Open Access Research Journal of
Science and Technology
Journals home page: https://oarjst.com/
ISSN: 2782-9960 (Online)

OARJ OPEN ACCESS RESEARCH JOURNALS

(RESEARCH ARTICLE)

Machine learning for secure API obfuscating in cloud applications with app cloaking techniques

Upakar Bhatta *

Information Technology Management, Central Washington University, Ellensburg, WA, USA.

Open Access Research Journal of Science and Technology, 2025, 15(01), 012–019

Publication history: Received on 29 July 2025; revised on 09 September 2025; accepted on 11 September 2025

Article DOI: https://doi.org/10.53022/oarjst.2025.15.1.0111

Abstract

App cloaking is a powerful technique for safeguarding cloud-based applications against cyber threats by obfuscating sensitive data. It enables organizations to secure their applications from public internet by concealing critical data and hiding application communication routes. With app cloaking, organization can isolate their applications and hide their IP addresses from the internet, making them invisible to unauthorized users. This paper leverages a machine learning (ML) approach to predict the effectiveness of cloaking techniques in cloud-based applications, enabling organizations to enhance their security measures. The proposed ML techniques analyze various features in a sample AWS dataset, including API Endpoint, AWS Service Type, Traffic Volume, Obfuscation Method, Detection Attempts, Obfuscation Complexity, Threat Detected, Response Time Impact, and Observed Attack Outcome, to evaluate the effectiveness of Cloaking techniques under different circumstances.

Keywords: Machine Learning; User Behavior Analytics; Cybersecurity; Anomaly Detection; Threat Identification

1. Introduction

With the rapid adoption of cloud technologies, protecting cloud-based applications and services against cyber threats has never been more important [1]. Security and privacy remain significant concern when adopting cloud-based services and applications [2–4]. Safeguarding these applications services from cyber threats requires a layered approach, including the implementation of granular access controls that aligns with Zero Trust principles. Application cloaking protects both the application and sensitive data by hiding the application from the internet, enabling organization enhance the security of cloud-based application and data services. App cloaking aligns with Zero Trust principles by enabling granular access control that limits application access to authorized users and devices. This research study explores machine learning techniques to predict the effectiveness of cloaking techniques in cloud-based applications, enabling organizations to enhance their security posture and restrict application access to authorized users and devices. This research study examines the AWS sample dataset features, including API endpoint, AWS service type, traffic volume, resources usage, obfuscations method, obfuscation complexity, threats detected, cloaking success, and observed attack outcome in cloud services, to predict the effectiveness of cloaking techniques in AWS -based applications.

2. Pre-requisite knowledge

2.1. Cloud computing

Cloud computing is the technology that connects virtual resources, such as computing power, storage, networking, and databases, over the internet to facilitate data sharing. It is one of the fastest growing technologies that utilizes compute, storage, networking, and database tools to share the virtual resources via the Internet. However, organizations face multiple challenges regarding data security and privacy when adopting cloud computing infrastructure [5].

*Corresponding author: Upakar Bhattta

The following are the three cloud computing services that the most organizations today are leveraging:

- Infrastructure as a Service (IaaS): This service offers virtualized computing resources over the internet and provides highest level of flexibility and management control over the infrastructure compared to other services.
- Platform as a Service (PaaS): It provides customer with an underlying infrastructure to deploy their own code.
- Software as a Service (SaaS): This service offers complete software packages over the internet on a subscription basis.

2.2. Big data analytics

Big data analytics is the process of examining large data sets to help businesses understand the hidden patterns and insights. Companies can leverage analytical tools and technologies to study big data in depth and uncover valuable business insights [6]. Data velocity, data variety, and data volume are key terms used in the big data analytics process. Data velocity refers to the speed at which the data is collected. More than 3.5 billion Google searches are conducted per day [7]. Data variety refers to various types of data, such as structured, unstructured, and semi-structured, that may be captured during the data analytic process. Big data analytics deals with a huge volume of data that can come from different sources. For instance, the average mobile traffic was 6.2 Exabyte per month, and Netflix has over 86 million members globally [7].

2.3. Machine learning

Machine learning is a mathematical technique that focuses on building algorithm and statistical models to enable computers to make predictions by analyzing hidden patterns in datasets. It has introduced a new approach for training algorithms using real-time datasets to identify specific patterns andanomalies in network traffic [8]. There are three main categories of machine learning: supervised learning, unsupervised learning, and reinforcement learning.

In supervised learning, the algorithm is trained using a labeled dataset. This approach is further divided into Classification and Regression [9]. In classification tasks, the output is a category, whereas in regression tasks, the goal is to predict a continuous value.

In unsupervised learning, the algorithm is trained using unlabeled data. Since unsupervised learning techniques deal with unlabeled data, the learning process is complex. Due to its complexity, unsupervised learning is used less frequently than supervised learning [10].

In reinforcement learning, a system learns through trial and error, receiving feedback in the form of rewards and punishments. Through this process, the algorithm determines which steps yield the highest rewards [11].

3. Methodology

In this paper, machine learning techniques are explored to predict the effectiveness of cloaking techniques in cloud-based applications, using sample AWS dataset. I demonstrated five machine learning algorithms to predict the accuracy of cloaking techniques. The five machine learning algorithms examined in this paper are Logistic Regression, SVM, Decision Tree, K-Nearest Neighbors, and Random Forest Classifier.

The experimental methodology used in this paper includes

Acquiring Sample AWS Dataset, Data Preprocessing, Exploratory Data Analysis, Split the Dataset into Training and Testing Segments, Apply SMOTE, Select ML Models, Train ML Models, Predict the Output, Find the Accuracy, and Evaluate the Model.

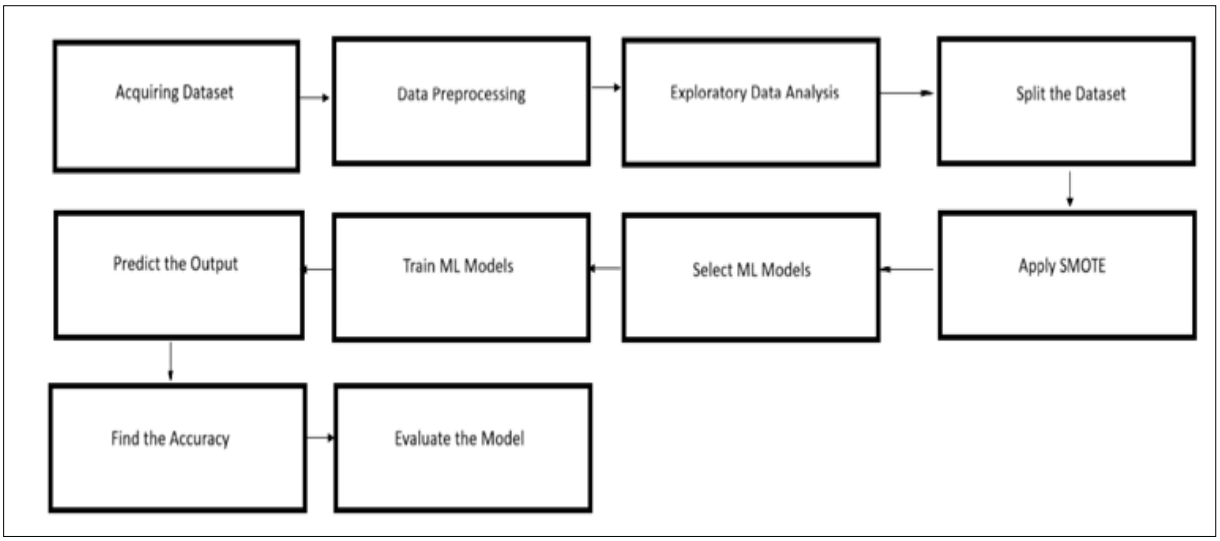

Figure 1 Machine learning implementation workflow

- Acquiring Dataset: Utilized Amazon CloudWatch Logs**,** AWS CloudTrail Logs**,** VPC Flow Logs**,** API Gateway Logs to create asample AWS dataset, collecting 5000 records
- Feature selection: A total of 14 features are included. The key features are API Endpoint, AWS Service Type, Traffic Volume, Obfuscation Method, Detection Attempts, Obfuscation Complexity, Threat Detected, Response Time Impact, and Observed Attack Outcome

3.1. Data preprocessing: Data cleaning, encoding, and normalization were performed

- Exploratory Data Analysis (EDA): To visualize patterns and identify relationship within the dataset, boxplots, histograms, and correlation heatmaps were utilized to detect outliers, observe feature distributions, and identify relations between features
- Split the dataset: The dataset was split into training (80%) and testing (20%)
- SMOTE: SMOTE was applied to address imbalance dataset
- Select the models: Appropriate algorithms were selected to train the ML models
- Train the models: Five ML models were trained. Hyperparameter tuning and cross-validation were employed to optimize model performance and ensure that the ML models are robust
- Model Accuracy and evaluation: Machine learning models were evaluated based on their accuracy.

4. Data analysis

The AWS dataset used in this research includes both numerical and categorical features. Numerical features include Traffic Volume, Resource CPU Usage, Obfuscation Complexity, and Detection Attempts. Categorical features include API Endpoint, AWS Service Type, Threat detected, and Obfuscation method.

The exploratory data analysis methods employed in this research are

4.1. Correlation Heatmap

It visualizes the numerical features and helps identify redundant features. Weakly correlated features, such as Traffic Volume, Resource CPU Usage, are ideal for training the ML models.

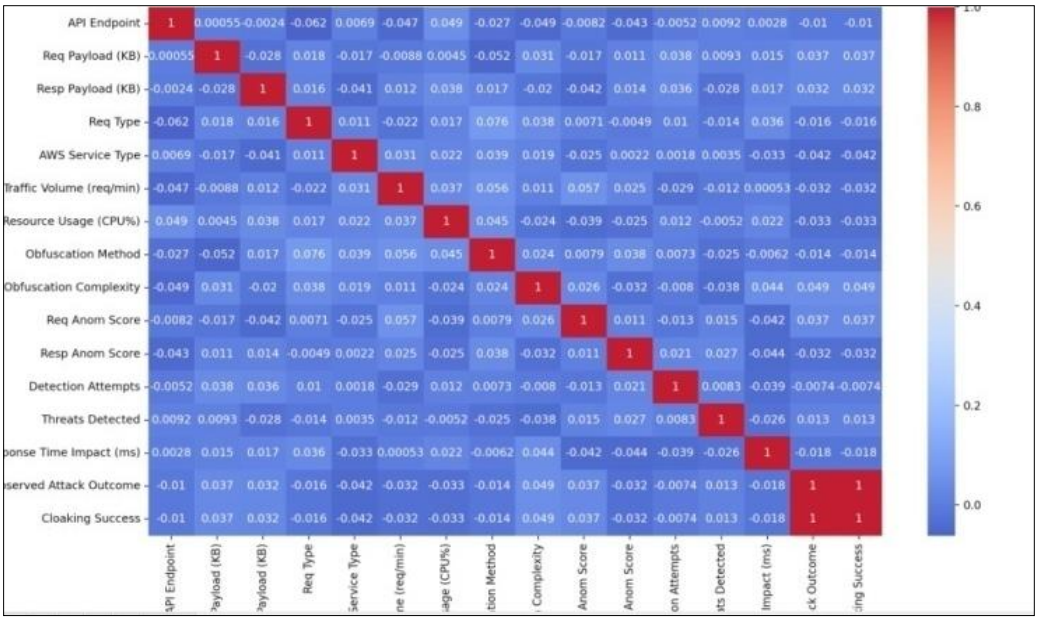

Figure 2 Correlation Heatmap

4.2. Histogram

It analyzes whether the data is skewed, uniform or normally distributed. The histogram of CPU usage and traffic volumes indicates that the data is fairly uniform with no major skew, suggesting that the variance in values positively contributes to model training.

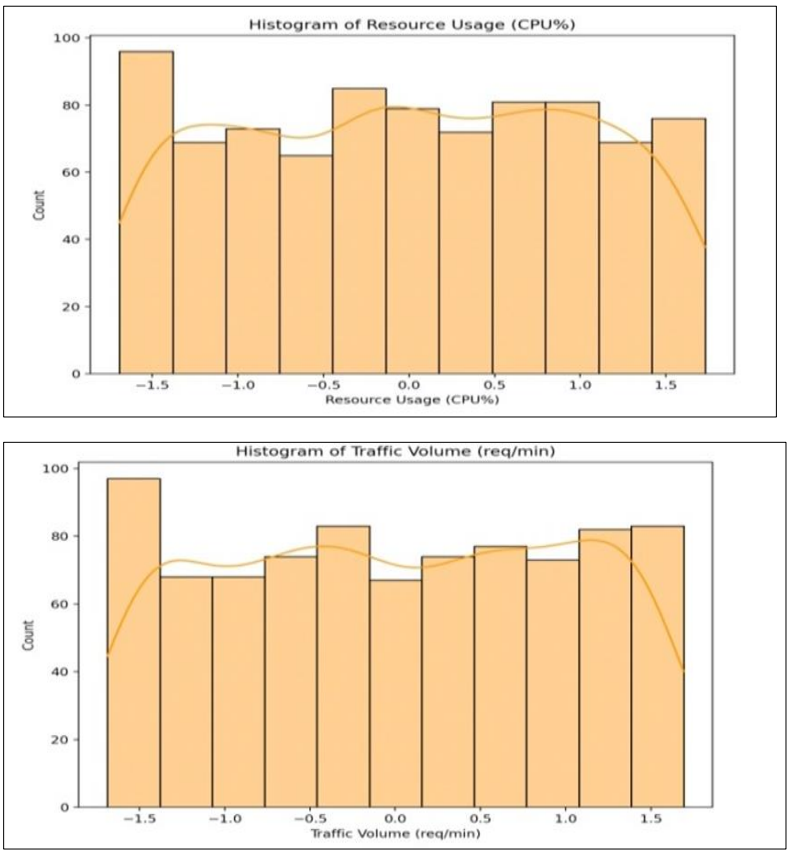

Figure 3 Histogram of Resource Usage and Traffic Volume

4.3. Box plot

The box plot of resource usage and traffic volume visualizes the distribution of CPU usage and traffic volumes across three cloaking success categories, labeled as 0, 1, and 2. It shows that the median is consistent for all three categories, indicating that fewer adjustments are needed for model training.

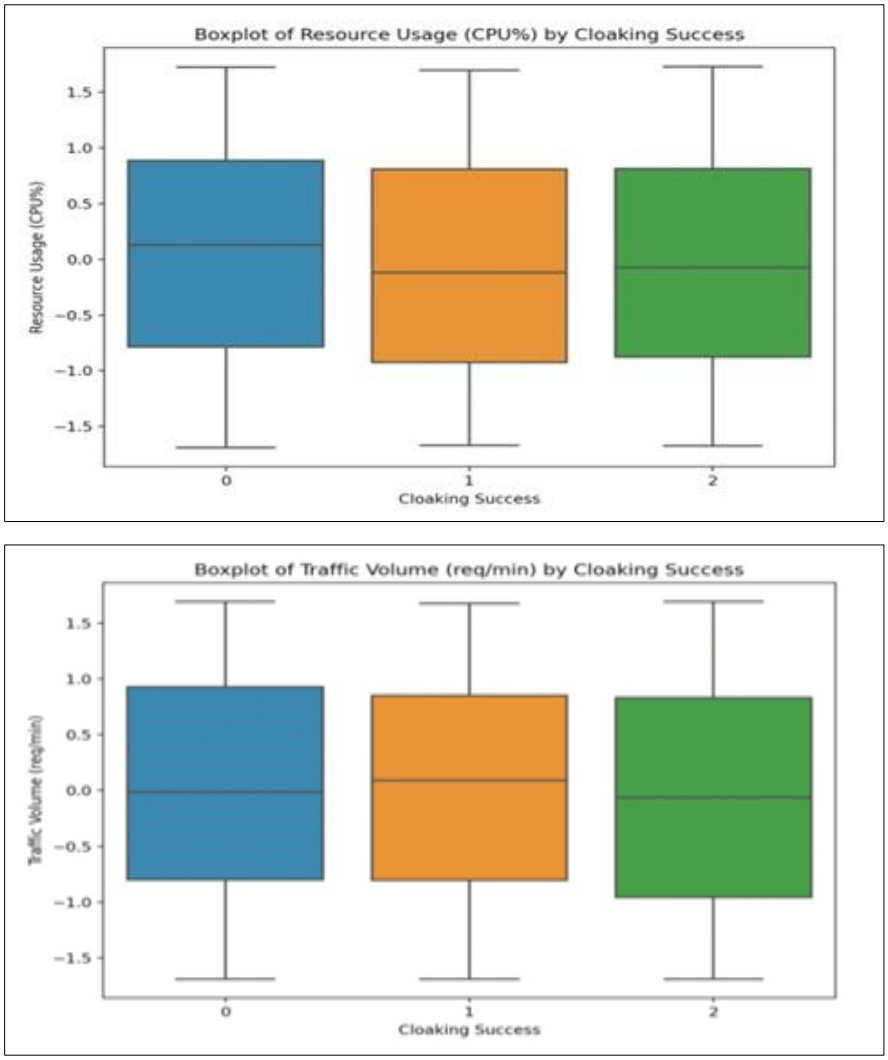

Figure 4 Boxplot of Resource Usage and Traffic Volume

5. Results

This paper leverages a sample AWS dataset created from Amazon CloudWatch Logs, AWS CloudTrail Logs, VPC Flow Logs, API Gateway Logs, containing 5,000 labeled instances of observed attact outcomes across 14 features. These include both numerical features, such as Req Payload (KB), Resp Payload (KB), Traffic Volume, Resource Usage, Obfuscation Complexity, Req/Resp Anom Score, Detection Attempts, Threats Detected, as well as categorical features, such as API Endpoint, Req Type, AWS Service Type, and Obfuscation Method, Obfuscation. The machine learning model demonstrated strong performance, achieving an accuracy of 90% in predicting the effectiveness of cloaking techniques.

Table 1 Comparison of machine learning models based on accuracy and evaluation metrics

ML Model	Accuracy	Metrics	Score
Logistic Regression	0.9043823442082550	Precision	0.89
		Recall	0.88
		F1-score	0.89
Decision Tree	0.9048893489082754	Precision	0.88
		Recall	0.89
		F1-score	0.88
Random Forest Classifier	0.9048893489082754	Precision	0.90
		Recall	0.90
		F1-score	0.89
SVM (Support Vector Machine)	0.9058895481022651	Precision	0.88
		Recall	0.87
		F1-score	0.89
K-Nearest Neighbors Classifier	0.588556607028439	Precision	0.89
		Recall	0.69
		F1-score	0.78

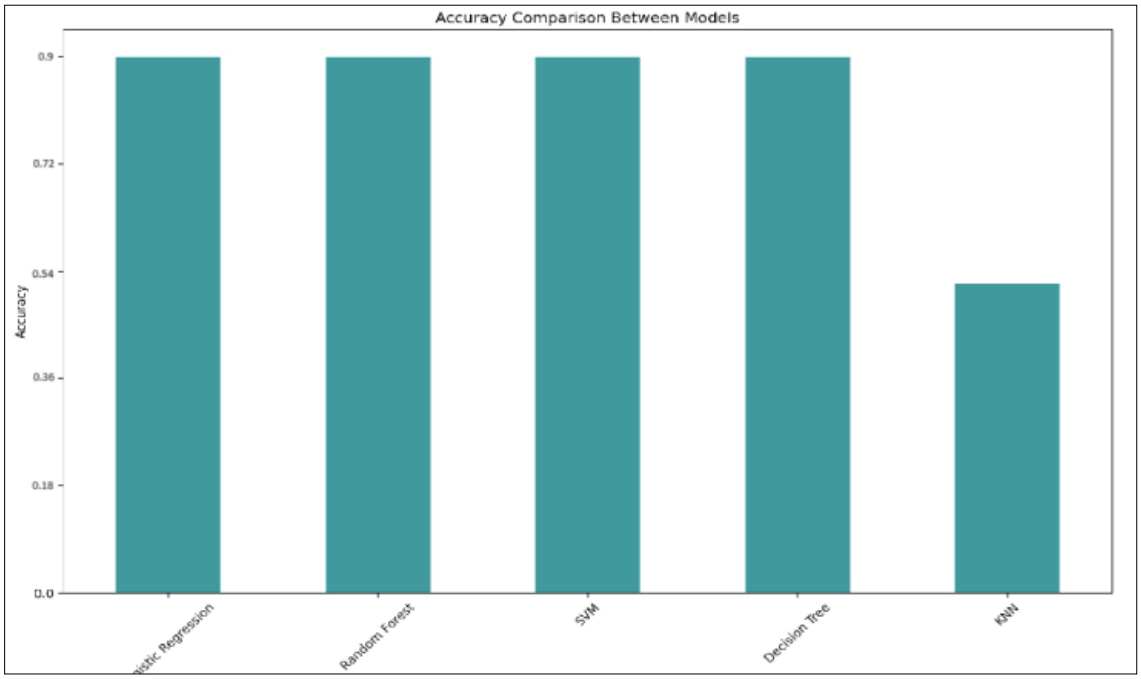

Figure 5 ML Model Accuracy

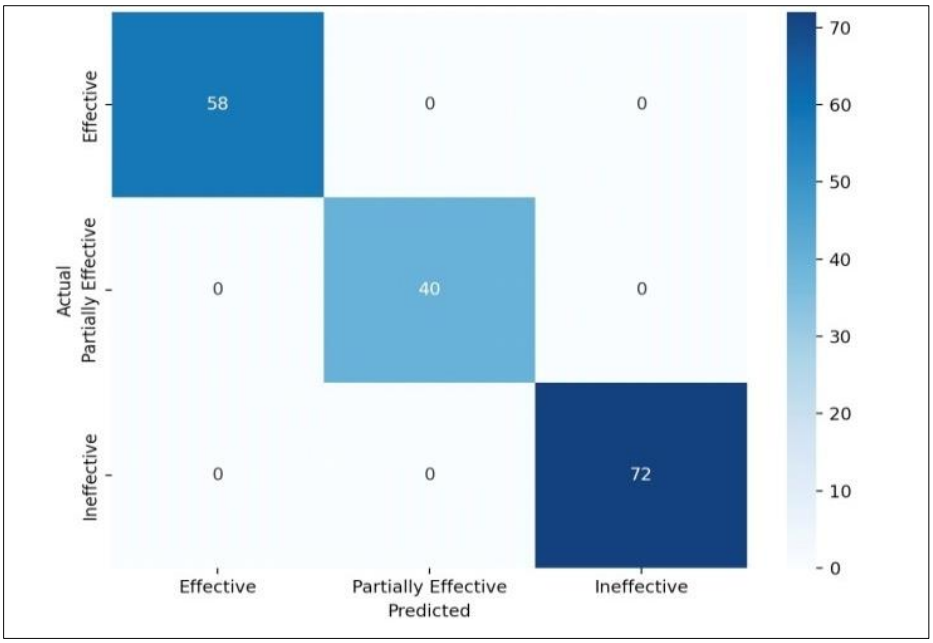

Figure 6 Confusion Matrix

6. Discussion results

6.1. Precision (0.91-0.94)

An ML model with a high precision in this range means few false positive and is correct 94% of the time. A precision rate of 94% is suitable for application where false positive is highly undesirable.

6.2. Recall (0.69-0.95)

An ML model with high recall, such as 0.95, identifies 95% of all actual positive instances. The lower range suggests that a recall of 69% means 31% of positive instances are missed.

6.3. F1 Score (0.78-0.93)

The F1 score ranges 0 to 1 with 1 being the best and 0 being the worst. It is considered as a balance between precision and recall, and a high score of 0.93 indicates that the model performs very well in maintaining this balance. The lower range suggests that there is a room for improvement.

6.4. Justification

Logistic Regression, Decision Trees, Random Forest, and Support Vector Machine classifier provided better accuracy. With precision of 0.93, recall of 0.95, and an F1 score of 0.93, Random Forest's performance indicates the model is highly effective.

Random Forest was chosen for anomaly detection in this research paper for following reasons

- Random forest is known for its high accuracy and works well with the network feature.
- Random Forest ensemble learning method is beneficial for this research as it improves performance compared to other model in detecting anomalies effectively
- Compare to other models, Random Forest can handle both numerical and categorical features without any extensive preprocessing, making it ideal for analyzing the diverse datasets used in this research.

7. Conclusion

App cloaking is a technique that protects cloud-based applications against cyber threats by obfuscating sensitive data. This paper demonstrates machine learning techniques to predict the effectiveness of cloaking methods in cloud-based

applications. It utilized the sample dataset created using AWS services logs, including Amazon CloudWatch Logs, AWS CloudTrail Logs, VPC Flow Logs, API Gateway Logs, to train the model to predict effectiveness of Cloaking techniques. With an accuracy of 90 percent, this paper highlights how effective the App cloaking techniques are in protecting cloud-based application against cyber threats.

Compliance with ethical approval

Acknowledgements

The author declares that there are no acknowledgements for this research study.

Disclosure of Conflict of interest

The author declares that there are no conflicts of interest related to the publication of this manuscript.

Statement of ethical approval

The present research work does not contain any studies performed on animals/human subjects by the author.

Statement of informed consent

The present research work does not involve participants, case studies, surveys, interviews, or any activities requiring informed consent.

References

[1] S. Prasanna, "Top 10 Cloud Computing Challenges in 2019," *Techiexpert*, [Online]. Available: https://www.techiexpert.com/top-10-cloud-computing-challenges-in-2019/

[2] A. Al Hadwer, M. Tavana, D. Gillis, and D. Rezania, "A systematic review of organizational factors impacting cloud-based technology adoption using technology organization-environment framework," *Internet of Things*, vol. 15, p. 100407, 2021. [Online]. Available: https://doi.org/10.1016/j.iot.2021.100407

[3] M.M. Salem and G. H. Hwang: "Critical factors influencing adoption of cloud computing for government organizations in Yemen," *Journal of Distribution Science*, vol. 14, no. 11, pp. 37–47, 2016. [Online]. Available: https://doi.org/10.15722/jds. 14.11.201611.37

[4] Y. Nugraha and A. Martin, "Towards a framework for trustworthy data security level agreement in cloud procurement," *Computers & Security*, vol. 106, p. 102266, 2021. [Online]. Available: https://doi.org/10.1016/j.cose.2021.102266

[5] D. Charalambous, "Cloud Computing and Security Issues that can affect users," Cardiff Metropolitan University, P. 3, 2017.

[6] P. Russom, "Big Data Analytics", *TDWI Best Practices Report*, Fourth Quarter, 2011, pp. 4-5, 22.

[7] A.S. Gills, 5V's of Big Data Management," Mar. 2021. Accessed: Jun. 202. [online]. Available: https://libraryguides.vu.edu.au/ieeereferencing/webbaseddocument

[8] V. Kumar, D. Sinha, A. K. Das, S. C. Pandey, and R. T. Goswami," An integrated rule-based intrusion detection system: Analysis on UNSW-NB15 data set and the real time online dataset," *Cluster Computing*, vol. 23, pp. 1397–1418, 2020. [Online]. Available: https://doi.org/10.1007/s10586-019-03008-x

[9] Wikipedia contributors, "Supervised learning," *Wikipedia, The Free Encycloperdia*, Jul. 12, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Supervised_learning&oldid=1033187514

[10] S. Mishra, "Unsupervised Learning and Data Clustering," *Towards Data Science*, May 20, 2021. [Online]. Available: https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a

[11] B.Osinski and K. Budek, "What is Reinforcement Learning? The Complete Guide," *deepsense.ai*, Jul. 5, 2018. [Online]. Available: https://deepsense.ai/what-is-reinforcement-learning-the-complete-guide/