International Journal of Applied Engineering Research ISSN 0973-4562 Volume 17, Number 4 (2022) pp. 332-337
 https://dx.doi.org/10.37622/IJAER/17.4.2022.332-337

Machine Learning based Security for Cloud Computing: A Survey

Munish Saran, Rajan Kumar Yadav, Upendra Nath Tripathi

Department of Computer Science, DDU Gorakhpur University, Gorakhpur-273001, Uttar Pradesh, India.

Abstract

Cloud computing is a computing paradigm that provides on-demand, scalable as well as measured services to the end uses. In today's era almost each and every business has huge dependency on this computing technology in terms of cost-saving, infrastructure, development platform, data processing, data analytics etc. The services provided by the cloud service providers (CSP) can be consumed by the end users anytime, anywhere by the web application over the internet. The security of the cloud infrastructure is of utmost importance and several research work involving various technologies are utilized so as to provide better and more accurate defence mechanism against cloud attacks. Machine learning is a technology that has proved to produce better results in securing the cloud environment in the recent times. Machine learning algorithms are trained on the various authentic datasets to build models that can automate the process of detecting the cloud attacks with higher accuracy in comparison with any other technology. This paper reviews some of the latest research papers that have employed machine learning as a security mechanism against cloud attacks.

Keywords - Cloud Computing, Machine Learning, Intrusion Detection System, Datasets, Supervised Machine Learning, Unsupervised Machine Learning, Reinforcement Learning, Deep Neural Network.

1 INTRODUCTION

The security of cloud environment is the biggest concern in the recent times. Even the big cloud service providers which has enough security measures such as Amazon, Google etc also suffers from several cloud attacks which are reported time to time on regular basis. Cloud security can be broadly categorized under five different categories namely information security, identity security, network security, infrastructure security and software security. Machine Learning as a Service (MLaaS) is the service model that is utilized by the cloud computing in order to enhance the defence strategy against several cloud attacks. Several Intrusion Detection System has been developed with the help of machine learning algorithms that has improved the accuracy of detecting the attacks and allowing the smooth business operations to carry on.

2 Theoretical Background

This section describes the background details for cloud computing and machine learning thus providing a brief idea about both of them.

2.1 Cloud Computing

Cloud computing is a computing paradigm that provides on-demand, scalable, measured and secure services to the end users over the internet. It is due to these benefits cloud computing paradigm finds a very large set of use cases. There are many cloud service providers in the market today that offer variety of cloud services to their customers. Some of them are Amazon Web Services (AWS), Microsoft Azure, IBM Cloud, Google Cloud, Oracle Cloud, Alibaba Cloud, etc.

2.1.1 Characteristics of Cloud Computing

According to NIST there are primarily five essential characteristics of cloud computing. [1, 2]

On-demand self-service- This characteristic states that the cloud services are made available to the end users on their demand and without the intervention of the cloud service provider.

Rapid elasticity- Cloud based applications are capable to handle the rapid increase or decrease in the demand of the services accordingly without the resource shortage or downtime in the business.

Measured Service- The services provided by the cloud service provider to the end user are billed on a measured basis, i.e. consumers availing the cloud services are charged only with for their service consumption and are free to stop consuming the services any time.

Broad network access- Cloud services are available and can be accessed from a wide range of thin clients such as mobile devices, laptop, desktop, PDAs etc.

Resource Pooling- Cloud computing resources such as memory, storage, processing unit, network bandwidth etc are pooled by the cloud service provider in order to serve the request of multiple customers.

International Journal of Applied Engineering Research ISSN 0973-4562 Volume 17, Number 4 (2022) pp. 332-337
© Research India Publications. https://dx.doi.org/10.37622/IJAER/17.4.2022.332-337

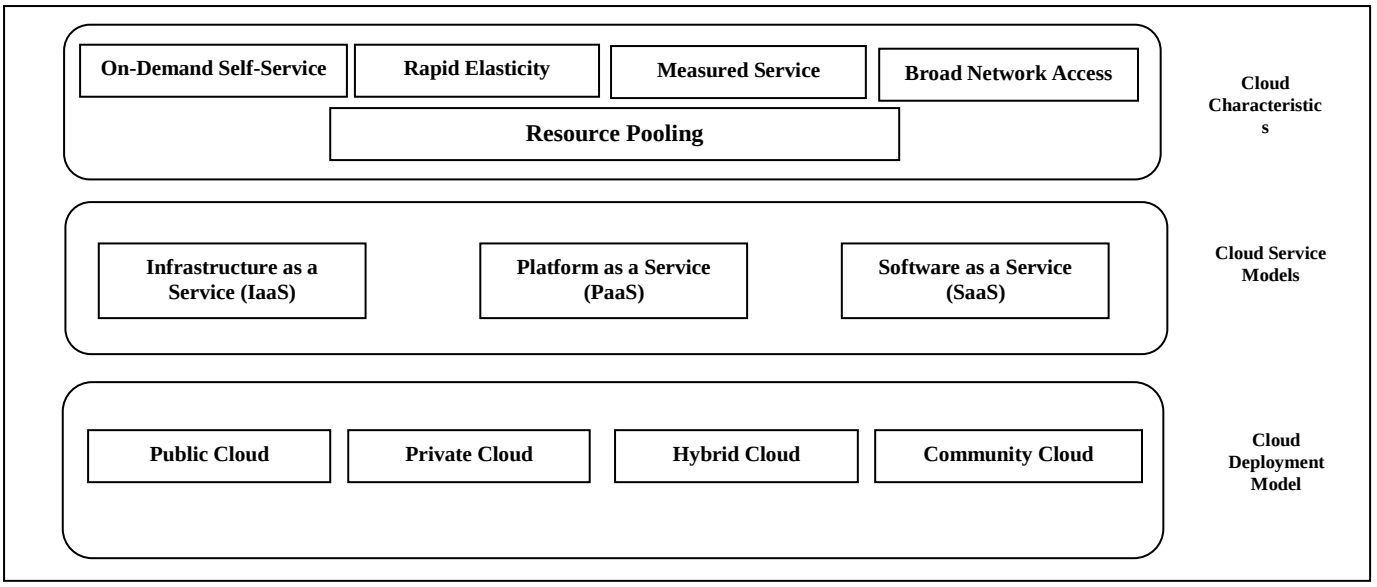

Fig. 1: Working Model of Cloud Computing

2.1.2 Service Models

IaaS- The fundamental hardware required to run the cloud application is provisioned by this service model. These resources include storage, network, processing unit, virtual machines etc. The cost of setting up and maintenance of these resources are very high and is saved by this service layer as the overall cost of setting up as well as maintenance is taken care by service provider. [3, 4]

PaaS- This layer provides the platform for the developers to develop the applications using the underlying cloud infrastructure. PaaS provides different tools, technologies, programming languages etc required in order to develop the cloud application. The end user doesn't have control over the underlying cloud infrastructure but has complete command over the application.

SaaS- This model allows the end user to use the cloud deployed applications over internet through wide range of available clients. End users have no command over the cloud infrastructure as well as the application itself except consuming the application.

2.1.3 Deployment Model

Public Cloud- Public cloud is made available to all the end users who just need to use the applications that is deployed over public cloud. Example includes Amazon EC2, Google App Engine, Microsoft Azure etc. [3, 5]

Private Cloud- Private cloud is totally dedicated to the individual private organizations for carrying out their business with high level of privacy as well as security and without the intervention of the outsiders. Examples include Microsoft ECI data center, Amazon Virtual Private Cloud, Ubuntu Enterprise Cloud, Eucalyptus etc.

Community Cloud- This type of deployment model is used when the organizations share the cloud infrastructure altogether. Based on the requirement community cloud model can be outsourced or on-site basis. Example include- Microsoft Government Community Cloud, Google Apps for Government.

Hybrid Cloud- Cloud structure that is based on the combination of other available deployment models is termed as hybrid cloud. Example may include VMware vCloud etc.

2.1.4 Cloud Attacks

Cloud computing paradigm suffers from several kinds of attacks. These attacks depending on their types may occur at different cloud service models, i.e. at IaaS, PaaS or at SaaS. [6]. Figure 2 displays some of the popular cloud attacks at the respective service models.

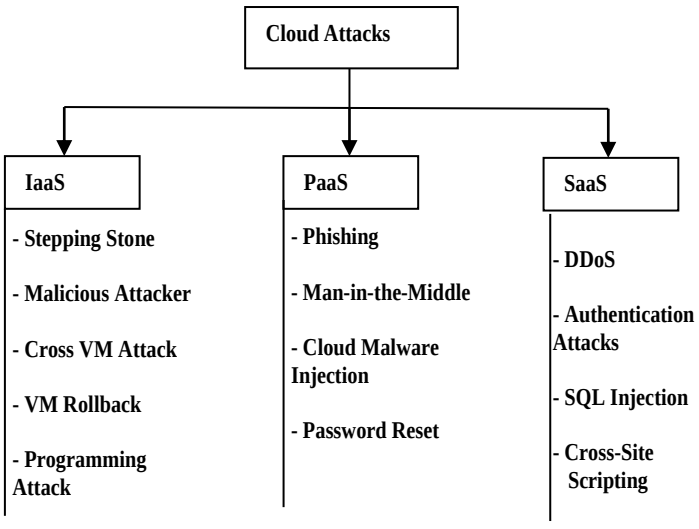

Fig. 2: Classification for Cloud Attacks

2.2 Machine Learning

According to Arthur Samuel, learning through past experience instead of learning through programming is termed as machine learning. Machine learning makes use of various types of algorithms to create models which when trained on large volume of dataset can predict the future outcome from the learning of the past historical data. The algorithms used to train the models are the backbone of machine learning. The choice of machine learning algorithm depends on the type of problem to be solved. The process of applying machine learning in order to solve a given problem starts with data collection and then follows the task of data preparation, data analysis, training the model, testing the model and finally deploying the model for actual use. [7, 8]

2.2.1 Types of Machine Learning

Supervised Machine Learning- Supervised ML algorithms are used to predict the future outcomes as they are trained on the datasets that are labeled and are mapped with corresponding output target values. The major task of supervised ML algorithm is to observe the given input data and allot an appropriate class for that data. This allotment of the class can only be achieved by getting trained beforehand using large volume properly labeled dataset with clear classes defined.

Supervised ML algorithm can solve two category of ML problem, namely Classification and Regression. The problem which has categorical (yes/no) target variable are solved using Classification Supervised ML algorithms where as when the target variable is not categorical but is continuous instead, such type of problems are solved using Regression ML algorithms. [9, 10]

Unsupervised Machine Learning- Unsupervised ML algorithms train the ML model with the datasets that is not labeled as well as not categorized. By examining the large dataset, unsupervised ML algorithm determines and learns all the data insights such as data patterns, classes, categories etc on its own. Clustering and Association are the two categories of unsupervised ML.

Clustering based algorithm form the groups of similar data which has similar characteristics. Whereas Association based algorithms finds the relation between the data that can be grouped together. [11, 2]

Semi-Supervised Machine Learning- The shortcoming of supervised and unsupervised ML algorithm is addressed by semi-supervised ML algorithms. Both labeled as well as unlabelled datasets are used to train the ML model in semi-supervised based learning. [9]

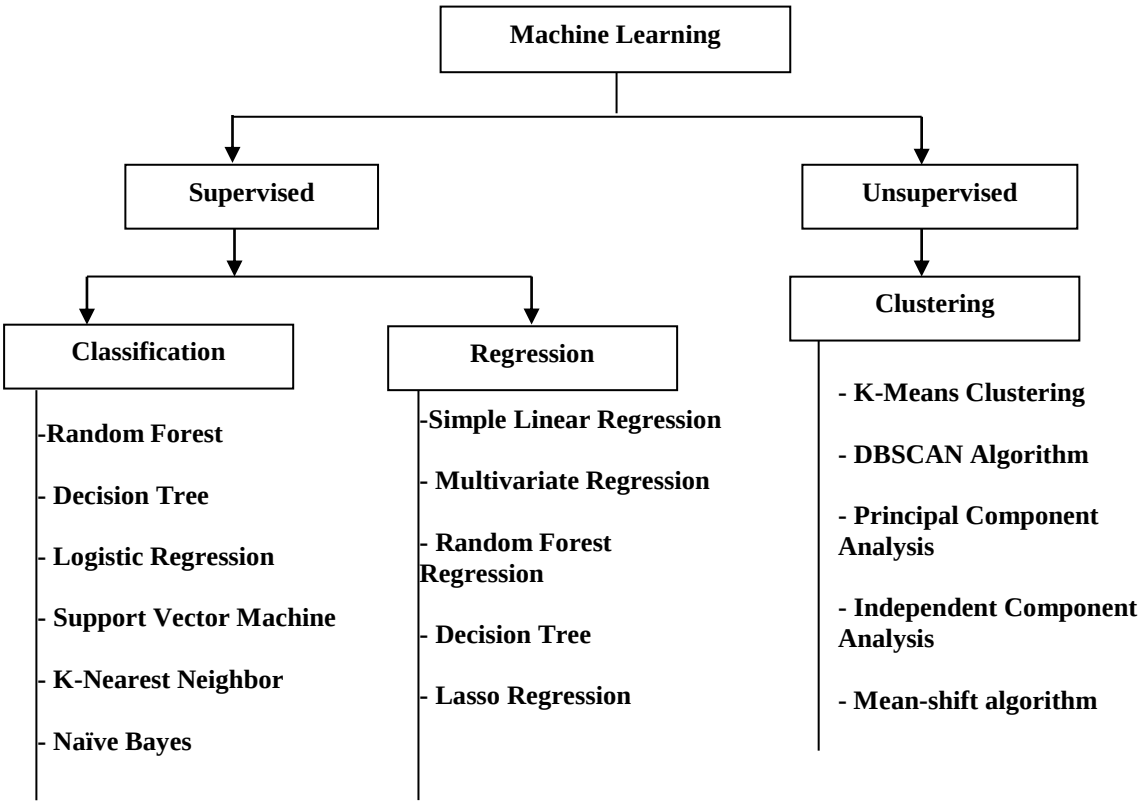

Fig. 3: Classification of Machine Learning Algorithms

Reinforcement Machine Learning- Feedback based learning methodology is utilized in Reinforcement ML. The agent learns from its own experience has no training over any type of supervised datasets and is rewarded or penalized for making the correct or incorrect decisions accordingly. Figure 2 shows the classification of machine learning algorithms.

3 LITERATURE REVIEW

Chkirbene et al. [13] proposed machine learning based intrusion detection system for cloud computing. The classifier in an intrusion detection system is most important component which fails to give high classification accuracy results due to the imbalance nature of the datasets available. In order to cater this problem weighted supervised decision tree algorithm is employed as classification algorithm in this proposed approach. High accuracy for the classifier is achieved as the proposed approach produces low scores for negative classification and high scores for positive classification.

Another security framework is proposed in the research work of *Bagga et al.* [14] based on the combination of SVM machine learning algorithm, Network Function Virtualization and Software Defined Network. This approach marks its importance as the security against different attacks for both NFV as well as for SDN in achieved. The proposed framework is divided into two levels. Firstly into "security enforcement plane" which is responsible for providing the security against both the internal as well as external attacks in IoT and is further sub-divided into three components namely MA (Monitoring Agent), IB (Infrastructure Block), CMB (Control and Management Block). And secondly into "security orchestration plane" for configuring the security policies at the run time. Better result is achieved in terms of accuracy, FRP, detection rate and training time as compared with other existing approaches.

The importance of data security in mobile cloud computing due to the involvement of heterogeneous network is depicted by *Dey et al.* [15] and an intrusion detection system that can handle such complex security constraints is thus proposed. K-Means and DBSCAN machine learning algorithm lays the foundation for such an IDS, which can guard defence against heterogeneous attacks such as MITM as well as DDoS. This approach trains the system on cluster basis and does the traffic classification on the basis of distance calculation. Better accuracy results for the proposed IDS is achieved as their is a reduction in the complexity due to the non requirement of updates in the rules regularly.

Concept for secure offloading using machine learning in multi-environment (Fog-Cloud-IoT) is given by *Alli et al.* [16]. Optimal selection of the fog node is done by PSO (Particle Swarm Optimization) which can be used as IoT data storage and then transfer of the data is done to the cloud which is selected via reinforcement learning. Private cloud is used for storing the sensitive data whereas the non-sensitive data not uploaded in the private cloud.

Another machine learning based scheme for monitoring the behavior of the user in the cloud for the CSP (cloud service provider) is given by the *Rabbani et al.* [17]. For the purpose of identification of unauthenticated user in the cloud, the approach utilizes the hybridization of PSO-PNN (particle swarm optimization and probabilistic neural network). Results showed the effectiveness of the proposed hybrid scheme by achieving high accuracy in terms of true positive rate, false negative rate, f-measure and precision.

Hesamifard et al. [18] utilizes machine learning capability for preserving the privacy. Data encrypted with homomorphic encryption is used for the purpose of training the neural network. The traditional sigmoid as well as ReLU (Rectified Linear Unit) activation functions of the neural network is substituted with the accurate polynomial approximations as an activation function of NN. The proposed approach produces more accurate in providing privacy in comparison with SMC (secure multiparty computation) and HE (homomorphic encryption).

Secure machine learning based sharing of data over cloud is achieved by *Singh et al.* [19] via mutual authentication protocol. The proposed mutual authentication protocol easily guards defence several types of cloud attacks such as DoS, DDoS, MITM, reply etc. ECC (Elliptic curve cryptography) as well as Schnorr's signature are used in combination for the purpose of encrypting the data with the benefit of small size keys and classification of threats or attacks are performed by voting classifier. The high accuracy of the proposed methodology is proved by the results from the ProVerif tool.

Salman et al. [20] gave a research paper suggesting the use of machine learning in order to mitigate different cloud attacks in multi-cloud environment via intrusion detection system. Linear regression and random forest supervised machine learning algorithms are employed by the intrusion detection system used in this proposed approach. Apart from the detection of the cloud threats, the main advantage of this approach is that it also makes sure to categorize the threats via a novel step-wise algorithm. 99.0% and 93.6% accuracy is achieved in terms of categorization as well as detection of the threats respectively.

With the hybridization of genetic and simulated annealing algorithms *Chiba et al.* [21] proposed an intrusion detection system based on deep neural network. The improved genetic algorithm used by this approach provides reduction in the convergence as well as in the execution time at the same time the optimization in the search process of genetic algorithm is achieved by the SAA algorithm. These algorithms improve factors of DNN including feature selection, activation function and thus enhancing the overall performance of the deep neural network.

Machine learning based authorization to allow only the authenticated user access the cloud services is proposed by *Khilar et al.* [22]. As the proposed approach improves the authorization mechanism of the cloud users and restricts the unauthorized access of the cloud resources, the trust between the service providers and the end users improves and also the overall data security reaches another level. The proposed approach gave better results in terms of MAE, time, recall, precision and f1-sccore when compared with traditional mechanism for user access to cloud resources.

Machine learning based IDS with improved accuracy is proposed in the work of *Aljamal et al.* [23]. SVM for classification along with K-Means clustering machine

learning algorithm is used in hybrid mode at the cloud hypervisor in order to detect the anomalies in the network. The proposed hybrid model performs network traffic investigation, unwanted feature reduction from dataset, clustering the data with K-Means algorithm and classification between normal as well as malicious requests via SVM.

Table 1. Summary of related work

Ref	ML Algorithm Used	Proposed Approach	Dataset
[13]	Decision Tree	Intrusion Detecting System based on weight optimization.	UNSW
[14]	Support Vector Machine	AI Framework based on the combination of ML, NFV, SDN.	NSL-KDD
[15]	K-Means and DBSCAN	Traffic filtration via distance calculation and training system via cluster basis.	Multiple datasets
[16]	Reinforcement Learning	Neuro-Fuzzy system for secure data offloading with PSO to select secure fog node in Fog-Cloud-IoT environment.	Multiple datasets
[17]	Multilayer Neural Network	Identification of unwanted user in the cloud with PSO and PNN.	UNSW-NB15
[18]	Deep Neural Network	Training NN with encrypted data and using accurate activation function for the NN.	Crab, Fertility and Climate Dataset
[19]	LR(Linear Regression) and KNN (K-Nearest neighbor)	ECC along with voting classifier for mutual authentication over multi cloud environment.	CICD
[20]	Linear Regression (LR) and Random Forest (RF)	Machine Learning based Intrusion detection system for detection of attacks in multi-cloud environment.	UNSW
[23]	K-Means clustering and SVM classification	The hybrid model is responsible for performing network traffic investigation, unwanted feature reduction from dataset, clustering the data with K-Means algorithm and classification between normal as well as malicious requests via SVM.	UNSW-NB15
[24]	Random Forest , Quadratic Discriminant Analysis, K-Nearest Neighbours, Gaussian Naive Bayes (GNB) and AdaBoost	Deep Reinforcement Learning model which has the host, agent and administrator network that predicts the affected virtual machines and also blocks them.	UNSW-NB15
[21]	Deep Neural Network	IGASAA, i.e. hybridization of genetic algorithm and simulated annealing algorithm for machine learning based network IDS.	CICIDS2017, NSL-KDD version 2015 and CIDDS-001
[25]	Linear Regression (LR)	EIDS for traffic analysis in which the past as well as current decisions are compared with each other using machine learning.	UNSW-NB-15
[26]	Decision Tree, Random Forest	Machine learning based classification TIDCS and detection TIDCS-A models for IDS.	NSL-KDD, UNSW
[22]	K-Nearest Neighbor, Decision Tree, Logistic Regression, Naive Bays	Authorization of the user is increased to provide better security to the cloud resources using machine learning approach.	User Dataset
[27]	DML	DML-DIV for retraining the integrity of the data in distributed cloud environment.	Advertisement Click Prediction

Security for cloud is enhanced by IDS based on reinforcement learning in the research paper of *Sethi et al.* [24]. The major drawback of the traditional IDS for cloud security is the incorrect classification accuracy, i.e. low FPR (false positive rate) is taken care by this approach. The proposed model involves three modules, the host network responsible for mitigating virtual machine based attacks, agent network responsible for detection between normal or malicious requests and the administration network for allowing the administrators to block the affected virtual machines.

Chkirbene et al. [25] suggested an "EIDS" scheme for traffic analysis in which the past as well as current decisions are compared with each other using machine learning based intrusion detection in order to provide cloud security. The current and the past decision comparison for the classification of attacks are performed in order to enhance the performance of the intrusion detection system. The security of IDS is increased as there is overall 24% increase (near about 90%), in the detection rate of the supervised learning classifier. *Chkirbene et al.* [26] gave two models for trust based IDS first classification model (TIDCS) and second one the accelerated model (TIDCS-A). The former model is responsible for the task of dimensionality reduction allowing only the required features to be processed by the machine learning algorithm from the UNSW dataset while the latter model takes care of

detection of the anomalies. Simulation results clearly depicts that both the proposed model with the help of machine learning algorithms, (TIDCS and TIDCS-A) are capable for attack classification as well detection with better accuracy.

When it comes to machine learning in the distributed cloud environment, the problem of data tampering increases. In order to solve this problem, *Zhao et al.* [27] proposed a verification methodology named as DML-DIV (distributed machine learning data integrity verification), for the data in distributed cloud environment so that the integrity of the data can be retrained. Also the simulation results clearly depicts that the proposed DML-DIV approach is better than the existing compared approaches in terms of privacy protection, forgery as well as tampering attack.

4 CONCLUSIONS AND FUTURE SCOPE

Client data over the cloud is very crucial and its security can't be compromised by any means. Several new technologies making use various security algorithms are applied by the researchers in order to enhance the security of the cloud ecosystem. Machine learning finds a huge space to provide more accurate as well as automate defence against the known and unknown cloud attacks. The main focus or the takeaway from this survey paper is to have a latest glimpse of the research work in the field of cloud security using machine learning.

In future, we propose an intrusion detection system that will make use of enhanced and optimized machine learning algorithm in order to provide more accurate cloud data security.

REFERENCES

[1] Alouffi, B., Hasnain, M., Alharbi, A., Alosaimi, W.: A Systematic Literature Review on Cloud Computing Security: Threats and Mitigation Strategies. IEEE Access, 9, pp. 57792-57807, 2021.

[2] Abdulsalam, Y.S., Hedabou, M.: Security and Privacy in Cloud Computing: Technical Review. Future Internet 2022, 14, 11.

[3] George, S.S., Pramila, R.S.: A review of different techniques in cloud computing. Materialstoday proceedings, 46, pp. 8002-8008, 2021.

[4] Attaran, M., Woods, J.: Cloud computing technology: improving small business performance using the Internet. Journal of Small Business & Entrepreneurship. 13. pp. 94-106, 2018.

[5] Basu, S., Bardhan, A., Gupta, K., Saha, P., Pal, M., Bose, M., Basu, K., Chaudhury, S., Sarkar, P.: Cloud computing security challenges & solutions-A survey. Annual Computing and Communication Workshop and Conference (CCWC), 2018.

[6] Dwivedi, R.K., Saran, M., Kumar, R.: A Survey on Security over Sensor-Cloud. In: 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, pp. 31-37, 2019.

[7] Butt, U.A.; Mehmood, M.; Shah, S.B.H.; Amin, R.; Shaukat, M.W.; Raza, S.M.; Suh, D.Y.; Piran, M.J. A Review of Machine Learning Algorithms for Cloud Computing Security. Electronics 2020, 9, 1379.

[8] Sarker, I.H.; Machine Learning: Algorithms, Real-World Applications and Research Directions. SN computer science volume. 2, 3 (2021): 160.

[9] Alzubi, J., Nayyar, A., Kumar, A.: Machine Learning from Theory to Algorithms: An Overview. Journal of Physics: Conference Series, Volume 1142, Second National Conference on Computational Intelligence 2018, Bangalore, India.

[10] Baraneetharan, E.: Role of Machine Learning Algorithms Intrusion Detection in WSNs: A Survey. Journal of Information Technology and Digital World, Vol. 02, pp. 161-173, 2020.

[11] Saranyaa, T., Sridevi, S., Deisy, C., Chung, T.D., Khan, M.K.A.: Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review. Procedia Computer Science, Vol 171, pp. 1251-1260, 2020.

[12] Sen, P.C., Hajra, M., Ghosh, M.: Supervised Classification Algorithms in Machine Learning: A Survey and Review. Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, vol 937, pp. 99-111, 2019.

[13] Chkirbene, Z., Erbad, A., Hamila, R., Gouissem, A., Mohamed, A., Hamdi, M.: Machine Learning Based Cloud Computing Anomalies Detection. IEEE Network, Vol. 34, pp. 178-183, 2020.

[14] Bagaa, M., Taleb, T., Bernabe, J.B., Skarmeta, A.: A Machine Learning Security Framework for Iot Systems. IEEE Access, Vol. 8, pp. 114066-114077, 2020.

[15] Dey, S., Ye, Q., Sampalli, S.: A machine learning based intrusion detection scheme for data fusion in mobile clouds involving heterogeneous client networks. Information Fusion, vol. 49, pp. 205-215, 2019.

[16] Alli, A.A., Alam, M.M.: SecOFF-FCIoT: Machine learning based secure offloading in Fog-Cloud of things for smart city applications. Internet of Things, Vol. 7, 2019.

[17] Rabbani, M., Wang, Y.L., Khoshkangini, R., Jelodar, H., Zhao, R., Hu, P.: A hybrid machine learning approach for malicious behaviour detection and recognition in cloud computing. Journal of Network and Computer Applications, Vol. 151, 2020.

[18] Hesamifard, E., Takabi, H., Ghasemi, M., Jones, C.: Privacy-preserving Machine Learning in Cloud. Cloud Computing Security Workshop, pp. 39-43, 2017.

[19] Singh A.K., Saxena, D.: A Cryptography and Machine Learning Based Authentication for Secure Data-Sharing in Federated Cloud Services Environment. Journal of Applied Security Research, 2021.

[20] Salman, T., Bhamare, D., Erbad, A., Jain, R., Samaka, M.: Machine Learning for Anomaly Detection and Categorization in Multi-Cloud Environments. IEEE 4th International Conference on Cyber Security and Cloud Computing, 2017.

[21] Chiba, Z., Abghour, N., Moussaid, K., Elomri, A., Rida, M.: Intelligent approach to build a Deep Neural Network based IDS for cloud environment using combination of machine learning algorithms. Computers & Security, Vol. 86, pp. 291-317, 2019.

[22] Khilar, P.M., Chaudhari, V., Swain, R.R.: Trust-Based Access Control in Cloud Computing Using Machine Learning. Cloud Computing for Geospatial Big Data Analytics, pp. 55-79, 2018.

[23] Aljamal, I., Tekeoğlu, A., Bekiroglu, K., Sengupta, S.: Hybrid Intrusion Detection System Using Machine Learning Techniques in Cloud Computing Environments. IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA), 2019.

[24] Sethi, K., Kumar, R., Prajapati, N., Bera, P.: Deep Reinforcement Learning based Intrusion Detection System for Cloud Infrastructure. International Conference on Communication Systems & Networks (COMSNETS), 2020.

[25] Chkirbene, Z., Erbad, A., Hamila, R.: A Combined Decision for Secure Cloud Computing Based on Machine Learning and Past Information. IEEE Wireless Communications and Networking Conference (WCNC), 2019.

[26] Chkirbene, Z., Erbad, A., Hamila, R., Mohamed, A., Guizani, M., Hamdi, M.: TIDCS: A Dynamic Intrusion Detection and Classification System Based Feature Selection. IEEE Access, vol. 8, pp. 95864-95877, 2020.

[27] Zhao, X., Jiang, R.: Distributed Machine Learning Oriented Data Integrity Verification Scheme in Cloud Computing Environment. IEEE Access, Vol. 8, pp. 26372-26384, 2020.