unfortunately, there is no single tool or straightforward solution available to effectively address these diverse losses of data. However, to effectively manage these risks, a solution must be developed that takes into account the various origins of data loss, which can be categorized into people, processes, and technology. (Ernst & Young, 2011). It must be noted that 70% of businesses that suffer a major data loss are out of business within a period of 18 months according to TMT prediction 2016. Figure 1 illustrates their assessments.

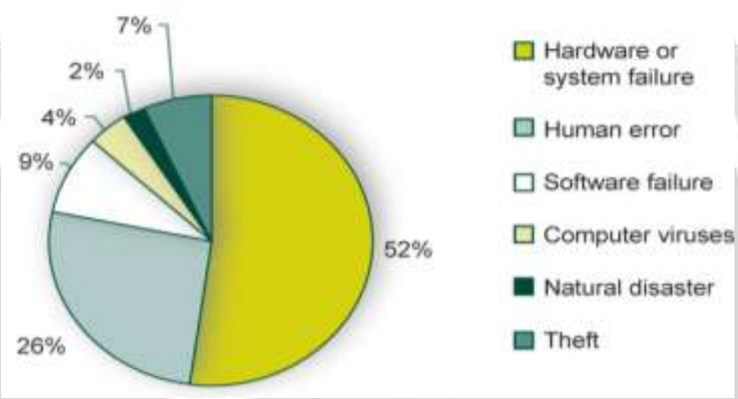

Fig -1: Causes of data leakage

2.2 Conventional security for data leakage

DLP systems are designed with the ability to analyze the actual nature of the sensitive data as well as the surrounding context. They also have the ability to provide protection of critical information at the different levels of data state, which are data-in-use (which uses endpoint protection), data-in-transit (that makes use of network monitoring) as well as a data-at-rest (which uses data classification). Lastly, DLP has the ability to protect confidential data through various policies and rules execution responses, such as notification, data auditing, active blocking, data encryption, and data quarantining. This makes DLP systems proactive and dedicated, which makes it different from conventional security mechanisms such as Intrusion Detection Systems (IDS), firewalls and Virtual Private Networks (VPNs). The convention security technologies have less dedication in terms of the content of the data they are protecting.

2.3 Machine Learning techniques for predicting and classifying data leakage

There are several deep learning methods that are being used in predicting and classifying data leakages, such as K-Nearest Neighbors (KNN), support vector machine, Naïve Bayesian Classifier (NBC), Decision Tree, and Random Forest. Here, we use DT tools for classification and regression classifiers.

2.4 Decision tree

One of the benefits of decision trees is their ability to extract information by inferring human-readable rules from data based on attribute values. The subsets are consistently divided into more refined subsets until the size of the split attains a satisfactory level. This entire process of modeling can be depicted as a tree-like structure, and the model is formulated through a sequence of 'if-then' rules. Decision trees are simple to understand, computationally inexpensive, and can function effectively with noisy data.

3. RELATED WORK

Numerous research studies have focused on data security to prevent the unauthorized leakage of sensitive data by company personnel or system insiders. An intrusion detection system is an automated tool that continuously monitors network or computer activities to identify potential intruders. Conversely, an intrusion prevention system is a comprehensive security solution that utilizes a variety of security technologies in order restrict potentially harmful network traffic and prevent real-time infiltration. The majority of non-manual defect detection systems rely on signal-based, model-based, or knowledge-based fault detection methodologies.

Fault detection systems commonly employ artificial neural networks (ANNs) or other classifiers to enhance detection rates. For example, in their study, [6], introduced and assessed data leakage detection in cloud computing environments using classification based on deep learning methods with the CERT dataset. Experimental results shows that the technique is computationally affordable that uses time and space efficiently that identified the data leaker in real time, which was able to defend against several aggressive and passive assaults. However, the primary drawback of this study lies in its inability to enhance the system's capacity to mitigate insider threats and its failure to adapt to a web environment where multiple users regularly access the same data item.

As a computer science/AI reviewer and researcher, I'd like to provide you with a proofread and corrected version of the text with proper grammar and style:

Similarly, in their study, [8] recommended the utilization of cloud computing to identify and prevent data leakage. The system was developed to assist in maintaining the integrity of formatted data and in safeguarding against the inadvertent leakage of unstructured data during the distribution process. The methodology encompasses a concise explanation of data leakage detection, along with a research approach for the identification of individual data leaks. However, it should be noted that the strategy did not specifically target particular institutions or economic sectors such as banking and finance.

Furthermore, [5] have presented an efficient security system in a cloud computing environment for intrusion detection that employs a hybrid deep learning method. They utilized the improved heap optimization (IHO) technique for data processing, ensuring data quality by removing extraneous data from the dataset. Additionally, they implemented the Chaotic Red Deer Optimization (CRDO) method for optimal feature selection. This method played a crucial role in reducing dimensionality when dealing with large datasets. The Deep Kronecker Neural Network (DKNN) was employed for the detection and classification of cloud attacks and intrusions. The system's effectiveness was evaluated using two datasets, namely, CSE-CIC-IDS2018 and DARPA IDS, with the results compared to those of other intrusion detection systems.

Also, [2] applied the centroid document classifier approach, which classified documents into the correct category resulting in confidentiality not being compromised and used cluster data properly to prevent data leakage. The approach could only obtain records on a specified issue, such as government borrowing and catastrophe.

Additionally, [9], analyzed data science notebook code from 100,000 analyzed publicly available notebooks in their work titled “Data Leakage in Notebooks: Static Detection and Better Processes”, and evaluated it using Sklearn fit and predict functions. It was able to develop a static code analysis approach that detects different forms of data leakage in notebooks. The method failed to identify repeated assessments that were not contained in the notebook and test data that were not present in the notebook. The methodology also misses some leakage while producing a few false positives and underestimating the reported leakage, according to the study.

Furthermore, [1] in his work “Cloud Computing Security for multi-cloud Service Providers: controls and Techniques in our modern threat landscape”, a comprehensive awareness of contemporary cyber dangers and how to address them was acquired, as well as a detailed exploration of cloud security elements and recommendations for multi-cloud service providers. The study delves into various facets of cloud security, offering detailed insights and proposing best practice recommendations for multi-cloud service providers. Challenges in providing security for a multi-cloud environment were discussed and solutions were offered to overcome these challenges. However, the work fails to address the issue of regulatory enforcement across contexts.

They [4] have devised a machine learning-based technique for detecting data breaches through large-scale distributed query processing. They have introduced the Dynamic Inference-based Rule Set Reduction algorithm, which is grounded in deep machine learning. This algorithm employs an adaptive threshold to generate rule sets and offers a method for dynamic inference-based Rule Set Reduction. This approach not only aids in improving time complexity but also reduces the loss of dataset attribute information. Its equivalence demonstrates a virtually 89% accuracy. The algorithm fails to achieve a safer and faster distributed query processing.

Table -1: Summary of related work

Reference	Method	Findings	Limitation
(Patil et al., 2022)	Data Leakage Detection in Cloud Computing Environment Using Classification Based on Deep Learning Architectures	The techniques are computationally affordable uses time and space efficiently that identify the data leaker in real-time, and are able to defend against several aggressive and passive assaults	The study was unable to enhance the system's capability to mitigate insider threats and also failed to adapt the system to a web environment where multiple users frequently access the same data item.
(V. Singh et al., 2023)	Data Leakage Detection and Prevention in Cloud Computing	This review offers a succinct summary of data leakage detection and presents a research methodology for pinpointing individual instances of data leakage.	The DLP strategy did not target large enterprises or specific business sectors such as finance and banking.
(Mayuranathan et al., 2022)	Hybrid deep learning method	Improved data pre-processing to improve data quality by removing unnecessary data. Optimal feature selection for best features was achieved. Intrusion detection, cloud attack and classification was illustrated using hybrid deep Kronecker neural network (DKNN)	Regulatory enforcement issues across contexts
(Gupta, Mittal, Tiwari, Agarwal, & Singh, 2022)	TIDF-DLPM: Term and Inverse Document Frequency Based Data Leakage Prevention Model	The centroid document classifier technique categorized documents correctly, ensuring that confidentiality was not jeopardized and properly employed clustered data to prevent data leakage.	Capable only of retrieving documents on specific topics, such as government loans and disasters.
(Yang et al., 2022)	Data Leakage in Notebooks: Static Detection and Better Processes	Developed a static code analysis method for detecting various types of data leaks in notebooks	The system failed to identify multiple evaluations that are not present in the notebook. It overlooks some leaks while also producing a few false positives and minimizing reported leakage.

(Achar, 2022)	Cloud Computing Security for Multi-Cloud Service Providers: Controls and Techniques in our Modern Threat Landscape	Efficiently identified modern threat to data confidentiality	Regulatory enforcement issues across contexts

4. RESEARCH GAP

The majority of current insider threat approaches primarily concentrate on detecting common insider attack scenarios. The approach proposed in [6] is designed to identify data leakage within cloud computing by employing deep learning architectures for data categorization. Input data is collected as network data and subsequently processed to reduce noise and smooth it out. The classification task is performed using the Generative Regression kernel Support Vector Machine (SVM). Experimental results are evaluated using the following metrics: RMSE (Root Mean Square Error), SNR (Signal-to-Noise Ratio), recall, F-1 score, precision and accuracy. The model presents realistic methods for addressing class imbalance and potential bias concerns to develop a system that identifies insider data leakage with 97% accuracy, 67% recall, 92% precision, 66% F-1 score, 61% SNR, and 62% RMSE. In other circumstances, however, accuracy might be deceiving. Precision, recall, and F-score can assist us in comprehending how accurate the accuracy provided is for a certain problem. Thus, some critical metrics in the research [6] such as recall and the f-score were too low (67% and 66% respectively). Hence, the network situation addressed by the study, needs the data leakage to be detected with high consistent results for all the metrics. Furthermore, it is generally recognized that the main disadvantage of SVM arises during its training phase. This is attributed to the necessity of solving a Quadratic Programming Problem (QPP) to train this classifier, which is a process that demands significant computational resources. Support Vector Machine (SVM), as utilized in [6], offers essential qualities such as a strong mathematical foundation and superior generalization capacity when compared to other classification approaches. On the other hand, the main disadvantage of SVM occurs during the training stage, which is computationally costly and heavily reliant on the input dataset size. SVM's overall performance often degrades with larger datasets, both in terms of accuracy and computing cost.

5. CONCLUSIONS

Cloud computing has introduced new security challenges, particularly in preventing data leakages. To address this issue, we propose an improved machine learning classification-based algorithm for detecting data leakages in cloud computing environments. Our algorithm leverages machine learning techniques to identify anomalous patterns and behaviors that may indicate potential leakages. We utilize a comprehensive feature selection process and consider factors such as file access patterns, user behavior, and network traffic to enhance detection capabilities. The algorithm is trained using a large dataset and employs state-of-the-art machine learning algorithms. Experimental results performed higher over the traditional methods in terms of accuracy and efficiency. The algorithm's real-time detection capabilities enable organizations to prevent and mitigate data breaches, safeguarding their sensitive information. Future research can focus on refining and integrating the algorithm into existing cloud security frameworks for enhanced data protection.

REFERENCES

[1]. Achar, S. (2022). Cloud Computing Security for Multi-Cloud Service Providers: Controls and Techniques in our Modern Threat Landscape. *International Journal of Computer and Systems Engineering, 16*(9), 379-384.

[2]. Gupta, I., Mittal, S., Tiwari, A., Agarwal, P., & Singh, A. K. (2022). TIDF-DLPM: Term and Inverse Document Frequency based Data Leakage Prevention Model. *arXiv preprint arXiv:2203.05367*.

[3]. Gupta, I., & Singh, A. K. (2022). A Holistic View on Data Protection for Sharing, Communicating, and Computing Environments: Taxonomy and Future Directions. *arXiv preprint arXiv:2202.11965*.

[4]. Kiranmai, M., & Haritha, D. (2021). Detection of Data Leaks through Large Scale Distributed Query Processing using Machine Learning. *International Journal of Advanced Computer Science and Applications, 12*(12).

[5]. Mayuranathan, M., Saravanan, S., Muthusenthil, B., & Samydurai, A. (2022). An efficient optimal security system for intrusion detection in cloud computing environment using hybrid deep learning technique. *Advances in Engineering Software*, *173*, 103236.

[6]. Patil, R. C., Kumar, A., Narmadha, T., Suganthi, M., Rao, A. V. S. R., & Rajesh, A. (2022). Data Leakage Detection in Cloud Computing Environment Using Classification Based on Deep Learning Architectures. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(2s), 281-285.

[7]. Singh, P., & Ranga, V. (2021). Attack and intrusion detection in cloud computing using an ensemble learning approach. *International Journal of Information Technology*, *13*, 565-571.

[8]. Singh, V., Raj, M., Gupta, I., & Sayeed, M. A. (2023). Data Leakage Detection and Prevention Using Cloud Computing *Sustainable Computing* (pp. 159-169): Springer.

[9]. Yang, C., Brower-Sinning, R. A., Lewis, G. A., & Kästner, C. (2022). Data Leakage in Notebooks: Static Detection and Better Processes. *arXiv preprint arXiv:2209.03345*.