JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2501-2511
https://publishoa.com
ISSN: 1309-3452

Cloud Services Anomalies Detection Using Network Flowdata Analysis

Sreenivasa Chakravarthi Sangapu[1], R. Jagadeesh Kannan[2], V. Anantha Natarajan[3]

[1]Associate Professor, Amrita Viswa Vidyapeetham, Chennai,
[2]Professor, School of Computing Science and Engineering, VIT, Chennai,
[3]Associate Professor, Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, Tirupati, India.
Email: sschakravarthi@ieee.org[1], jagadeeshkannan.r@vit.ac.in[2], vananthanatarajan@vidyanikethan.edu[3]

Received 2022 April 02**; Revised** 2022 May 20**; Accepted** 2022 June 18

ABSTRACT
Cloud computing paved an excellent platform for the emergence of cost effective technological solutions. However, security and privacy issues still remain as a stringent challenge during service catering. Explicitly, the service utility anomalies are liable to cause severe privacy and security issues in cloud service delivery. So, the overall performance of Cloud service consumption and end-user applications' service levels utility is degraded. The open access and distributed nature of the cloud computing is the major reason for its vulnerability to intruders. The security and privacy in cloud services have many challenges and problems still open for research. This paper proposes an intrusion detection method capable of detecting nine categories of attacks in two stages. This paper focuses on establishing a network based intrusion detection mechanism using machine learning techniques. A model will be constructed with a supervised learning methodology using historical network flowdata and flowdata collected from Internet.
keywords: Cloud Services Anomalies, Malicious Users, Services Exploitation, Network Flowdata, Network Monitoring.

INTRODUCTION
Intrusion detection system (IDS)monitors the traffic in the network for detecting any vulnerable exploit against or within the target service. Thus, the foremostintentfor any intrusion detection systems (IDS)becomes the safety measures& privacyaugmentationin the network. The IDS can be designed using various approaches to detect the malicious nodes in the network. The three of IDS are Host based, Hypervisor based, and network based. In hosts based approach the IDS runs inside individual hosts and monitors the inbound and outbound data packets. The hypervisor based IDS,primarily, sits in between the host and the guest operating system, so as to detect the vulnerabilities through the kernel and its components. In Network based approach all the incoming and outgoing traffic is scanned in the network for detecting the attacks in the virtual networks. The cloud network is a multi-tenant environment where IDS introduces an optimal method to safeguard resources and services amongst the known and unknown attacks. Attackers launch an attack by utilizing the advantage of vulnerabilities in the virtual machine and are capable of destroying the host machine and the corresponding network. As each cloud service's architecture is usually

uncommon from other service oriented architectures, the IDS techniques to be used in each of such cloud services architectures shall also differs. The IDS techniques can be categorized into any one of the following types namely Artificial Neural Network (ANN) based Detection, Fuzzy rules based Detection,Anomaly Detection, and Signature based Detection.

The Signature based IDS refers to approach in which a specific pattern in the network traffic like byte sequences, or known malicious intrusion sequences followed by malware are detected. In general the efficiency of the signature based method is high but it fails to detect new attacks whose pattern is unknown. Anomaly based detection is used to detect new attacks whose pattern is unknown or not available. The approach is simple as it involves modeling using machine learning algorithm to recognize genuine activity within the network. This approach easily detects even a new attacks pattern but it suffers from false positive alarms. The probability of recognizing the known genuine traffic pattern as unknown malicious pattern is high. Most successful IDS tool or application consumes more time during detection which reduces its overall performance in the detection process even though accuracy of the detection is high. To make the detection more reliable an efficient feature selection algorithm can be used before modeling.

ANN based IDS have the capacity to analyze a large volume of data and efficiently detect intrusion patterns [14]. ANN based IDS can recognize attacks by learning from false detections. In [16] the proposed ANN based IDS approach detects 99.9% of the malicious patterns which falls under any of the following category namely DoS, Probe, Remote to Local, and user to root.

The IDS can be also be categorized as Host based or Network based IDS methods depending on the place where detection takes place. A cloud network may consist of one or more NIDS placed at strategic points where the data traffic to and from all the connected nodes can be monitored. The detection system analyses the every traffic in the network and informs administrator once it identifies any abnormal patter in the network.

The cloud service has unique some characteristics and differs entirely from other conventional service architectures. In the same way IDS techniques used for each type cloud services must be differ from conventional intrusion detection. The conventional IDS used in internet and other network environments lack are not scalable and they cannot adopt manage themselves on a distributed environment. Most of the existing IDS techniques don't have the potential to identify the intruders in real time and they work in offline in detecting intruders by analyzing the log files or records.

In cloud architecture the IDS can be deployed at various instance such as the network boundary, in a host node / virtual machine/ distributed across all nodes. This paper focuses on developing a network based intrusion detection model deployed at the network level to monitor all the inbound and outbound traffic. The proposed method can efficiently detect the known attack patterns with low false alarms. The approach is flexible as new attacks can be detected by adding those patterns in training dataset. The focus of this research is to detect the different possible attacks on the cloud network which possibly affects the valuable services provided in the respective cloud environment.

The focus of this research is to detect the different possible attacks on the cloud network which possibly affects the valuable services provided in the respective cloud environment. The aim is to analyze the suitability of applying machine learning techniques to discriminate between various attacks rather than just detecting the anomalous

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2501-2511
https://publishoa.com
ISSN: 1309-3452

traffic alone. A hybrid method is employed which comprises of both anomaly and signature based approach to detect the type of attack. The intrusion detection is carried out in two stages in the first stage a model screens the traffic and recognizes any abnormality and in the second stage the abnormality is further classified using a signature based method.

The rest of the paper is organized as follows. The Section 2 presents a detailed systematic survey of various literatures related to the proposed methodology. The Section 3 describes the dataset used for training and evaluation of the intrusion detection model. The Section 4 gives an overview of the methodology adopted in the work. The Section 5 gives the details of experiment conducted and analysis of the results obtained. Finally the section 6 concludes the paper with the future scope of work.

RELATED WORK

The present IT industry is experiencing a paradigm shift to cloud environment and hence new security solutions are essential to support the business functionality. This paper also studies the various taxonomies and classification of various attacks in different services delivery mode of cloud environment.

Software as a Service is kind of cloud service wherein the software is owned, delivered and managed by a service provider remotely. Many of the security feature aspects are very similar to security features of a web service. According to recent research survey 70% of the security attacks in SaaS are caused by internal nodes which can be eliminated by various security measures including anomaly based intrusion detection. From a detailed literature survey it is observed that the each of the layer in the cloud network suffers from various kinds of vulnerabilities. Hence an Intrusion Detection System (IDS) is essentially required to protect the cloud services or nodes against various attacks at different service modes.The general types of attacks in different cloud service modes are presented in Fig. 1

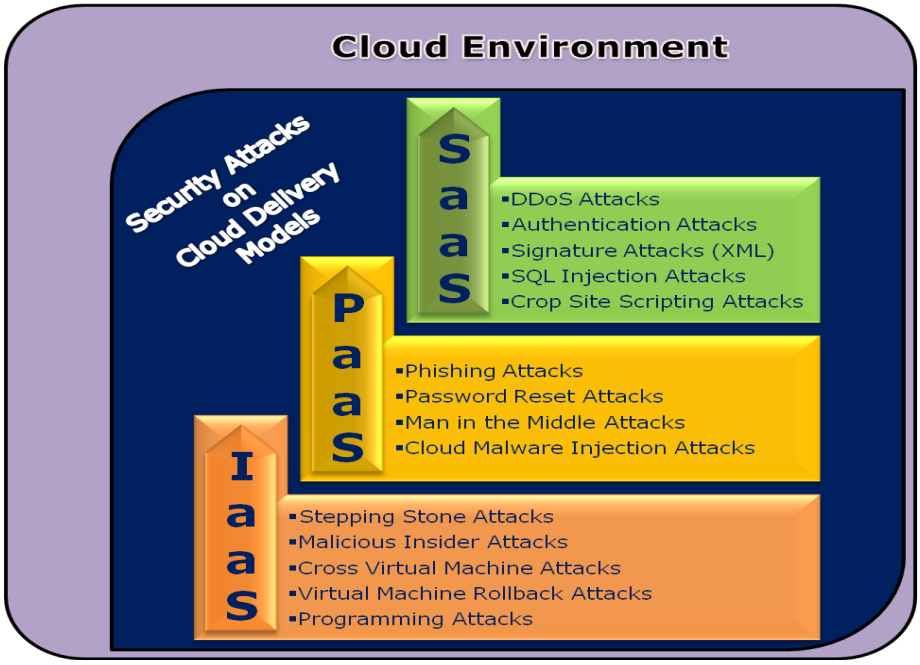

Fig 1. Types of attacks in different cloud service mode

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2501-2511
https://publishoa.com
ISSN: 1309-3452

Training an intrusion detection model on a imbalanced dataset poses various issues and challenges and it has received attention of many researchers [1]. From the results presented in various literatures, it was observed that the class imbalance problem was more serious in the case of the high-dimensional dataset [2]. In the high-dimensional dataset, the unfairnessduring the classification of majority classes is high though there are more cases of discrimination among the classes. The high-dimensionality influences the performance of various classifiers differently. All the proposed solutions in the literatures to overcome the class-imbalance problems are not `efficient in the case of high-dimensional data. The simple oversampling solution cannot resolve the class-imbalance problem as it does not change the classification rule [3]. The Synthetic Minority Over-sampling Technique(SMOTE) proposed in [9]follows an oversampling approach by synthesizing samples of minority classes. The classification results on these synthesized data samples are better when compared to the results of classification on the dataset oversampled using simple oversampling approach. SMOTE based oversampling approach was used for marking sentence boundary in continuous speech [6], intrusion detection [5], species distribution prediction [7] and breast cancer detection [8]. SMOTE was also used in bio-medical research for predicting gene [9, 10], annotating histopathology images [13], identification of binding specificity of proteins [11].

In [19] the authors have categorized the intrusion detection in a cloud environment into five categories based on the deployment location. They are InGuest agent based approach, InVMM agent based approach, Network Monitor based approach, Collaborative agent based approach, and Distributed approach. As per the given categorization the proposed method in this paper can be categorized as Network Monitor based approach. An intrusion detection model based on ensemble based multi filter feature selection is proposed in [20]. The selected features are used to train a decision tree based classifier. For training and evaluation of the detection model the authors have used the NSL-KDDD dataset.

DATASET

The various attack type detected using signature based approach using a deep neural network is discussed in this section. The dataset used for training and evaluation of the intrusion detection model is obtained from UNSW dataset [18]. The dataset is constructed using the packets generated from the IXIA Perfect Storm tool for both normal and various attack categories possible in the cloud network. The TCP logs were used to extract the relevant attributes which better represents the attacks. Using k-fold cross validation the dataset is divided for training and testing the detection model. Description about the various attacks along with normal data is summarized in Table 1.

Table 1: Dataset Records Distribution.

Attack Category	Records Count	Narration
Ordinary	2,218,761	Normal traffic pattern data.
Fuzzers	24,246	Attempt to suspend a service by injecting the random breeded data.
Analysis	2,677	This attack includes attacks on ports to scan, penetrate spam on cloud nodes.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2501-2511
https://publishoa.com
ISSN: 1309-3452

Backdoors	2,329	A mechanism to have unauthorized access to virtual machine bypassing the security features.
DoS	16,353	This attack makes a service unavailable to authorized user by causing momentarily interrupts or delaying the relevant service.
Exploits	44,525	Security vulnerability in operating system / software running in a virtual machine is utilized to leverage attacks.
Generic	215,481	A technique works against all blockciphers (with a given block and key size), without consideration about the structure of the block-cipher.
Reconnaissance	13,987	Comprises of all attempts to collect information by simulating
		attacks.
Shellcode	1,511	A tiny piece of code snippet which may be used as the payload for creating software vulnerability exploitations.
Worms	174	Attacker regenerates itself so that theystretchinto other computers in the cloud network. Repeatedly, it makes use of computer networks to stretch, and causing security failures on the target computer during their accessing.

METHODOLOGY

The intrusion detection is accomplished as a two stage process in first stage the malicious traffic is discriminated from the normal traffic using a binary classifier. The idea is to use a binary classifier to discriminate the normal and abnormal traffic. During the training process all category of attacks are labeled as negative samples and normal traffic patterns are labeled as positive. This approach is already adopted in various literatures and proved to be efficient in detecting the abnormality at the first stage [21]. Support Vector Machine (SVM) is used for discrimination of normal and

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2501-2511
https://publishoa.com
ISSN: 1309-3452

abnormal patterns and it constructs a decision boundary in the form of a hyperplane. Further, it is built in such a way that the distance from the origin is maximized. Fig.3 gives the detailed picture of the scheme to be adopted in the proposing model.

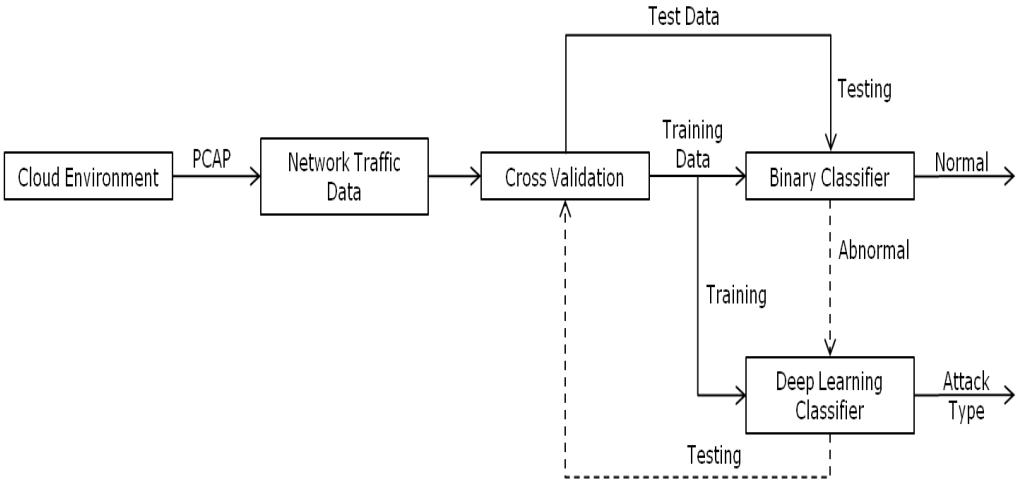

Fig. 3 Schematicview of the proposed model

Scientifically, let us assume x_i as one of the training examplein the dataset $X = \{x_1, \ldots, x_m\}$ in the input space. Let ϕ be a mapping function, that will map the input feature space *X* towards considerably high dimensional feature space *H*. Hence, the dot product in *H* is computed using following gaussian kernel function:-

$$\boldsymbol{K(x, x')} = \boldsymbol{e^{-\gamma\|x-x'\|^2}} \quad \textbf{Eq. 1}$$

where K represent the kernel of the classifier. The input vector and the support vector is represented as x and x' respectively. The value of the '$\boldsymbol{\gamma}$' is fixed by calculating the reciprocal of the number of attributes considered for intrusion detection.

The next step is to detect the category of the attack if the traffic pattern is observed to be abnormal in the first stage. At this stage the detection model is already trained to classify abnormality into any one of nine categories (listed in Table 1.) of attacks which targets different services offered in the cloud network. Convolution Neural Network is used to classify the abnormality as CNN and its variants have showed better performance in classification when compared to the conventional machine learning algorithm based classifiers. CNN performance better due to the fact that it has the capacity to extract higher level feature representation from the low level attributes of the network. The proposed CNN architecture contains five sequential 1D convolution layers apart from a fully connected layer at the end. The series of convolution layer extracts a better representation of the given attributes and send it to the fully connected layer for classification. During the training process the weights of the convolution layers are updated to extract an optimal representation of the input.

The Table 1 shows the distribution of samples of ten categories of classes considered for classification. It can be observed that the number of samples belonging to each class is not same – the classes are imbalanced. SMOTE is employed in the experiments to overcome the class imbalance in the dataset. SMOTE is one of the popularly used oversampling techniques that generate synthetic samples from the minority class available in the data. For each of the sample belonging to minority class, five random samples

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2501-2511
https://publishoa.com
ISSN: 1309-3452

having least Euclidean distance from the respective original sample is derived. Out of the five synthesized samples one was chosen randomly. This process is repeated for each of the sample until the dataset is balanced.

When applied on a high-dimensional data, SMOTE do not alter the mean value of the synthesized minority class samples but the variation between them. As a future work a feature selection method has to be designed to reduce the dimensionality of the input data and increase the overall performance of the intrusion detection model [17]. SMOTE does not change the importance of variable or ranking of the variable so the results of feature selection method before and after applying SMOTE will be same.

Let X denotes the set of samples and each of the samples belongs to class 'c' where c = 1 or 2 in case of binary classification or the value of c can be from the set C = {1,2,3,….,p } in case of multiclass classification. Let 'x'denotes a random sample from the set X and x_{ij}denotes the value of j^{th} (j = 1,……..n) variable for the i^{th} sample where i = {1,……..,m}. The proportion of samples is represented as $k_c = n_c/m$. To oversample random samples are taken from the set X and its k nearest neighbors are found. Find the vector between one of the k-nearest neighbor and current sample data point. Add the current sample and the vector after multiplying it with a random number between 0 and 1.

During training the CNN model overfitting can occur which makes the model to memorize training data and reduces the generalization capability of the model. Dropout is a regularization technique used to reduce overfitting in the network by dropping out certain neural units from the network. The hyper parameters of the model such as learning rate, decay rate of learning, initial weights of the network, the number of hidden units, batch size were tuned such that the overall accuracy of the network is increased.

EXPERIMENTS AND RESULTS

Before training the machine learning models on the chosen dataset the class imbalance issue is resolved by oversampling the dataset using SMOTE method described in section 4. Next the Support Vector Machine is trained to discriminate between normal and abnormal traffic patterns. The performance of the binary classifier is analysed based on the parameters namely precision, recall, and F1- score.

Table 2 summarizes the performance of the SVM classifier on the test dataset in detecting intruders after training. The test dataset consists of 82,332 samples out of which 37000 samples belong to normal traffic patterns and remaining 45332 are abnormal traffic patterns. The model has the capacity to detect the normal patterns more accurately when compared to the abnormal patterns. From the confusion matrix shown in Fig. 4 above plot the following statistics can be inferred.

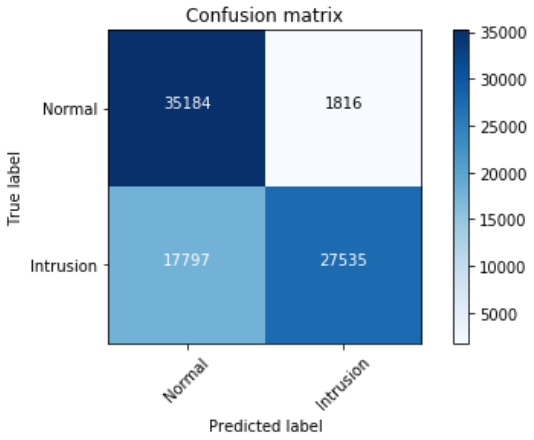

Fig. 4 Classification results of SVM on test dataset

Accuracy in detecting the Normal patterns = 95.09%

Accuracy in detecting the abnormal patterns = 60.7%

False Alarm Rate = 4.91%

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2501-2511
https://publishoa.com
ISSN: 1309-3452

Table2. Performance of SVM on test dataset

	Precision	Recall	F1-Score
Normal	0.66	0.95	0.78
Abnormal	0.94	0.61	0.74
Weighted Avg.	0.81	0.76	0.76

The abnormal traffic pattern categorization is implemented using a Convolution Neural Network which is trained with nine categories of abnormal traffic patterns. The hyper parameters of the model are tuned and the convolution neural network is trained with cross validation. The learning rate, loss function, density of hidden layers and the volume of units in hidden layer, momentum, weight decay rate, dropout regularization value are the some of the hyper-parameters of the network. A deep neural network configured with optimal set of hyper parameters will yield minimum value for chosen loss function. The conventional hyper parameter optimization technique uses a Grid Search method. The search method is an exhaustive and time consuming and it is guided by the cross validation score on the training data or evaluation on a separate validation data [13]. To eliminate the problem of overfitting dropout regularization is used thereby increasing the generalization ability of the network. As recommended in many literatures, a dropout value of 0.25 is chosen. Activation functions are used to give non-linearity to the model which makes the deep neural network model to learn nonlinear classification boundaries. Rectified Linear Units (ReLU) activation function is used in the hidden layer and softmax function is used in the output layer. ReLU activation is the simplest activation function and it can be represented by $f(x) = max(x, 0)$.Softmax activation gives the probabilities that a category is true. It is represented as

$$\sigma \quad Z_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \qquad \textbf{Eq. 2}$$

The hyperparameters of the model are selected using cross validation approach. The training dataset is divided into k-folds in such a way that each subset each subset is mutually exclusive and the union of all the subset gives the whole training data set. For hyperparameter selection k models are considered and trained with k-1 subsets of training data. k^{th} model is validated with the subset d_k and find the average error on all the model. For different combination of hyperparameter values repeat the above procedure and estimate the average error. Finally chose the set of hyperparameters for which the average error was low.

The Random Forest classifier and Naïve Bayes classifiers were also trained and their results were compared with the results of the CNN. The Fig. 5 presents the plot of accuracy of the CNN model during training and testing. The results obtained from other classifiers are tabulated in Table 3.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2501-2511
https://publishoa.com
ISSN: 1309-3452

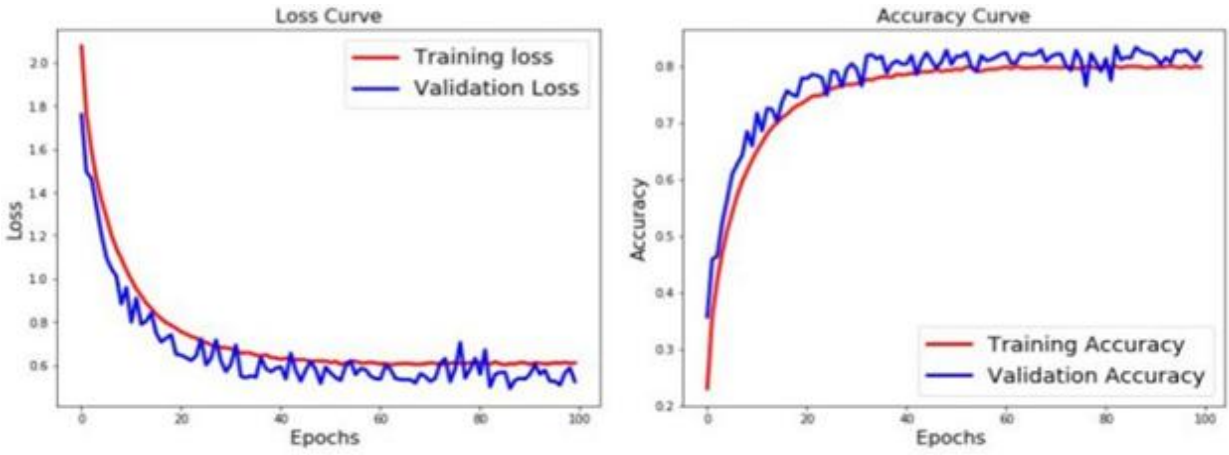

Fig 5. Plot of CNN Accuracy on training and test data set

From the Fig. 6 it is clear that the CNN achieves high detection rate and the Naives Bayes classifier which is a probabilistic model gives the lowest detection rate.

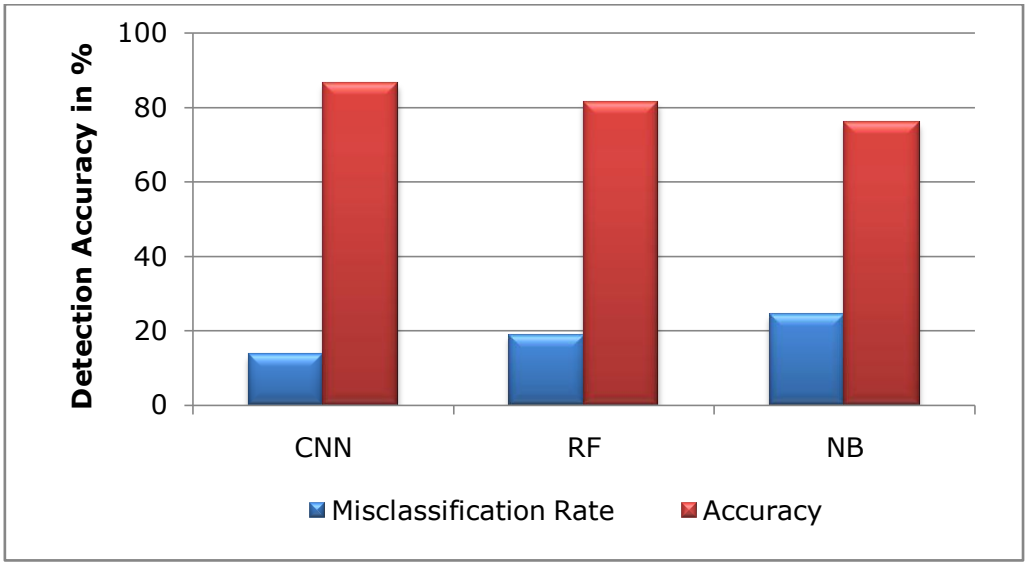

Fig. 6 Comparison of Classifiers in Detecting Attacks

CONCLUSION

This paper analyzed the start-of-art techniques used to detect the abnormal traffic patterns andproposed a deep learning based intrusion detection strategy. Probability of intrusion attacks in cloud environment is high when compared to other environment. As the traffic flow inside a cloud network isenormous a deep learning based intrusion detection scheme is an optimal one. The CNN baseddetection model hasdemonstrated varied level of efficiency in detecting the abnormalities. It was also observed that some of the features in the dataset less role in discriminating the different category of attacks. Future work will aim at selecting

an optimal set of features which will increase the overall detection rate.

REFERENCES

1. He H, Garcia EA: Learning from imbalanced data. IEEE Trans KnowledgeData Eng 2009, 21(9):1263–1284.
2. Blagus R, Lusa L: Class prediction for high-dimensionalclass-imbalanced data. BMC Bioinformatics 2010, 11:523+.
3. Hulse JV, Khoshgoftaar TM, Napolitano A: Experimental perspectives onlearning from imbalanced data. In Proceedings of the 24th internationalconference on Machine learning. Corvallis, Oregon: Oregon StateUniversity; 2007:935–942.
4. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: syntheticminority over-sampling technique. J ArtifIntell Res 2002, 16:341–378.
5. Cieslak DA, Chawla NW, Striegel A: Combating imbalance in networkintrusion datasets. In Proc IEEE Int Conf Granular Comput. Atlanta,Georgia, USA; 2006:732–737.
6. Liu Y, Chawla NV, Harper MP, Shriberg E, Stolcke A: A study in machinelearning from imbalanced data for sentence boundary detection inspeech. Comput Speech Lang 2006, 20(4):468–494.
7. Johnson R, Chawla N, Hellmann J: Species distribution modelling andprediction: A class imbalance problem. In Conference on IntelligentData Understanding (CIDU); 2012:9–16. doi:10.1109/CIDU.2012.6382186.
8. Fallahi A, Jafari S: An Expert System for Detection of Breast CancerUsing Data Preprocessing and Bayesian Network. Int J Adv Sci Technol2011, 34:65–70.
9. Batuwita R, Palade V: microPred: effective classification of pre-miRNAsfor human miRNA gene prediction. Bioinformatics 2009, 25(8):989–995.
10. Xiao J, Tang X, Li Y, Fang Z, Ma D, He Y, Li M: Identification of microRNAprecursors based on random forest with network-levelrepresentation method of stem-loop structure. BMC Bioinformatics2011, 12:165+.
11. MacIsaac KD, Gordon DB, Nekludova L, Odom DT, Schreiber J, Gifford DK,Yung RA, Fraenkel E: A hypothesis-based approach for identifyingthe binding specificity of regulatory proteins from chromatinimmunoprecipitation data. Bioinformatics 2006, 22(4):423–429.
12. Wang J, Xu M, Wang H, Zhang J: Classification of imbalanced data byusing the SMOTE algorithm and locally linear embedding.In International Conference on Signal Processing. Guilin, China; 2006.
13. Doyle S, Monaco J, Feldman M, Tomaszewski J, Madabhushi A: An activelearning based classification strategy for the minority class problemapplication to histopathology annotation. BMC Bioinformatics 2011,12:424+.
14. Garzia, Fabio; Lombardi, Mara; Ramalingam, Soodamani (2017). An integrated internet of everything — Genetic algorithms controller — Artificial neural networks framework for security/safety systems management and support. 2017 International Carnahan Conference on Security Technology (ICCST). IEEE.
15. Vilela, Douglas W. F. L.; Lotufo, Anna Diva P.; Santos, Carlos R. (2018). Fuzzy ARTMAP Neural Network IDS Evaluation applied for real IEEE 802.11w data base. 2018 International Joint Conference on Neural Networks (IJCNN).
16. Dias, L. P.; Cerqueira, J. J. F.; Assis, K. D. R.; Almeida, R. C. (2017). Using artificial neural network in intrusion detection systems to computer networks. 2017 9th Computer Science and Electronic Engineering (CEEC).
17. Dudoit S, Fridlyand J, Speed TP: Comparison of discriminationmethods for the classification of tumors using gene expressiondata. J Am Stat Assoc 2002, 97(457):77–87.
18. M. Nour, and J. Slay, “UNSW-NB15: a comprehensive data set for network

intrusion detection systems (UNSW-NB15 network data set),” in IEEE Military Communications and Information Systems Conference (MilCIS), 2015, pp. 1-6.

19. Mishra, Preeti, et al. "Intrusion detection techniques in cloud environment: A survey." *Journal of Network and Computer Applications* 77 (2017): 18-47.
20. Osanaiye, Opeyemi, et al. "Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing." *EURASIP Journal on Wireless Communications and Networking* 2016.1 (2016): 130.
21. Kang I, Jeong MK, Kong D. A differentiated one-class classification method with applications to intrusion detection. Expert Systems with Applications. 2012;39(4):3899–3905.