

IMPLEMENTATION OF HYBRID MACHINE LEARNING TECHNIQUE FOR INTRUSION DETECTION SYSTEM IN CLOUD COMPUTING

¹Dr. E. POORNIMA, ²Dr. C. SASIKALA

¹Associate Professor, Dept of CSE, G Pulla Reddy Engineering College, Kurnool, Andhra Pradesh, India

²Associate Professor, Dept of CSE, Srinivasa Ramanujan Institute of Technology (Autonomous) Anantapuramu, Andhra Pradesh, India

ABSTRACT: The emergence of new networking models, such as Cloud Computing and the Internet of Things, has created new security challenges, requiring the development of new mechanisms to ensure the integrity, availability, and security of information and service data. In order to create a secure and trustworthy cloud computing environment, intrusion detection is a critical tool. Accordingly, an intrusion detection system is needed in cloud environments to detect new and unknown attacks with high accuracy. One of the most researched strategies that meets these requirements is the use of hybrid machine learning techniques to automate the intrusion detection process. This paper describes how to implement a hybrid machine learning in a cloud computing intrusion detection system (IDS). The intrusion detection system is implemented as a network space at the cloud hypervisor level, which improves the intrusion detection system's accuracy. The system employs a hybrid algorithm that combines K-means clustering and SVM (Support Vector Machine) classification algorithms. The analysis shows that ML hybrid technology has higher detection accuracy than other ML technology for model development work that can detect well-known attacks.

KEY WORDS: Cloud Computing, Intrusion Detection System (IDS), Hybrid Machine Learning, K-means clustering and SVM (Super Vector Machine).

I. INTRODUCTION

Over the years, the Internet has expanded significantly. It connects not only computer networks but also a global network of devices using Big Data. The Internet enables a wide range of innovations in any sector, including education, healthcare, public services, financial technology and digital commerce [1].

Despite its many benefits, the Internet can harbour malicious activities and cyber-attacks that can harm anyone who connects to it. Intrusion detection system (IDS) is used to detect and identify any cyber-attacks entering the network [2]. Intrusion detection systems detect threats using two methods: signature-based detection and anomaly-based detection. The process of training data affects the quality of the resulting machine learning model in anomaly-based detection. The most challenging aspect of the machine learning methods is developing an appropriate model to represent the data set [3].

Machine learning is a promising solution to address the security challenges in the technology field today [4]. It is a type of artificial intelligence that makes use of different learning algorithms to train devices without explicit programming. Its use of mathematical models, data sets, and dynamic (regular and irregular) data behavioural patterns and learning algorithms that require no human intervention, makes it applicable for the defence against new threats. A few challenges have been identified which are limitation in computational resources and the need for new data sets required for learning [5]. Machine learning has been applied in recent times to solve cyber security challenges, most notable of these challenges are software application, system and network vulnerabilities using network security systems in the form of firewalls, antivirus software and network intrusion detection systems (IDS) [6].

This study is unique in that it uses a feature extraction technique that uses multiple parameters and selects only the most useful features to improve the results of the clustering algorithm [7]. The high false positive rate when detecting anomalies in the intrusion detection system is due to the following factors: The number of data samples used to train the model is usually insufficient to make it accurate enough for deployment in a production environment. In intrusion detection systems, even a small increase in false positive rates can have a large negative

impact on system performance in a production environment. Anomalies are defined differently by each organization, making it difficult to establish a global standard for normal or anomalous traffic [8]. Network traffic is highly variable due to the variety of network applications. A new combined detection method for network-based anomaly detection in cloud computing networks [9] is proposed with these difficulties. The feature selection method, which represents supervised learning, and the data reduction method, which represents unsupervised learning, are combined in this study to give a combined machine learning approach to develop an appropriate model.

II. LITERATURE SURVEY

Bostani and Sheikhan [10] According to the paper, The Internet and wireless sensor networks, which are critical components of IoT, are both insecure and vulnerable to a variety of attacks. The same author proposes a new real-time intrusion detection framework based on an anomalous intrusion detection module as well as a specification for detecting two types of routing attacks: IoT swarm and selective routing attacks. Alvarenga et al. [11] as cyber security dangers have permeated most daily activities, this paper discusses security challenges, especially the connection of IoT and real devices to the Internet. Attacks on critical infrastructure such as power plants and public transport can devastate cities and countries as a whole. The author presented a study of approaches to IoT intrusion detection systems and a classification method for classifying the articles used in this study based on criteria such as detection methods, IDS deployment strategies, security threats, and validation strategies.

Yang et al. [12] presented According to one study, the Internet of Things is envisioned as a vast network of small gadgets. To overcome the limitations of previous studies, Using state estimation and sequential hypothesis testing, an anomaly detection-based approach to protecting data aggregation from false data injection (FDI) attacks has been presented. The results show that their method can reliably detect compromised aggregators even if the aggregators perform low-frequency and intensive FDI attacks. In terms of IoT security issues, the search for any attacks or vulnerabilities is still on.

Airehrour et al. [13] describes an interest in learning more about IoT routing protocols and their vulnerabilities to attacks. This study is one of the best in our knowledge, as it provided a comprehensive overview of the various results and potential solutions to the problem of secure routing protocols between IoT devices.

As a result of their inherent ability to detect an intrusion in real-time, Intrusion Detection Systems (IDSs) have grown in popularity over the last few decades. For example, the authors of [14] provided a useful overview of the importance of security properties in cloud computing platform monitoring. The authors of [15] proposed an outstanding signature-based intrusion detection system that significantly improved the detection rate of injections within a database. Furthermore, by extracting only patterns with suspicious content, this approach significantly reduces throughput in the signature database. However, it is only intended for database intrusion and cannot be generalised as misused intrusion detection for a traditional network environment with varying behaviour and functionalities.

III. HYBRID MACHINE LEARNING TECHNIQUE FOR IDS

This section describes the hybrid detection model as well as the developed technique. Hybrid approaches outperform traditional methods in Intrusion Detection systems. Therefore, the focus of this research is the application of hybrid techniques for analyzing contextual relationships between data flows in a network. The focus here is to use the K-means clustering algorithm to automatically generate labels and to use SVMs to create learning models. Fig.1 shows the Framework of a hybrid machine learning technique for IDS.

3.1 Packet analysis

As a result, dynamic traffic analysis and traffic classification are important ideas in traffic analysis. A network flow is a series of network packets sent between two endpoints that the hypervisor intercepts. Monitoring and network management technologies are used to perform flow analysis. Wireshark is used in this study as a free and open source packet analyzer to generate TCP dump files. A dump file is a collection of packets that were sent between hosts via a network connection. After monitoring network flows and extracting features, the output data set will be focused on removing noisy and unreliable data.

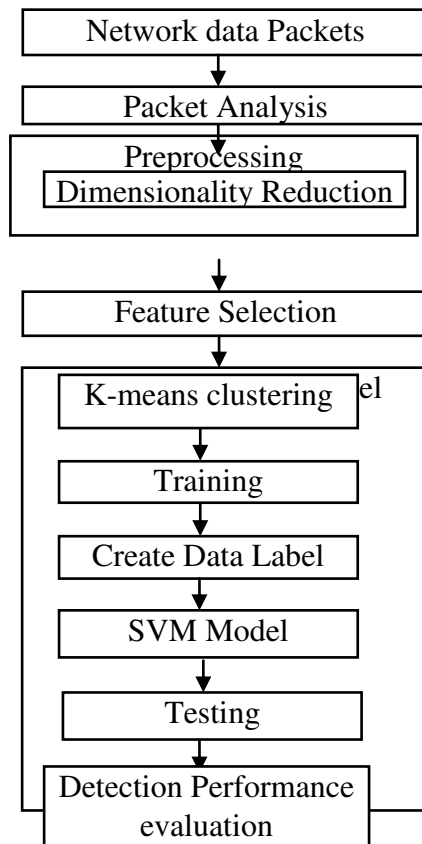


Fig. 1: Framework Hybrid Machine Learning Technique for IDS

3.2 Data Pre-Processing

Intruder detection pre-processing activity is essential to improve accuracy. Dimensionality reduction: Some features do not improve detection and may be difficult to detect. As a result, deleting characteristics with a little variance may even improve outcomes because these dimensions are essentially noise, reducing the weight of features with more information. To remove these features, Principal Component Analysis (PCA) is used. As a result, each instance's detection method will only take a portion of the feature vector's components as input. PCA is a technique for reducing the number of features in a dataset from all of them.

3.2 Feature selection

The feature selection approach reduces data size and provides faster and more accurate models by determining which aspects of the data set actually provide information to the ML model. Here, progressive function reduction (GFR) is used for function selection. Feature selection takes the entire set of features, temporarily removes one, and evaluates the model's performance along with the rest. Features that are found to be less important in terms of performance will be permanently removed. This method is run 'n' times. Where 'n' is the total size of the feature set. When the last iteration was complete, the methodology determined the most important features, sorted the other features in order of importance, and removed the least important features during the first iteration of the algorithm.

3.3 Hybrid detection Model

The study of computer programmes' ability to learn without being specifically designed to do so is known as machine learning. If the performance of the program improves at T, as measured at P as a result of experience E, then it is said that you are learning from experience E in terms of task T and the performance measurement P. As a result, machine learning models for anomaly detection can be widely used in this type of IDS. In contrast to supervised learning, unsupervised learning is an ML technique that does not have a set of labels Y that can identify

every entry in the set of predicting variables X. Instead, this method focuses on finding patterns that allow the dataset to be described. Supervised learning is a type of machine learning in which a set of predictors X and a set of labels Y are presented, and each item in X is identified as a member of the set Y. As a result, ML models have teachers. The form of a set of Y that guides the model through the training process. Classification and regression problems are usually solved using this type of machine learning. Because supervised learning methods are faster to test than unsupervised learning methods, we choose Support Vector Machines (SVMs) as the test detection model and K-means clustering, and unsupervised machine learning model for training.

K-means clustering: K-means is one of the well-known data mining clustering algorithms based on centroids and has been used to detect abnormal network user behavior in network traffic. Supervised learning is a type of machine learning in which a set of predictors X and a set of labels Y are presented, and each item in X is identified as belonging to the set Y. As a result, ML models have teachers. The shape of a set of Y directs the model through the training process. On the other hand, The algorithm requires a number of centroids to be specified at startup, is susceptible to noise (outliers), and can produce significantly divergent results depending on how the centroids are initialised.

Support Vector Machine (SVM): Some of the most used supervised learning algorithm is SVM. Support Vector Machines (SVMs) are a popular, reliable and accurate machine learning approach. The new label data is fed into the SVM using the polynomial kernel to generate the trained model. The learnt model will then be applied to the new network traffic to detect any additional anomalies. SVM is a machine learning technique that connects input data to a higher-dimensional characteristic space before constructing the best separating hyperplane in that space. The separating hyperplane is chosen by maximising the distance between the support vectors that are closest to the separating hyperplane and reflect the boundaries of different classes. Kernel functions are used to create the hyperplanes, which allow for a variety of separations.

3.4 Detection Performance Evaluation

Confusion matrix is used in this work as a detection performance evaluation method. One of the most widely used performance indicators among machine learning researchers is the confusion matrix. Confusion matrices and tables can be used to graphically represent the performance of a machine learning model for a specific task. It is based on four main factors. 1) true Positive (TP) and 2) true Negative (TN), which represent the number of elements correctly identified as negative or positive. 3) False Negatives (FN) and 4) False Positives (FP) represents the number of records incorrectly classified as negative or positive.

Table 1: CONFUSION MATRIX

	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 1 shows a confusion matrix for the two classes that can be observed, with the actual values of the inputs in the first column and the model predictions in the first row. Then, diagonally conforming entries reflecting model success were distributed between TP and TN, while misclassifications, distributed between FN and FP, were outside of the diagonal.

IV. RESULTS

The KDDCup99 dataset is used to evaluate the anomaly detection method in the intrusion detection system using hybrid machine learning techniques. The dataset was used to develop an intrusion detection kit for the 3rd International Knowledge Discovery and Data Mining Tool Competition, which is a predictive model capable of discriminating between results. "bad" connections, such as an intrusion or attack, and normal "good" connections. This competition resulted in the collection of a large number of Internet traffic records, which were then collected into the KDDCup99 dataset. There are 41 properties in the KDDCup99Dataset, along with a label or class. There are five main classes in the KDD99 data set: probe attacks, denial of service (DOS) attacks, root user (U2R) attacks, and Remote-to-Local (R2L) attacks (normal class and four main intrusion classes, and probe attack, denial of service (DOS), user root attack (U2R), and Remote-to-Local (R2L) attack). The KDD dataset is mostly made up of four

different sorts of attacks. The confusion Matrix is created for each dataset based on the labels supplied by the testing algorithm to evaluate the performance of a hybrid machine learning approach intrusion detection system. The following parameter metrics are evaluated using the confusion matrix method.

Accuracy is the percentage of entries successfully categorised by the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \text{ --- (1)}$$

- The error rate is the percentage of entries that are incorrectly classified.

$$ErrorRate = (100 - HitRate) \text{ ---- (2)}$$

- The accuracy of the model is determined by the percentage of correctly classified entries.

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \text{ ----- (3)}$$

- The percentage of entries identified as negative that were correctly classified is referred to as specificity.

$$Specificity = \frac{TN}{TN + FP} \times 100 \text{ ----- (4)}$$

- Precision is the percentage of hits in the positive class that are classified as positive.

$$Precision = \frac{TP}{TP + FP} \times 100 \text{ ---- (5)}$$

The following Fig. 2 shows the detection performance analysis of Hybrid ML IDS. It can be seen that the detection performance is analyzed by using five parameter metrics such as accuracy, error rate, sensitivity, specificity and precision.

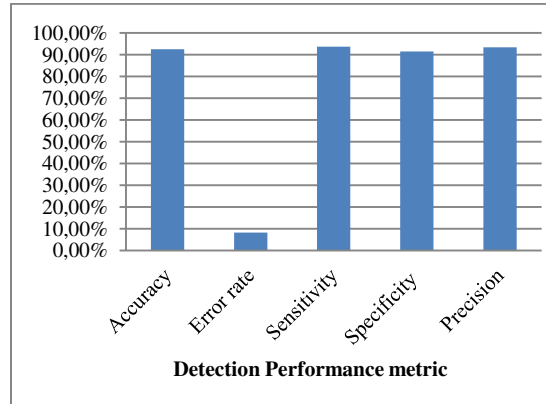


Fig. 2: Detection Performance Analysis of Hybrid ML IDS

A higher accuracy more than 90% and lower error rate less than 10% is achieved for the detection of intrusions in networks. Similarly high sensitivity, specificity and precision rates are achieved with the hybrid machine learning technique (SVM+K-means) used.

Fig. 3 shows the detection accuracy of hybrid ML technique for different intrusion classes of KDD-99 dataset. It can be seen that the DOS class is detected with high accuracy where as probe class can detect with less accuracy by using hybrid machine learning technique in intrusion detection.

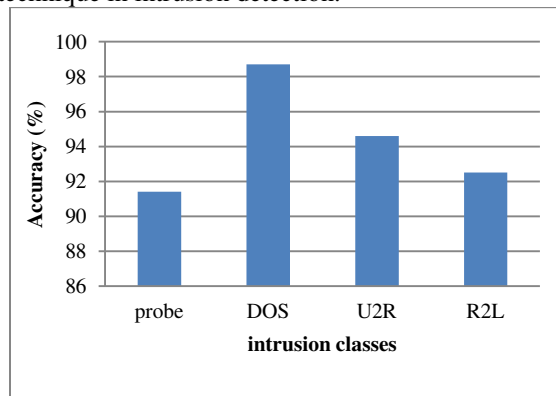


Fig. 3: Detection Accuracy of Hybrid ML Technique for Different Classes

Fig. 4 shows the intrusions detection accuracy and error rate analysis of ML techniques. Here, various machine learning techniques such as SVM, Artificial Neural Networks (ANN), Naïve Bayes (NB) and K-means clustering are compared with our hybrid machine learning technique (SVM+K-means) in detecting intrusions in terms of accuracy and error rate parameters. It can be seen from Fig. 4 that hybrid model has a highest accuracy of 92.3% and lowest error rate of 8.3% compared to all among techniques where as ANN has a lowest accuracy of 74.5% and highest error rate of 25.3%.

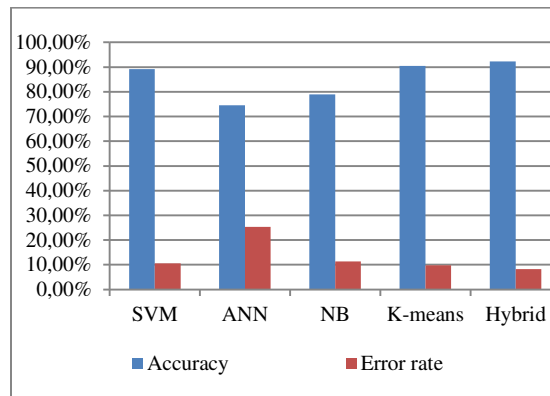


Fig. 4: Detection Accuracy and Error Rate Analysis of ML Techniques

V. CONCLUSION

This paper presents a hybrid machine learning technique for intrusion detection systems in cloud computing. In this case, hybrid machine learning was used to detect intrusions in a cloud computing environment by combining unsupervised machine learning and supervised machine learning methods. Unsupervised model of K-mean clustering is used for training and the supervised model of SVM is used for testing in the hybrid detection phase. Hybrid intrusion detection systems are network-based techniques that use the K-means algorithm to group network flows into clusters. The cluster is then divided into two categories: normal events and abnormal events. Then use the SVM approach to train a monitored detection model based on the data instance. Our detection model is an SVM trained model that detects anomalies in instances. The KDD-Cup99 Dataset is a publicly accessible data set used to test the detection algorithm. The overall performance of hybrid machine learning intrusion detection system is observed with five evaluation parameter metrics.

VI. REFERENCES

- [1] Indrajit Das, Shalini Singh, Ayantika Sarkar, “Serial and Parallel based Intrusion Detection System using Machine Learning”, 2021 Devices for Integrated Circuit (DevIC), 2021
- [2] Lin Chen, Xiaoyun Kuang, Aidong Xu, Siliang Suo, Yiwei Yang, “A Novel Network Intrusion Detection System Based on CNN”, 2020 Eighth International Conference on Advanced Cloud and Big Data (CBD), 2020
- [3] Azar Abid Salih, Maiwan Bahjat Abdulrazaq “Combining Best Features Selection Using Three Classifiers in Intrusion Detection System” at International Conference on Advanced Science and Engineering (ICOASE), University of Zakho, Duhok Polytechnic University, Kurdistan Region, Iraq in 2019
- [5] Aditya Phadke, Mohit Kulkarni, Pranav Bhawalkar And Rashmi Bhattad “ A Review of Machine Learning Methodologies for Network Intrusion Detection” at 3rd National Conference on Computing Methodologies and Communication (ICCMC 2019) IEEE Xplore Part Number: cfp19k25-art; isbn; 978-1-5386-7807- 4 in 2019.
- [6] Hassan Azwar, Muhmmad Murtaz, Mehwish Siddiquie, Saad Rehman “ Intrusion Detection in Secure Network for Cyber security Systems Using Machine Learning and Data Mining” at IEEE 5th International Conference on Engineering Technologies \$ Applied Sciences, 22-23 Nov 2018, Bangkok Thailand in 2018
- [7] Xiaoyan Wang, Hanwen Wang “A High Performance Intrusion Detection Method Based on Combining Supervised and Unsupervised Learning” at IEEE Smart World, Ubiquitous Intelligence \$ Computing Advanced \$ Trusted Computing, Scalable Computing, Internet of People and Smart City Innovations in 2018.

- [8] Pinjia He, Jieming Zhu, Shilin He, Jian Li and Michael R. Lyu, "A Feature Reduced Intrusion Detection System Using ANN Classifier", *ELSEVIER Expert Systems with Applications*, vol. 88, pp. 249-247, December 2017
- [9] Wathiq Laftah Al-Yaseen, Zulaiha Ali Othman and Mohd Zakree Ahmad Nazri, "Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System", *ELSEVIER Expert System with Applications*, vol. 66, pp. 296-303, Jan 2017.
- [10] H. Bostani, M. Sheikhan, Hybrid of anomaly-based and specification-based ids for internet of things using unsupervised opf based on mapreduce approach, *Comput. Commun.* 98 (Supplement C) (2017) 52–71.
- [11] B.B. Zarpelao, R.S. Miani, C.T. Kawakani, S.C. Alvarenga, A survey of intrusion detection in internet of things, *J. Netw. Comput. Appl.* 84 (2017) 25–37. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1084804517300802>.
- [12] L. Yang, C. Ding, M. Wu, K. Wang, Robust detection of false data injection attacks for data aggregation in an internet of things-based environmental surveillance, *Comput. Networks* 129 (2017) 410–428. Special Issue on 5G Wireless Networks for IoT and Body Sensors. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1389128617302372>.
- [13] D. Airehrour, J. Gutierrez, S.K. Ray, Secure routing for internet of things: a survey, *J. Netw. Comput. Appl.* 66 (2016) 198–213. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1084804516300133>
- [14] Muñoz, A.; Maña, A.; González, J. Dynamic Security Properties Monitoring Architecture for Cloud Computing. In *Security Engineering for Cloud Computing: Approaches and Tools*; IGI Globa: Hershey, PA, USA, 2013; pp. 1–18.
- [15] Zhang, Y.; Ye, X.; Xie, F.; Peng, Y. A Practical Database Intrusion Detection System Framework. In *Proceedings of the 2009 Ninth IEEE International Conference on Computer and Information Technology*, Xiamen, China, 11–14 October 2009; Volume 1, pp. 342–347
- Result