Machine Learning Defence Mechanism for Securing the Cloud Environment

Girish L[1*] and Raviprakash M L[2*]

[1*] Shridevi Institute of Engineering Technology, Karnataka, India.
[2*] Kalpataru Institute of Technology, Karnataka, India.

*Corresponding author(s). E-mail(s): girishltumkur@gmail.com; raviprakashml@gmail.com;

Abstract

A computer paradigm known as ”cloud computing” offers end users on-demand, scalable, and measurable services. Today’s businesses rely heavily on computer technology for a variety of reasons, including cost savings, infrastructure, development platforms, data processing, data analytics, etc. The end users can access the cloud service providers’ (CSP) services from any location at any time using a web application. The protection of the cloud infrastructure is of the highest significance, and several studies using a variety of technologies have been conducted to develop more effective defenses against cloud threats. In recent years, machine learning technology has shown to be more effective in securing the cloud environment. In recent years, machine learningtechnology has shown to be more effective in securing the cloud environment. To create models that can automate the process of identifyingcloud threats with better accuracy than any other technology, machine learning algorithms are trained on a variety of real-world datasets. In this study, various recent research publications that used machine learning as a defense mechanism against cloud threats are reviewed.

Keywords: Cloud Computing, Cloud Security, Machine Learning, Cloud Attacks

1 Introduction

Cloud computing is the term that is normally used to describe many different computing concepts which are involved in the more number of computers which are providing computer power through the Internet. The above definition which is referred to software that is hosted in remote locations with providing clients with network provisioning. Cloud computing allows the realization of a favorable computing model that includes the desire for quite powerful and widely available resources.

The cloud computing systems are classified by NIST [16] in four deployment models which are: private cloud, hybrid cloud, public cloud, and community cloud. Additionally, three service models can be provided via cloud computing: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a service(IaaS) are the three services which are provided by the cloud computing. In the terms of layers, the administrative duties involved with the system's cloud deployment and delivery models, as well as their interaction.

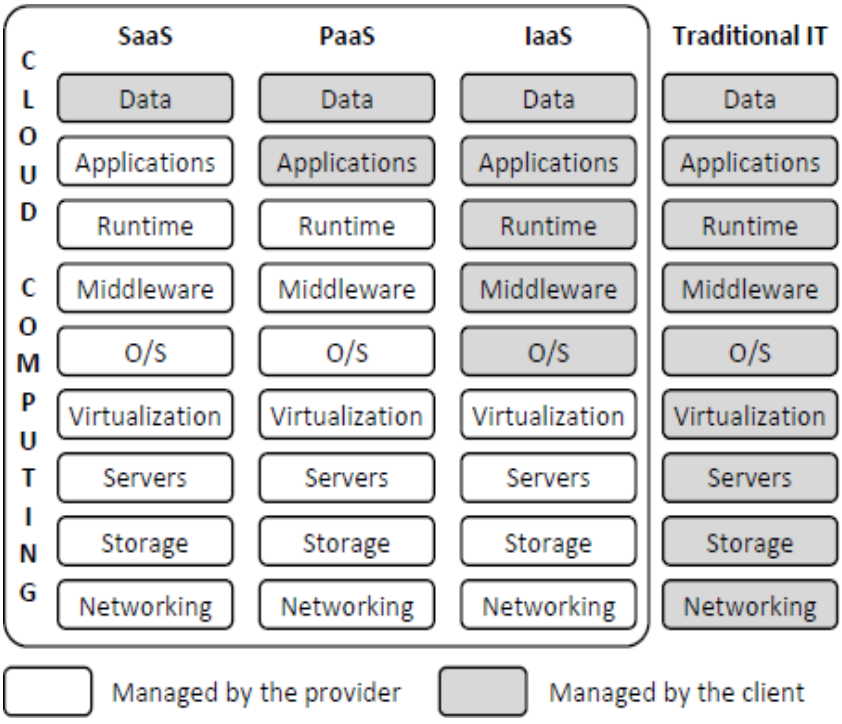

Fig. 1 Architecture of Cloud computing and Types of Services

There are basically four different cloud deployment methods, and each of them comes with unique trade-offs for companies moving their services and

operations to cloud-based environments. The models for cloud deployment are shown in Figure 2. Here are several examples:

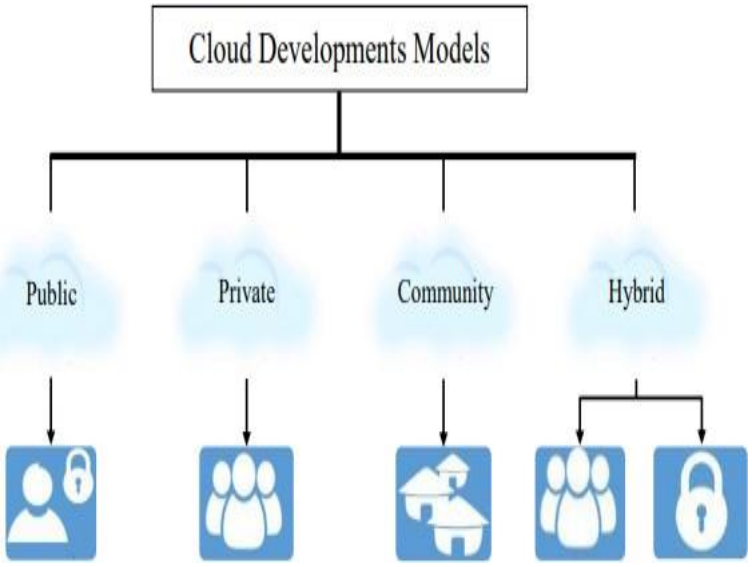

Fig. 2 Cloud Deployment Models.

- Private cloud – In this approach to cloud deployment, a cloud infrastructure is managed solely for one company. However, it may be handled by by a third party or by the organization as such

- Public cloud – The organization that owns the cloud infrastructure is one that offers cloud services, but it is also accessible to the general public or a major industry group.

- Community cloud – A particular community with common problems and objectives is supported by the cloud infrastructure, which is shared by a variety of organizations. (e.g., e-commerce, business, security, compliance considerations and requirements).

- Hybrid cloud – It is distinguished by a cloud infrastructure made up of two or more clouds with various model types. (private, community or public).

Delivery models are the models whereby various services are provided to the customer [17] [18]. There are already three commonly used and recognized structured cloud delivery models namely Platform as a Service, Infrastructure as a Service and Software as a Service. They are described as follows:

- SaaS – includes programmes that may be accessed using a thin client interface, such as a web browser, on a variety of client devices. (such as web email) One of the biggest benefits for these clients is the potential to reduce IT support costs by outsourcing hardware, software maintenance, and sup- port. In short, It is a model of software deployment in which services are offered to users on demand over the Internet while being remotely hosted by the service provider.
- PaaS – comprehends a platform layer made up of operating systems and application frameworks that help clients reduce the workload associated with installing new apps directly on virtual machines. Programming languages and tools supplied by the providers may be used to create applications. In this delivery model, the network, servers, operating systems, and storage of the underlying cloud infrastructure are not under the user's control; instead,the user has control over the deployed programmes and configurations of the application hosting environment.
- IaaS – includes both infrastructure layer and the hardware layer, as well as basic computer resources like processing, storage, networks, and other components that allow the user to deploy and run any softwares, such as operating systems and applications. The client may often be in control of handling some of the physical resources. The client may often be in control of handling some of the physical resources. (e.g., servers, routers, switches). By using virtualization techniques, the virtualization layer provides instances of the physical infrastructure to various clients.

In this paper, The Machine learning (ML) algorithms are used to discuss security concerns, difficulties, and related solutions in cloud computing. Using machine learning techniques, security problems can be resolved and data can be managed more effectively [23]. ML is the application of artificial intelligence that enables frameworks to naturally take in and improve without being significantly modified [24]. The evolution of computer programs that can utilize it to learn for themselves is the main emphasis of machine learning. The method for learning begins with perceptions or data, such as heading, direct understanding or models, to channel for informational structures and choose better options later on in relation to the models that are provided. The objective of this work is to give a research of the security risks and ethical issues related to distributed computing using machine learning techniques.

The remaining parts of the paper is arranged as follows. The background research of CC models, threats, and attacks, as well as an overview of ML algorithms, are presented in section 2. The associated survey articles are discussed in section 3 together with the review paper. Along with machine learning techniques for cloud computing security, the objectives, methods, advantages, and drawbacks of many papers are presented in section 4. Section 5 describes the concluison of the work.

2 Cloud Network Security

The ability to monitor a cloud network enables a cloud provider to analyse a variety of network traffic characteristics, including throughput, response time, jitter, lost data, etc. The majority of monitoring solutions on the market nowadays offer a graphical view of network information from which issues may be found and fixed [19].

These statistics can also be used by cloud administrators to carry out normal procedures that could or might not reveal anomalies in the traffic. These abnormalities may eventually be identified as an issue, such as a server down or a server receiving an unusually high volume of queries.

Monitoring cloud traffic makes it possible to continuously maintain track on network connectivity and application availability, making it easier to spot issues with hosts, networks, or servers. After gathering all of this data, it is possible to spot a number of patterns in the network traffic that define these abnormalities. We can develop strategies to avoid future occurrences of similar issues [20]. These statistics can be used to draw conclusions about how future network traffic will behave as they accumulate over time. As a result, the network administrator have enough time to take the action when an anomaly arises before it becomes worse. It is well known that there are distinctions between network traffic that behaves normally and that which results from an attack or anomaly. The line between these two extremes is blurry, so we are unclear when network traffic stops being a normal usage and starts to represent an attack or abnormal.

The Cloud Network Attack (CNA) consists of activities typically carried out by software created with malicious purposes. The effects are fairly diverse;some of them are intended to infect or penetrate another person's computer in order to eventually delete the data and destroy the hardware or software, altering the device functions, or even leaving the computer open to other formsof attacks.

2.1 Identifying the Cloud Attacks

A network anomaly is the situation in which the behaviour of the network deviates from its typical operating pattern [21]. The nature of typical traffic behaviour must be understood by the network administrator before they can identify an abnormality. This is a crucial aspect to make the system more robust and safeguard network traffic against threats: in order to assure a sufficient level of security, one needs a thorough understanding of the computer network. There is no quick and simple way to determine whether the network is transporting legitimate, typical traffic. This typically calls for intensive net- work monitoring in order to derive the characteristics of the network's typical behaviour.

2.2 Different Types of Attacks in Cloud

The network data may be attacked without the proper security measures and a strong control mechanism. The attacks that have already occurred may be divided into two classifications: passive and active.

During a passive attack, the data is just visualised without being changed. Traffic analysis, monitoring of communications that are not protected, decryption of traffic that is weakly encrypted, and the capture of authentication data like passwords are all examples of passive attacks.

Instead, information is altered with the intention of destroying or causing damage to the data in an active attack. Attempts to cheat or circumventsecurity measures, introduce malicious code, and steal or alter sensitive data are all examples of active attacks.

The most frequent attacks in cloud computing over networks or iso- lated virtual machines have been the subject of numerous articles in the literature.[22].

- **Denial of Service:** A DoS attack aims to prevent legitimate users of a service from using that service instead of stealing information. The attack's objective is to make the World Wide Web's(WWW) hosted sites or services unavailable and typically targets sites or services hosted on the Internet.
- **Brute Force Atatck:** consists of an attacker using a terminal to attempt to guess a password, frequently with the assistance of scripts and dictionaries that run automatically.
- **Sniffer Attack:** This attack is carried out by software or hardware that may intercept and record data flowing over a network. Its original goal was to assist in controlling and locating network issues. Hackers employ snif-fers to gather private data from networks, including passwords and account information.
- **Remote to Local Attack:** The attacker in this attack doesn't have a user account on the target computer. He then attempts to take advantage of security flaws in order to access the remote machine across the network as a local user.
- **Flash Crowd Attack:** Despite the fact that this isn't really an attack, the abnormalities caused by flash crowds are quite similar to DDoS attacks. Flash crowds are significant surges in real traffic that are focused on particular websites or newly released software that is made available online for a brief period of time.
- **User to Root Attack:** Due of the attacker's local access to the victim PC, this attack is unique. He yet tries to illegally obtain super user rights.
- **Probe Attack:** In a Probe attack, the attacker seeks to learn more about the target host. In order to make it easier to find vulnerabilities, an attacker can try to obtain relevant information about the computers and services that are accessible on the network.
- **Distributed Denial of Service:** DDoS and DoS are similar. The primary difference between DoS and DDoS is that DDoS is carried out simultaneously

by multiple locations while DoS is carried out by a single machine. Similar to a DoS assault, a DDoS attack aims to temporarily disable a system or network resource. Typically, DDoS attacks are more successful than DoS attacks because they may employ thousands of computers to target a single system. It often happens when a lot of Internet packets from infected hosts (zombies) overrun a single target's bandwidth or resources (victim).

These kind of assaults are particularly dangerous with cloud computing and may have a severe impact on consumers in two different ways. Let's start by visualising an attacker who has accessed a cloud infrastructure without authorization. Once hundreds of virtual computers can send this harmful operation at once, he will be able to launch an attack of significant size. The second need is that the service provider ensure effective management and traffic control.

Serious repercussions may result if the provider fails to complete this task. When handling a legitimately large request for resources, the provider could mistake when dealing with an irregular traffic control. The provider traffic control system can associate this with a deliberate intent to do harm. Asa result, many genuine users would experience resource blocking and a DoS attack's consequences.

3 Related Work

This section describes the various works done in the literature to under- stand the gap analysis. We briefly mentioned various related work details and comparision of the works has been presented in the Table1.

Munish saran, [1] Proposed that cloud providers like Amazon, Azure, Google cloud providers are facing issues of cloud attacks even if they are hav- ing enough cloud security. In this regard they are using Machine Learningas a Service(MLaaS) service models to make the strategies to minimize the cloud attacks. Using Machine Learning algorithms, intrusion detection systems have been developed to increase the accuracy of detecting attacks. Weighted supervised decision tree classifier is used to achieve high accuracy in intrusion detection by Chkirbene et al. Bagga et al. proposed a research work with Combination of Support Vector Machines, Network function virtualization and software defined Networks as a security framework to manage security against different attacks. Dey et al. illustrate the significance of data security in mobile cloud computing due to the inclusion of heterogeneous networks.

Deval Bhamare, [2] - With the development of machine learning techniques, learning-based methods for security applications are currently becoming more and more prominent in the literature. The main difficulty with these strategies is, however, locating current and objective datasets. Due to privacy concerns or a potential lack of specific statistical qualities, many datasets are internal and cannot be shared. There are few studies that concentrate on buildingand testing machine learning models over various datasets. In fact, doing sois essential to evaluating the robustness and usefulness of machine learning algorithms in practical situations. In this study, the performance of important

supervised machine learning algorithms using two separate datasets, namely UNSW and ISOT, in order to address this specific problem in current studies is examined. First, using the UNSW training dataset, the models are trained. The trained models are then put to the test using datasets from ISOT and UNSW. It should be emphasized that while the ISOT dataset was acquiredusing an entirely different experimental context than the UNSW training and test datasets.

Gopal Krishna Shyam, [3] - A rising range of dynamic threats and attacks, including as data breaches, insecure interfaces, account theft, shared technology vulnerabilities, advanced persistent threats, and distributed attacks, target the cloud in an attempt to disrupt cloud services and damage security. A variety of control-based technologies, including next-generation firewalls, cryptographic approaches, intrusion detection systems, software-defined networks, and machine learning techniques, among others, can be used to address various security challenges, threats, and attacks. These are used to provide cloud security. It is important to take the proper precautions to address the security issues. Using Machine Learning and non Machine Learning techniques, several cloud security issues, threats, attacks and suggested solutions are analyzed. Machine learning methods are crucial for identifying threats and attacks.

Jeffrey C Kimmell [4], Mahmoud Abdelsalam, Maanak Gupta - The IT workforce may now automatically supply resources in the cloud thanks to cloud automation solutions, which have become the norm. Such automation is made possible by technologies that allow for the creation, modification, and deletion of cloud resources using configuration scripts. In cloud infrastructure Malware is a crucial threat. Numerous malware detection techniques have been suggested, each with advantages and disadvantages. The signature of an executable is examined and contrasted with a database of known malware signatures in the common technique known as static malware detection. Attackers have used strategies like obfuscation and packing to try and reduce the effectiveness of static analysis. The activity of malware can be accurately and effectively captured using machine learning (ML) and neural network approaches. When compared to other Machine Learning models, the DenseNet-121 (CNN) model is best at detecting malware attacks online.

Ayesha Sarosh [5] - Cloud computing domain is replacing old technologies and conventional software models nowadays. By fulfilling the infrastructure demands, cloud computing is becoming standard day by day. The anomaly detection technique is immensely used for project identification, prediction & detection and behavioral analysis. Intrusion detection is an essential and reliable tool for building a secure environment in cloud computing. In Machine Learning, using Support vector machines and K-means clustering algorithms Hybrid model has been introduced to work on intrusion detection for virtualized infrastructure which is better than using individual models. With the help of the anomaly detection model, average detection time has been calculated and used as a best model for Intrusion detection. The dataset which is used is the UNSW-NB15 standard dataset that has rich labeled data.

The importance of data security in mobile cloud computing due to the involvement of heterogeneous network is depicted by Dey et al. [6] 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.

Salman et al. [7] 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. based on deep neural network proposed an intrusion detection. 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 [8].

Machine learning based authorization to allow only the authenticated user access the cloud services. 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 [9].

4 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

Table 1 Summary of The Related Work

Algorithm / Methodology	Detailed Approach	Dataset
Decision Tree	Intrusion Detecting System based on weight optimization.	UNSW
Support Vector Machine	AI Framework based on the combination of ML, NFV, SDN	NSL-KDD
K-Means and DBSCAN	Traffic filtration via distance calculation and training system via cluster basis.	Multiple datasets
Multilayer Neural Network	Identification of unwanted user in the cloud with PSO and PNN.	UNSW-NB15
K-Means clustering and SVM classification- 23	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
Decision Tree, Random Forest	Machine learning based classification TIDCS and detection TIDCS-A models for IDS.	NSL-KDD, UNSW

the task of data preparation, data analysis, training the model, testing the model and finally deploying the model for actual use [10].

4.1 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 [11].

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.

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 onits 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 [12].

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. Figure 3 shows the Classification of Machine Learning algorithms.

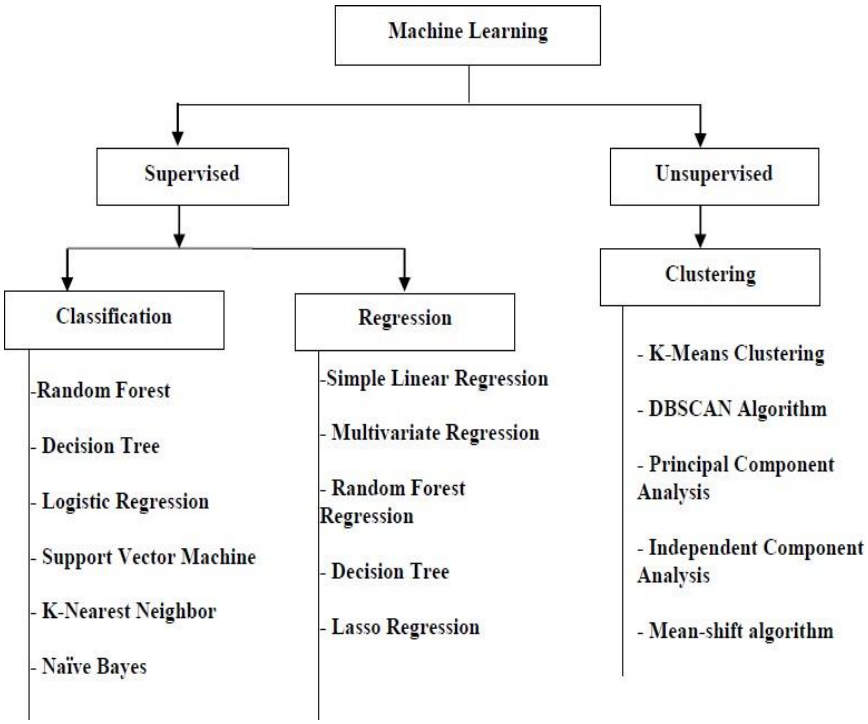

Fig. 3 Classification of Machine Learning Algorithms

4.2 Machine Learning Model for Attack Detection

Figure 4 depicts the two phases of a machine learning-based attack detection: the training phase and the attack detection phase. The machine learning-based detector first collects the various CPS data relevant to cyber-physical security on each CPS layer in both unattacked and attacked scenarios during the training phase. Safety-relevant data can be gathered and used to train the detector's machine learning model on the physical system layer. Examples of safety-critical data include vast quantities of sensor measurements from several physical systems, control input signals, and control period information. The application layer is where information about the computer system that could be harmful to system software and hardware can be obtained. This informationincludes how often a given command is executed, how much CPU and RAMare being used, and what files are being stored. Upon data collection, the machine learning-based detector classifies the data according to whether it wasproduced under normal or unusual situations.

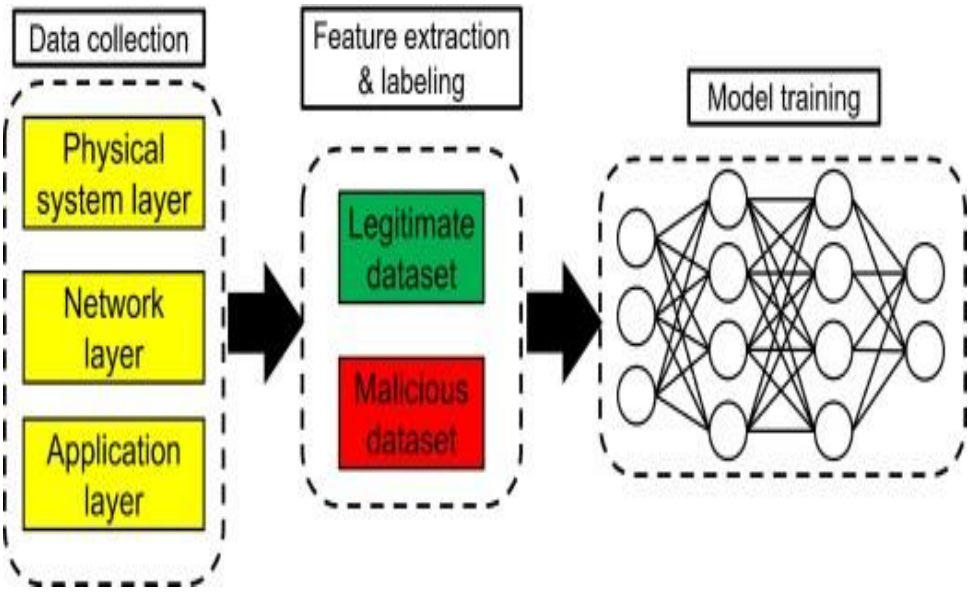

Fig. 4 Training the ML model for Attack Detection

4.3 Machine Learning and Deep Learning

Machine learning algorithms tend to fail to environmental changes and are not robust to noisy data, and low quality data [13]. Deep learning is robust to environmental changes by constant feedback. It structures the algorithms in the form of layers and creates artificial neural network which can learn from the data automatically and make intelligent decisions on its own. Each layerin the neural network applies its own transformation of the data and learns multiple level representations of the data. The features are learnt progressivelyfrom lower layers to higher layers from raw data. The entire process can be thought of finding the meaning of the data. Deep learning is used in many applications across industry.

4.3.1 Neural Network

Neural network is the imitation of human brain. It is made of hundreds or thousands of neurons, connected together much in the similar way as neurons are connected inside the human brain. The neurons in the NN work together to produce desired result [14]. Each neuron is responsible for solving only a part of the entire task. Neuron holds a number that lies between 0 and 1. This number is called the activation. Neural networks have the ability to learn and model many complex and non-linear data being generating in today's real life. There are many types of neural networks, such as, multilayer perceptron, Shallow neural network, long short term memory, recursive neural network, convolutional neural network, etc. [15].

The structure of the biological neuron and the mathematical neuron are shown in Figure 5 and Figure 6 respectively. The mathematical neuron is the imitation of the biological neuron. The biological neuron consists of a cell body, dendrites and an axon. Dendrites extend from the cell body a few hundred micrometres. Axon extrudes and carries impulses away from cell body. The tip of the axons is the synapses connected to another neuron. With the same analogy, today there is a mathematical neuron consisting of a cell body as shown inthe figure. The inputs coming to the cell body are the dendrites in the biological terms that are carrying the input values multiplied by the weights assigned. These product values are summed along with the bias. The summation value will be the value of the neuron. An activation function is applied on this summation value and the resulting value is sent on the output axon. Thus the cell body is activated with a specialized function called activation function.

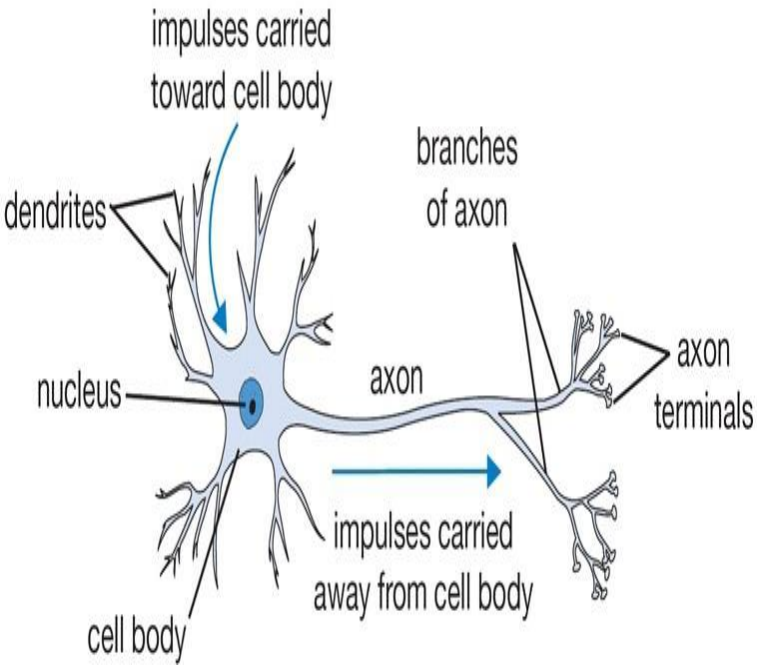

Fig. 5 Biological Neuron

The neural network is made up of three types of layers. The input layer, the output layer and the middle layers called the hidden layers. The input layer takes the input signals and transfers them to the hidden layer. A hidden layer can be a single layer or multiple layers. This layer does all the calculation and output of the feature. Input values are transferred to the hidden layer. Randomweights provides the interaction to the hidden layer from the input layer. Inputvalues from input to hidden layer are multiplied by weight. These are product

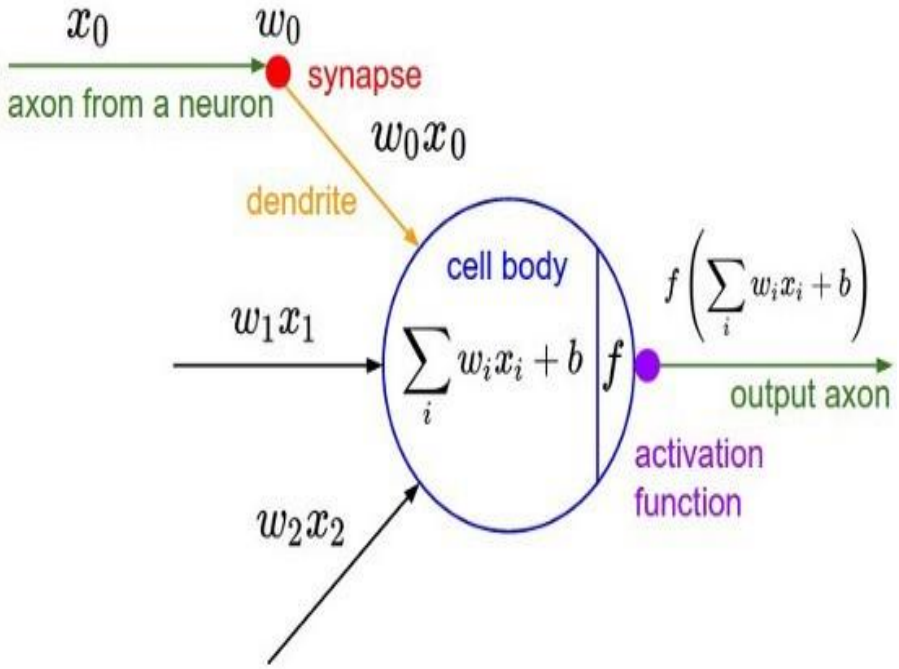

Fig. 6 Mathematical Neuron

summaries and selections. Summary of weighted inputs and bias is included as input to activation function in hidden layers. Interactions from the hidden layer to the extraction layer are also provided by random weights. The output values of the hidden layer are multiplied by the given weights. The result of the hidden end layer is transmitted as an input to the output layer. The output layer uses the activation function for this summary input and predicts the result. The predicted output of the network is matched with actual output. The difference between the both is the error. The error is now back propagated through the network and weights and biases at every interconnection are adjusted based on their contribution to the error. This is found using a cost function. In each iteration the output obtained is compared with the actual output; error is found and again back propagated to adjust the weights. The iterations are continued until maximum accuracy in the result is reached. The different typesof activation functions used in the hidden layers re explained in the following section.

4.3.2 Recurrent Neural Network

A Recurrent Neural Network or RNN contains recurrent layers. These layers process the inputs sequentially. RNNs are pretty flexible; they are capable of processing all kinds of sequential inputs. They can be used to process time series. Figure 7 shows a basic architecture of recurrent neural network. It

comprises of two recurrent layers and one dense layer to serve as the output. RNN takes the input sequence in batches and the output is also received in batches for the corresponding inputs. The input dataset is of three dimensional; having batch size as the first dimension, timestamps as the second and the dimensionality of input and every time step is taken as the third dimension. Example, if it is a univariate time series then the value will be one and the value will be more for multivariate.

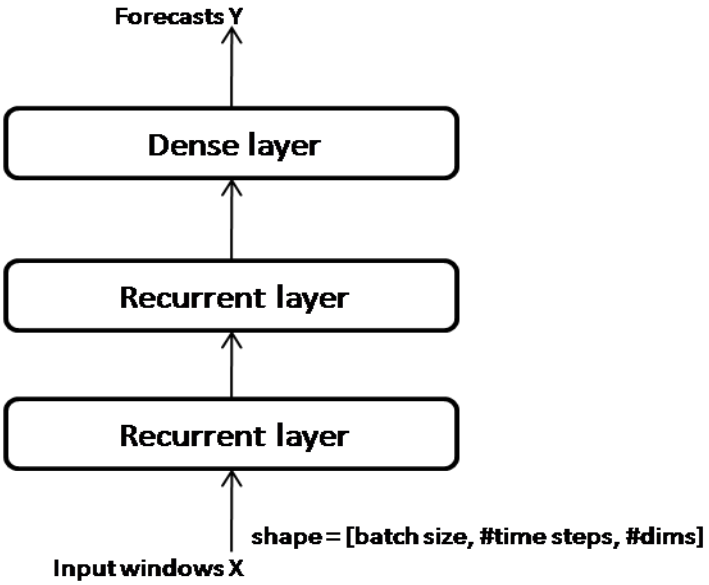

Fig. 7 Recurrent Neural Network

In Figure 8, it looks like as though there are so many RNN units, but actually it is only one unit being reused multiple times by the layer. In each time step, its corresponding input is fed to the memory cell. In this case zerois passed as input at time step 0. The output of time step 0 is computed. In figure, the output at time 0 is Y0. And the state vector is H0. The state vector H0 is fed as input to time step 1. At the same time the new input X1 is servedto the memory cell. The output of time step 1 is Y1 and its state vector is H1. Now the state vector H1 is fed as input to the next time step 2. This produces Y2 as the output and H2 as the state vector. The process continuous until the last time step in the given input dataset is reached. In figure 2.4 the numberof time steps is fixed to 30. This is the reason that this kind of architectureare called recurrent neural network.

The inputs are three dimensional. So for instance, if the window size is of 30 timestamps and they are batched in sizes of four, the shape will be 4x30x1. And at each timestamp, A four by one matrix will be used as the memory

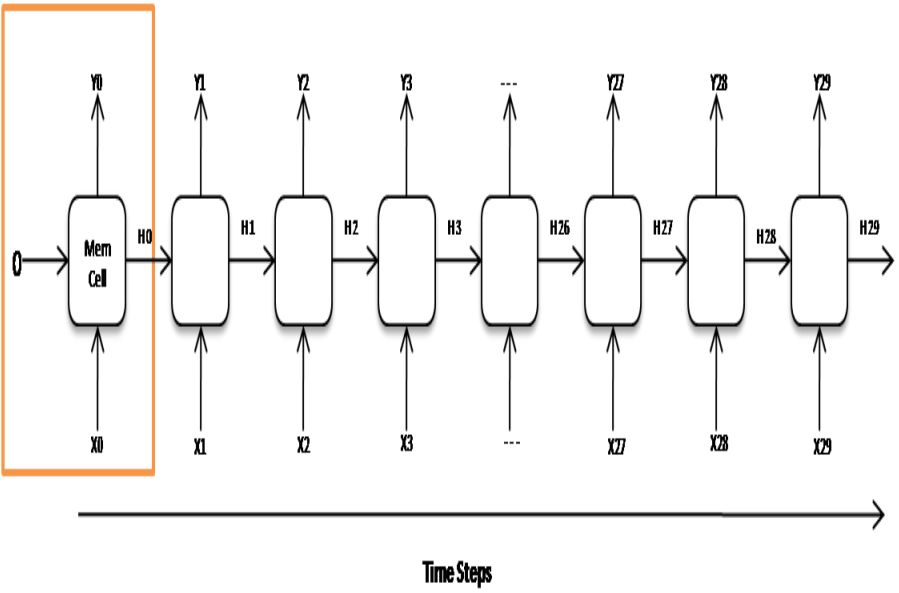

Fig. 8 Unfolding RNN with Memory Cell

cell input. The state matrix from the prior step will also be fed into the cell. Naturally, this will be zero in this instance and in the first step.

5 Dataset

This section describes the datasets used in the cloud network security issues. It also mentions the companion websites for these datasets.

- **UNSW-NB15** [25] UNSW is a dataset on network intrusion. It contains nine different attacks, include fuzzers, backdoors, Danial of service attacks, and worms. The collection also includes raw network packets. The testingset consists of 82,332 records from the attack and normal kinds, while the learning set is made up of 175 341 records.
- **Dropbox Dataset** [26] The experiment can make use of the Home 1 and Campus 2 Dropbox monitoring datasets. Customers of a national ISP who subscribe to Fiber to the Home and ADSL pay for the Home 1dataset, but they could be able to share the connection at home with WiFirouters. Instead, data from Campus 2 was gathered in academic settings like wired workstations in administration and research buildings and wi-fi access points all over campus.
- **NASA Web Server Logs** In order to identify web server-based threats, we may consider two samples of publicly available month-long Apache web-server logs from July 1995 and August 1995 at NASA [27]. The data in the logs (in hours) were processed to produce a time series showing the numberof HTTP requests sent to the server as a function of time.

- **Google Traces** Users frequently conduct particular kinds of tasks on their computers, which causes certain patterns in the data they collect about resource utilisation. We may leverage published user level data gathered from a Google 12.5k-machine cluster to capture this. The trace covers the entire month of May 2011 [28]. Since each trace had hashed usernames (of Google engineers) and timestamps for metrics like mean CPU utilisation rate, canonical memory consumption, and mean disc I/O time, among others, it was possible to create a user level time-series for each of these measures.
- **Kubernetes Cluster Data** [29] The collection contains numerous traces from systems and software used by Google to manage clusters. Four microservice benchmarks were deployed on our dedicated Kubernetes cluster, which has 15 different nodes. The dataset is not sampled and is drawn from a subset of requests for each benchmark, including requests to create posts on social networks, write reviews for media services, book rooms in hotels, and reserve tickets for trains.
- **Alibaba Cluster Trace** [30] gathers comprehensive statistics for the co-located workloads of batch and long-running operations over a 24-hour period. Three sections make up the trace: (1) statistics of the homogeneous cluster of 1,313 machines under investigation, including system configurations and runtime CPU, memory, and disc resource utilisation for 12 hours (the second half of a 24-hour period); (2) a record of all container deployment requests and actions, as well as a resource use trace for the previous 12 hours; (3) Co-located batch job workloads, providing a reduced resource usage trace and a history of all batch job requests and actions;

6 Conclusion

The most difficult problems in cloud computing were examined in this study, specifically security threats and attacks. Customer data stored in the cloud is extremely important, and its security must never be compromised. The researchers deploy a number of innovative technologies and security methods to increase the security of the cloud environment. There is a lot of room for machine learning to improve accuracy and automate defence against both known and unidentified cloud assaults. The primary objective of this review study is to provide an overview of recent machine learning-based cloud security research. In future, we propose an attack detection system using machine learning techniques that will make use of enhanced and optimized machine learning algorithms in order to provide more accurate cloud data security.

References

[1] Munish Saran, Rajan Kumar Yadav, Upendra Nath Tripathi, “Machine Learning based Security for Cloud Computing: A Survey”, IJAER ISSN 0973-4562 Volume 17, Number 4 (2022) pp. 332-337

[2] Deval Bhamare, Tara Salman, Mohammed Samaka, Aiman Erbad, Raj Jain - "Feasibility of Supervised Machine Learning for Cloud Security",

[3] Gopal Krishna Shyam, Srilatha Doddi - "Achieving Cloud Security Solutions through Machine and Non-Machine Learning Techniques: A Survey" ISSN: 1791-2377 © 2019

[4] Jeffrey C Kimmell, Mahmoud Abdelsalam, Maanak Gupta - "Analyzing Machine Learning Approaches for Online Malware Detection in Cloud", arXiv:2105.09268v1

[5] Ayesha Sarosh, "Machine Learning Based Hybrid Intrusion Detection ForVirtualized Infrastructures In Cloud Computing Environments" Jour- nal of Physics: Conference Series, doi:10.1088/1742-6596/2089/1/012072 Tanja hajemann, katerina Katsarou - "A Systematic review on anomaly detection for cloud computing environments"

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Lars Buitinck, Gilles Louppe, "API design for machine learning software: experiences from the scikit-learn project", arxiv, Sep 2013

[14] Keider Hoyos-Osorio, Jairo Castaneda-Gonzalez, Genaro Daza-Santacoloma, "Automatic epileptic seizure prediction based on scalp EEG and ECG signals", IEEE, 2016

[15] Khanum, Salma, and L. Girish. "Meta heuristic approach for task scheduling in cloud datacenter for optimum performance." International Journal of Advanced Research in Computer Engineering \& Technology (IJARCET) 4.5 (2015): 2070-2074.

[16] Mell, P. and Grance, T. (2011). The NIST definition of cloud computing. National Institute of Standards and Technology.

[17] Kandukuri, B., Paturi, V., and Rakshit, A. (2009). Cloud security issues. In IEEE International Conference on Services Computing,2009. SCC '09. pages 517–520.

[18] Hwang, K., Kulkareni, S., and Hu, Y. (2009). Cloud security with virtualized defense and reputation-based trust mangement. In Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009. DASC '09., pages 717–722.

[19] Plonka, D. and Barford, P. (2009). Network anomaly confirmation, diag- nosis and remediation. In Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on, pages 128–135.

[20] Dainotti, A., Pescape, A., and Claffy, K. (2012). Issues and future directions in traffic classification. Network, IEEE, 26(1):35–40.

[21] Rashmi TV, Prasanna MK, Girish L (2015) Load balancing as a service in Openstack-Liberty. Int J Sci Technol Res 4(8):70–73

[22] Subashini, S. and Kavitha, V. (2011). A survey on security issues in service delivery models of cloud computing. Journal of Network and Computer Applications, 34(1):1–11.

[23] Le Duc, T.; Leiva, R.G.; Casari, P.; Östberg, P.O. Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey. ACM Comput. Surv. 2019, 52, 1–39

[24] Li, K.; Gibson, C.; Ho, D.; Zhou, Q.; Kim, J.; Buhisi, O.; Gerber, M. Assessment of machine learning algorithms in cloud computing frameworks. In Proceedings of the IEEE Systems and Information Engineering Design Symposium, Charlottesville, VA, USA, 26 April 2013; pp. 98–103.

[25] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network dataset)," 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 2015, pp. 1-6, doi: 10.1109/MilCIS.2015.7348942.

[26] Drago, I., Mellia, M., Munafò, M. M., Sperotto, A., Sadre, R., and Pras, A. (2012). Inside Dropbox: Understanding Personal Cloud Storage Ser- vices. In Proceedings of the 12th ACM SIGCOMM Conference on Internet Measurement. Berlin, Germany., IMC'12, pages 481–494.

[27] J. Dumoulin, "Nasa http webserver logs." [Online]. Available: http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html

[28] C. Reiss, J. Wilkes, and J. L. Hellerstein, "Google cluster-usage traces: format + schema," Google Inc., Mountain View, CA, USA, Technical Report, Nov. 2011, revised 2014-11-17 for version 2.1. Posted at https://github.com/google/cluster-data.

[29] Girish L, Rao SKN (2020) Quantifying sensitivity and performance degradation of virtual machines using machine learning. J Comput Theor Nanosci. https://doi.org/10.1166/jctn.2020.9019

[30] Girish L, and Raviprakash M L. "Data Analytics in SDN and NFV: Techniques and Challenges". International Journal of Advanced Scientific Innovation, vol. 4, no. 8, Dec. 2022, doi:10.5281/zenodo.7657569.