EAI Endorsed Transactions
On Scalable Information Systems
Research Article EAI.EU

Optimization of Deep Generative Intrusion Detection System for Cloud Computing: Challenges and Scope for Improvements

Nitin Wankhade[1,*], Anand Khandare[2]

[1]Research scholar, Thakur College of Engineering & Technology, Computer Engineering Department, Mumbai, Maharashtra, India.
[2]Professor, Thakur College of Engineering & Technology, Computer Engineering Department, Mumbai, Maharashtra, India.

Abstract

The large amount of data and its exponential increase result in security problems which subsequently cause damage to cloud computing and its environments. The Intrusion detection system (IDS) is among the systems that monitor and analyse data for malicious attacks in the cloud environment. High volume, high redundancy, and high dimensionality of network traffic in cloud computing make it difficult to detect attacks by contemporary techniques. To improve the performance of IDS features selection and data imbalance issues need to be resolved. This paper includes techniques and surveys of cloud-based IDS with ML techniques and IDS performance on the different types of cloud-based datasets. It also analyses the gaps and scope for enhancement of evaluation parameters of IDS. It provides a cloud-based IDS system which will produce a good performance result as compared to the other contemporary system. Moreover, this paper offers a current overview of cloud-based IDS, Data imbalance technique, Dataset and proposed cloud IDS system architecture.

Keywords: Cloud Intrusion detection system, Data Imbalance, Machine Learning, Ensemble methods

Received on 01 July 2023, accepted on 18 September 2023, published on 26 September 2023

doi: 10.4108/eetsis.3993

1. Introduction

With the incremental advancement of the computer network, the cloud computing system plays an important role as a service on the internet. Intrusion detection systems with a cloud-centric approach attract the research community. The IDS is one of the systems that monitors and analyses data for malicious attacks in the cloud environment. The high volume, high redundancy and high dimensionality of network traffic in cloud computing make it cumbersome to detect attacks by existing traditional techniques. The machine learning-based-IDS is one of the approaches in which unknown i.e. novel attacks are detected. The IDS can use a machine learning-based approach [1]- [6] and a non-machine-based approach and approach [7]- [9]. The non-machine-based IDS is highly dependent upon prior information about the signature of a given attack and it is incompetent to detect unknown attacks for which signatures are not available. The proposed system uses machine learning to detect known and unknown attacks. Non-ML-based cloud-based attack detection systems, previous studies propose the snort along with SDN technology for detecting malicious attacks. The signature-based attack detection technique and Snort-based technique are unable to detect unknown attacks. The anomaly detection technique can differentiate the attack traffic from normal traffic with the help of the use of different machine learning classifiers however; it produced a false alarm rate which is very high in amount.

Corresponding author. Email: nitin.wankhade@gmail.com

One of the critical challenges is the lack of malicious samples in the development of cloud-based IDS. In the cloud environment, it is revealed that the out of collected samples the majority of samples belong to the normal class while only a few samples belong to the attacked class [15]. This can be solved in our system by generating the labelled malicious samples with the help of different ML techniques like deep generative models. The proposed system designs an effective algorithm to address the imbalance problem of data in the cloud-based IDS environment.

Currently existing available dataset has a number of features that are challenging. To train the cloud-based IDS, to enhance accuracy, the features in the cloud-based dataset need to be reduced. The proposed system designs a novel approach to feature dimensionality reduction. The existing cloud-based IDS is unable to detect a simultaneous attack, but it is capable of detecting only specific attacks at one time. In the proposed system we are trying to detect simultaneous cloud base attacks like distributed denial of service attack (DDoS), Cache base side channel attack (CSCA), and Structure query language (SQL) injection attack. In DDoS, three types are there, which are ping of death, TCP flood attack, and Slowloris attack. The proposed system designs an efficient algorithm that not only detects the attack but also prevents the same attack.

The dataset used to train the cloud-based IDS is imbalanced in nature [24]- [26]. Our system tries to make the existing dataset balanced by different techniques. The data imbalance can be solved by generating data samples. However, it leads to the loss of original data. The performance of the ML algorithm gets compromised if data samples of generated data show different distributions from the original data of the dataset. The proposed cloud-based IDS system tries to get rid of it. The proposed system will be trained on more than one dataset because the ML classifier trained one type of dataset not suitable for other types of datasets. The proposed system designs an algorithm that will improve the attack detection rate and reduce the rate of false positives. It is doing so by employing a different ML-based technique.

The proposed system's performance will be evaluated by comparing it to the contemporary existing system.

2. Related Works

In this section we will discuss the review of existing cloud-based IDSs, different techniques to handle imbalance challenges of datasets, cloud base attack detection techniques and techniques for feature dimensionality reduction.

2.1. Intrusion Detection Bases on Cloud Scenarios

The intrusion detection system response to network security problems is different from the traditional security system. An IDS monitors the real-time network, and it issues a warning upon attack detection so that the necessary corresponding measures are taken [10]- [12]. Cloud-based Intrusion detection system's main purpose is to identify malicious behaviours on the cloud at an early stage. Attacks like as Distributed denial of service (DDoS), TCP flood, cache side-channel attacks (CSCA) and Structured Query Language (SQL) injections attacks are regarded as serious cloud-based attacks that affect the cloud environment. The DDoS attack affects the service quality and hamper the resources in the network. The sharing of resources in a cloud-based environment among many customers leads to the CSCA attack. Structured Query Language (SQL) injections result in the prevention of unauthorized access to resources of the cloud environment.A number of solutions has been proposed to detect as well as reduce cloud-based attacks.

Wenjuan Wang et al [13] this paper, authors proposed cloud base intrusion detection for the detection of attack and features extractions purpose. This system attempts to gather the network data of the Xen cloud by using SDN technique. The collected traffic is then applied to the model which is the combination of SCAE plus SVM (support vectors machine) for attack detection and feature extractions. It detects attacks like DOS, Probe, R2L, and U2R on the data plane. After reviewing this paper, it is identified that the classifier used in this can be optimized to a further level and the deployed SVM classifier does not detect some existing cloud-based attacks, so optimization and detection of all attacks is found out the scope for this paper.

Sahi et al. [14] in this paper, authors proposed techniques that not only detect but also prevent attacks like as DDoS TCP flood attacks. This technique employs the attributes of data packets to detect the attack. Source and destination internet protocol (IP) address attributes are used to separate malicious users from normal users. In this authors use the CS_DDOS and LS_SVM systems to identify the attack accurately. However, the proposed model does not identify the DDoS attack committed using a spoofed IP address.

Yixuan Wu et al. [15] this paper, authors proposed IoT-based edge computing techniques for attack detection. They propose an intelligent system with fuzzy logic for feature selection. After an overview of this paper, it is found that a robust feature selection method is required to deal with existing datasets as feature extraction in this is not as effective as required. The cache side-channel attack (CSCA) is one of the serious attacks on a network. The CSCA is caused due to the sharing of hardware resources among customers; it has been observed that CSCA attacks are more prominent in the cloud than in a traditional network system.

Chauhan et al. [16] authors proposed a methodology to handle side channel attacks that uses a Bloom Filter-based detection technique. The Bloom Filter reduces the performance overhead to the minimum level. Bloom Filter seems to be effective while predicting the cache behaviours caused by this attack. It also detects SQL injection attacks to prevent unauthorized access to resources [17]. The SQL statements analysis is done to generate a rule tree and then uses the rule tree to detect attacks. However, these methods suffer from data imbalance and do not detect other cloud-based attacks.

Nguyen et al. [18] in this paper, based on the authors, proposed a hybrid model to improve the attack detection rate in the cloud, the model is a combination of Principal Component Analysis with the Restricted Boltzmann Machine. In this model, PCA is used for dimension reduction and RBM is used for feature selection. However, both proposed dimension reduction and feature selection lead to increased computation costs.

Pandeeswari and Kumar [19] in this paper, authors, proposed a model for the detection of cloud base anomaly for the hypervisor layer. This technique is an ensemble of clustering algorithm techniques Fuzzy C-Means along with an Artificial Neural Network (FCM-ANN). The technique Fuzzy C-mean takes the datasets and forms the clusters of it into equal parts. Then Artificial Neural Network is then trained on each cluster of the datasets. The results from the Artificial Neural Network (ANN) models are then aggregated to form the final prediction score.

Dey et al. [20] authors proposed a two-layer attack detection system using machine learning for mobile cloud-based environments which depends upon data fusion techniques. In the first layer, the traffic of the cloud is screened and in the second layer, the decision is made for the detection of the attack. The results reveal that this technique is effective for detecting two important attacks, distributed denial of service (DDoS) attacks and Man-in-the-Middle (MITM) attacks.

Kira et al. [21] authors employed a technique that uses a two-step approach for the detection of DDoS attacks. This technique uses ensemble techniques based on clustering algorithm and filter-based technique. The filter-based methods, use the signature of the DDoS to identify the malicious packets in the mobile cloud environments. To detect the malicious packets a clustering algorithm is used which passes through the filter from the filter-based attack. matching. Two different approaches are employed to detect the DDoS attack. In the first technique, the IP address of both types i.e. static and dynamic of the attacker is blocked completely to protect the further damage from the said attack. The host base IDS techniques are put into practice to detect the pattern-matching attack. In this router and server in the cloud, the base environment is protected by using different IDSs like a host-based intrusion detection system (HIDS), signature base intrusion detection system (SIDS), and network-based intrusion detection system (NIDS.) It has been revealed that in this paper no algorithm is employed to detect the mentioned attack.

Zhang et al [23] authors proposed methods for the detection of cloud-based attacks using an algorithm known as many objective evolutionary algorithms (MaOEA). The MaOEA algorithm technique results in the optimization of the evaluation parameters like accuracy, recall, precision, and false alarm rate and it also reduces the number of features in the datasets which leads to improved time and space complexity. The KNN classifier is used to detect the attack. It has been observed in this paper, that as the classifier (KNN) degrades the performance there is a need to enhance the performance of IDS detection by improving classifier performance.

There is different cloud base network intrusion detection technique that has been deployed to detect the different cloud base attack which is elaborated in Table 1. In Table 1, we have tried to summarize the different techniques, attacks, and techniques employed in cloud-based IDS in recent years.

Table 1. Different Cloud IDS Systems based on cloud

Sr. No	Ref.	Attack Detected	Datasets	Technique Employ	Adv.	Challenges
1	[15]	-Any kind of cyber attack	-NSL-KDD, KDD CUP 99	-Fuzzy algorithm for big data mining -Generative adversarial network(CNN) -Convolution neural network (CNN)	-Fuzzy rough set extracts the low dimensional features -It also maintains feature validity which achieves precise detection of attacks. -Multiple cyber-attacks detected.	-Optimization of features selection method required. -Lightweight and high precision methods required.
		Brute force, DDoS,	CIC-	-Synthetic Over Sampling Technique --Borderline - SMOTE with sampling methods	- Effective for detecting intrusion for edge	- Need to evaluate evaluatingmodel

2	[45]	DoS, Heartbleed, Infiltration, portScan, Web attacks	IDS2017, CSE-CICIDS2018	Random Under Sampling -Bayesian Optimization and Tree Parzen Estimator (BO-TPE) use to optimize the CNN model.,	devices -Having a very low false positive rate with an accuracy of 100%.	ondifferent network datasets for evaluation purposes.
3	[44]	Slowloris, TCP land, ping of Death	NSL-KDD, UNSW-NB15, CTU-13s, Cloud IDS datasets	-Deep neural network -Conditional denoising adversarial autoencoder(CDAAE) -Hybrid of CDAAE with KNN algorithm (CDAAE-KNN) - classifier like DT, RF and SVM.	-Generated malicious data can help the classifier achieve high performance. -Quality of the generated sample is high -The data imbalance problem of IDS datasets is somewhat resolved.	- CDAAE implanted only for data which follow Gaussian distribution not for other distributions. -CDAAE attempts to learn from external information -CDAAE and CDAAE-KNN effectiveness can be enhanced.
4	[13]	DOS, Probe, R2L and U2R	KDD Cup 99 and NSL-KDD.	-Deep learning +Shallow learning techniques. -Stacked contractive AutoEncodersare proposed for feature reduction. -SCAE-SVM combinesapproachesto improve detection performance.	-It extracts essential features by using SCAE. - It improves the attack detection rate as compared to the existing technique.	- Need to improvethe classifier to detect all existing attacks. -The SVM classifier is unable to detect some new attacks like buffer overflow, nmap, and pod. -To evaluate the performance needed to implement in the real cloud environment.
5	[46]	Stealthy malware in the cloud	UNM datasets and BareCloud dataset	- VMI with dynamic analysis. - for features extraction Bag of n-gram use. - Binary particle swarm optimization employ forfetching and discardingrelevant features. - RF for classification of malicious samples and K-fold validations.	- Storage requirements are reduced by VMShield. - Provides the computationally efficient feature selection approach	- Not detect the other kind of attack
6	[14]	DDoS TCP flood attack	Incoming packets	- K-fold cross-validation. - Classifiers like LS-SVM, Naive Bayes, K-Nearest, and	-improve the security of records. - reduce bandwidth -With the Kappa coefficient under single and	- It cannot detect the DDoS attack performed using spoofed IP.

				multilayer perceptron. -CS_DDoS system	multiple attacks accuracy is good. -Important for the security of e-health records during emergencies.	
7	[17]	SQL injections	---	- Input filtering and dynamic taint analysis, Triple tree	- The common attack and attack of replacing code is detected. -Detection accuracy is improved.	---
8	[18]	Cyberattack	KDD Cup 1999, NSL-KDD, UNSW-NB 15	- Deep learning offline mode -Principal component analysis used for dimension reduction -Restricted Boltzmann Machine (RBM) is used for feature selection	- high accuracy in detecting cyber attack -Robust and flexible	-not implanted for real-time device data - Dimension reduction and feature selection methods lead to increased computation costs.
9	[10]	----	MAWIflow	-NSGA-II for feature selection. -RF classifier for classification of malicious samples.	- Suitable for real-time IDS for high-speed network. - This IDS has a high lifespan.	- The lifespan can be improvedby other suitable techniques.
10	[22]	-DDoS -Brute force -Pattern matching	--	-Cloud flare for the cloud server. -Host base intrusion detection system -Network base intrusion detection system -Signature-based intrusion detection system	-Design a network topology where DDoS attacksare carried out inside the cloud server. -Brute force and pattern matching carried out outside the server.	- Thealgorithm requiredfor the cloud which checks the effect of various components of the system.
11	[23]	-DOS, Probe, U2R and R2R	NSL-KDD	-K-Nearest Neighbours (KNN) -Many-objective evolutionary algorithm (MaOEA-ABC)	-The employ techniques reduce the features - It also improves theaccuracy and reduces the false alarm rate.	As the KNN classifier contains a number of parameters it leads to affect the attackdetection performance.

From Table 1 it is clear that the current cloud-based IDS detects different cloud attacks like DOS, DDOs, Probe, U2R, R2R, Slowloris, TCP land, ping of Death, etc. The different classifier is employed to improve the detection rate, precision, recall, receiver operating characteristic, and F-measure parameters. The recent system uses the K-Nearest neighbours (KNN), Support Vector Machines (SVM), Random Forest (RF), Naive Bayes (NB), and deep learning approach. Some IDS systems also employ the ensemble classifier to improve the performance of IDS. Along with advantages the existing system suffers from some challenges like, it is unable to detect all the cloud-based attacks, the performance of the classifier needs to be improved, optimizing feature selection methods, etc. The proposed intrusion detection system will address all the challenges that have been a part of the current existing system.

Intelligent intrusion detection for Internet of things security: a deep convolution generative adversarial network-enabled approach overall attack detection for CSE-CIC 2018 and CIC-DDOS-2019 seen in figure 1 and figure 2. For the CSE-CIC 2018 dataset the accuracy- is 95.22%, precision is 98.22%, recall is 95.42%, F-measure-97.09%, and false alarm rate is 5.78%. For the CIC-DDOS-2019 dataset, the accuracy is 98.62%, precision is 99.60%, recall is 98.98%, F-measure is 99.29% and false alarm rate is 12.84%.

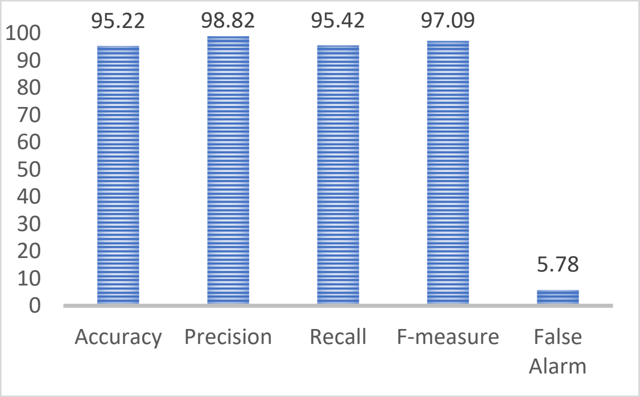

Figure 1. Performance for CSE-CIC 2018 Datasets

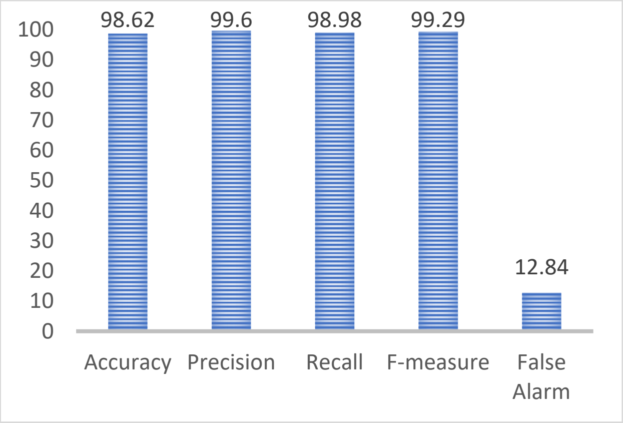

Figure 2. Performance for CIC-DDOS 2019 Datasets

Deep generative learning models for cloud intrusion detection systems and overall attack detection for NSL-KDD datasets are seen in Figure 3. For the NSL-KDD dataset the F1 score- 84.2%, the AUC score- 75.3%, and GEO score—76%.

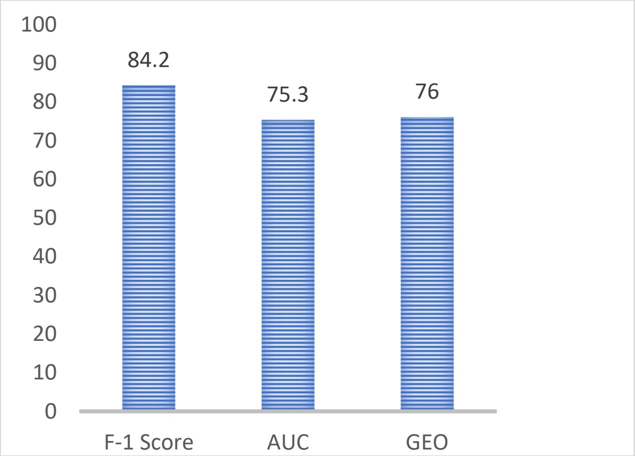

Figure 1. Performance for NSL-KDD Datasets

Cloud Intrusion Detection Method Based on Stacked Contractive Auto-Encoder and Support Vector Machine overall attack detection for NSL-KDD and KDD-CUP 99 dataset can be seen in Figure 4 and Figure 5. For the NSL-KDD dataset the accuracy- is 87.33%, Precision-87.96%, Recall-87.33%, and F-measure-85.01%. For KDD-CUP 99 dataset the accuracy-98.11%, Precision-98.21%, Recall-98.11%, F-measure-98.13%. We consider the overall performance for the 5-class classification.

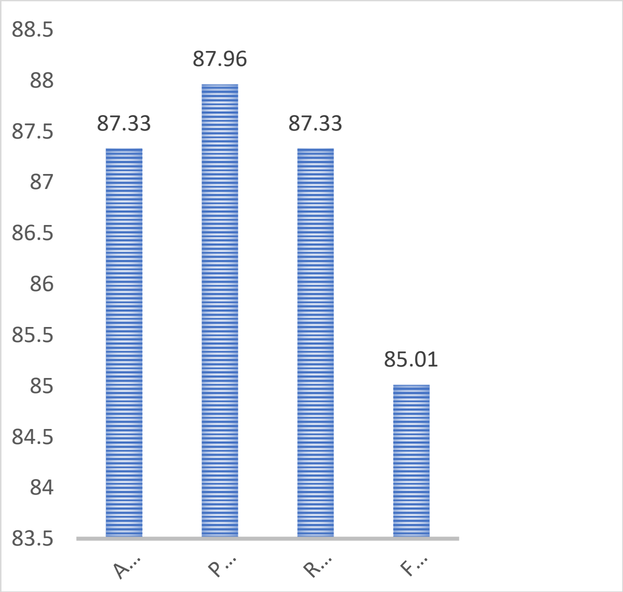

Figure 4. Attack detection performance for NSL-KDD Datasets

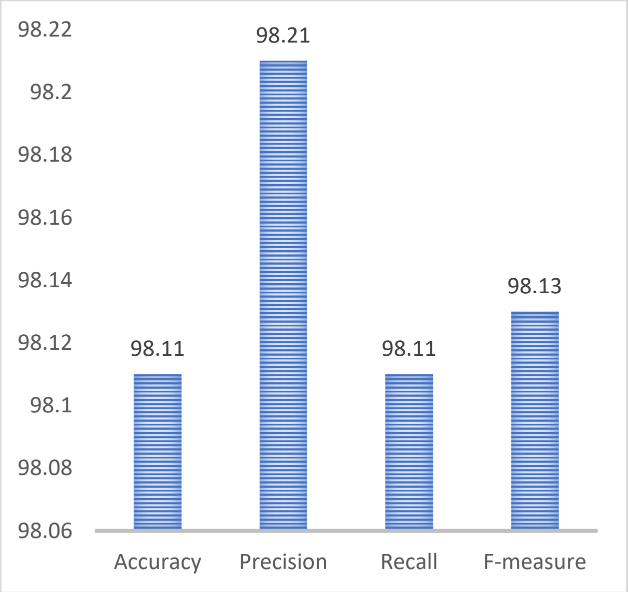

Figure 5. Attack detection performance for KDD-CUP 99 Datasets

2.2. Techniques to Handle the Imbalance of Data Problem

This technique is categorized into three kinds. That categorization depends upon how the technique handles the class imbalance issue. The first one is the Algorithm approach or cost-sensitive learning system that creates or modifies the algorithm that already exists to address the issue of class imbalance in datasets. In the algorithm-level technique, the misclassification cost is assigned to the classes of the datasets. The higher misclassification cost is assigned to the minority class and the lower misclassification cost is assigned to the majority class [24]- [26].

Zhang et al. [24] authors proposed the SVM algorithm to address the data imbalance problem. In which the first imbalance data is divided by using the sampling techniques into smaller numbers of subsets. After this, the cost-sensitive SVM techniques are used to assign the misclassification cost to the minority and majority classes.

Li et al. [25] authors proposed the Adaboost

Algorithms technique for dealing with imbalanced data. Adaboost is an ensemble-based machine learning algorithm; during training of the Adaboost algorithm it put the highest weightage to the minority class as compared to the majority class.

Chung et al. [26] the authors proposed a regression loss function technique approach to address the deep neural network data imbalance problem. Wang et al. [27] and Raj et al. [28] authors employed a loss function that first calculates errors of the minority class and majority class of datasets. Then it calculates the averages of errors together for minority and majority classes.

The second approach to address the data imbalance issue is based on data-level approach. In this technique, the data is first undergone through the data pre-processing stage. This attempts to change the contents of the training data to balance the majority and minority classes. The random under-sampling technique is used to reduce the size of the majority class of datasets while the over-sampling technique is used to enhance the size of the minority class of the imbalance data [29], [30].

The under-sampling technique may remove useful information from the majority class data while oversampling can add the samples which results in overfitting problems [30]. To overcome this limitation, the Synthetic Minority over Sampling Technique (SMOTE) [30] technique is employed which synthesizes the minor class of datasets to enhance the size. The extrapolation and interpolation methods are used to synthesize the minority class from the neighbourhood's ones. The Synthetic Minority over Sampling Technique (SMOTE) is available in different variances to get rid of the issue. 1) Borderline SMOTE [32], 2) SMOTE SVM [33], 3) SMOT -ENN [34] and 4) SMOTE Tomek.

In Borderline SMOTE, the minority class which is only near the borderline is over sample. This technique divides the minority class samples into safe, noisy, and dangerous groups on the basis of the neighbourhood samples. The SMOTE-based ENN technique employs the over-sampling as well as under-sampling techniques of sampling. Here the SMOTE is used for an over-sampling purpose while Edited Nearest Neighbours (ENN) is used for an under-sampling purpose. The SMOTE-based Tomek is a combination of SMOTE and Tomek link under-sampling techniques, here SMOTE is used for an over-sampling purpose while Tomek link is used for under-sampling purposes [34].

The hybrid method is also used to resolve data imbalance issues which is also known as the Ensemble method. This method combines data resampling techniques with an algorithm approach to find out the major and minor class distributions in a dataset to make the necessary decision for the application of resampling techniques. The BalanceCascade (BC) [35] proposed an ensemble of classifier techniques to balance the data in the dataset. It selects the major class from the datasets and under sampling is then performed on it. EasyEnsemble [35] learns various things from the major class of the datasets by using the unsupervised machine learning technique. It creates a different number of data training sets by using a random under-sampling technique. After that, the model is trained on each dataset, and then the bagging technique is employed for the combination of the prediction.

Tomek Link [36] proposed a method that removes samples from the negative class that is close to the positive region in order to return a dataset that presents a better separation between the two classes. The techniques SMOTE-SVM, BalanceCascade, and EasyEnsemble have been in practice but these techniques have lost the original datasets distribution which may lead to the original data loss which may degrade the performance of the machine learning algorithm.

3. Dataset Use

This section describes the datasets we will use in our proposed system. We will use existing different IDS datasets, including one cloud IDS dataset NSL-KDD, and UNSW-NB15 network base datasets. We will also use some malware datasets from the CTU-13 dataset system. These datasets were used in previous research on network IDSs and cloud IDSs. We will also plan to test our system on datasets like UNSW-NB15 and CICIDS-2017 to evaluate the performance metric of our proposed system. The number of data samples in NSL-KDD, and UNSW-NB-15 datasets, and their descriptions are elaborated in Table 2.

Table 2. Number of data samples in Network datasets.

Dataset	Descriptions	Class	Sample Specifications
NSL-KDD	-Four different classes of attack - Total of 39 attack -Total features 41 out of which 4 are categorical, 6 are binary, 23 are Discrete and 10 are Continuous features The training set consists of 125973 and the test set consists of 22544 samples	Normal	67373
		Attack	58630
		DoS	45927
		U2L	52
		R2L	995
		Probing	1156
UNSW-NB-15	-Nine different classes of attack - Total features 49 which consist of The training set consists of 175341 and the test set consists of 82332 samples	Normal	37000
		Attack	45332
		Generic	18871
		Exploits	11132
		Fuzzers	6062
		DoS	4089
		Analysis	677
		Worms	44
		Shellcode	378
		Reconnaissance	3496
		Backdoor	583

One of the Cloud-based environment IDS datasets is used by Kumar et al. [37] using an open-source cloud platform. We will select three types of DOS attacks including Ping-of-Death, TCP Land, and Slowloris attacks from this dataset. NSL-KDD is a network-based IDS dataset [38] which overcomes the issue of the KDD'99 datasets.

The sample of the dataset consists of 41 numbers of features which are classified as attack samples or normal samples. The attack samples are divided into four types of attack denial of service (DOS), user to root(U2R), remote to user (R2L), and Probing attack. Each of these attacks also consists of the attack subtypes. UNSW-NB15 was created in the cyber lab of the Australian Centre of Cyber Security (ACCS). It consists of a total 49 numbers of features and a total nine numbers of attacks like DoS, Fuzzers, Analysis, Backdoors, Exploits, Generic, Reconnaissance, Worms and Shellcode [40]. The number of samples in each class of the cloud base and malware base dataset can be elaborated in Table 3.

The CTU-13 is a malware-based dataset that is available publicly and it was captured at the Czech Technical University (CTU) University, Czech Republic. This dataset consists of normal traffic with a number of botnet traffic. We will choose some scenarios from the dataset that correspond to three kinds of malware, including Men-ti, NSIS.ay, and Virus [40].

Table 3. Number of data samples in Cloud and Malware datasets

Dataset	Class	Samples Specifications
Cloud Base	Slowloris	Normal Samples – 3077 Attack Samples - 293
	TCP Land	Normal Samples - 3072 Attack Samples - 34
	Ping of death	Normal Samples - 3070 Attack Samples - 48
Malware Base	CTU13.6-Menti	Normal Samples - 18904 Malware Samples - 230
	CTU13.12-NSIS.ay	Normal Samples - 92485 Malware Samples -1420
	CTU13.13-Virus	Normal Samples - 37000 Malware Samples - 37000

There are, different datasets used for the IDS However, in this paper, we will be planning to use a few. We will implement the IDS system on the larger datasets so that the performance of the system will be compared with the existing datasets.

4. Proposed System

4.1 Objective

From the preceding literature survey, it is discovered that there is a scope for advancement in cloud-based IDS. The purpose of this paper is twofold. First to make cloud base IDS optimized and efficient by using optimized techniques for data imbalance; and a number of feature selection methods and second purpose is to detect the cloud base attack by reducing the false positive rate and maximizing the detection rate. From the survey, this paper formulated the following objectives which are listed below.

- To design and optimize an algorithm to address the data imbalance problem in cloud-based IDS.
- To design a novel approach for feature dimensionality reduction for cloud-based datasets.
- To design efficient algorithms to detect and prevent different cloud-based attacks.
- To improve the efficiency of the classifier, use for cloud-based IDS
- To design a novel approach to achieve maximum detection rate and minimal false positive rate.
- This is a bulleted list.

Motivated by these challenges of cloud intrusion detection, we propose the cloud-based deep generative IDS system which uses the six stage approach to design the cloud-based IDS. There is, different parameters for the improvements of the cloud base IDS this paper is emphasizing on various aspects of the IDS. These different parameters of improvements with its descriptions can be seen in Table 4.

Table 4. Parameters for improvement in cloud base IDS

Sr. No.	Parameters for Improvement	Characteristic of Parameter
1	Dataset classification	NSL-KDD, UNSW-NB15, CTU-13 dataset are classified into attack and normal samples.
2	Malicious data sample generation	The attack samples are generated to solve the data imbalance problems.
3	Merging of data samples	The normal samples and generated attack samples merge together to form a new datasets.
4	Features Analysis (Selection)	The particulars features from the new datasets are select to optimize the number of features.
5	Attack detection	The different cloud base attack are detected
6	Performance evaluations	The performance is evaluated against different evaluation metrics

4.2 Guidelines

To enhance the performance of cloud-based IDS, this paper is focused on various aspects such as classification of the original datasets, malicious sample generation, merging of data samples, Feature selection, attack detection, and performance evaluations. The guideline for the proposed methodology is given as follows.

(i) Dataset classification: The datasets are divided into normal datasets and attack datasets. The datasets are divided to know the number of samples belonging to the attack and normal samples. We will use the different network-based datasets like NSL-KDD, and UNSW-NB15, and malware-based datasets like CTU13s and different cloud datasets for experiment purposes. The purpose of using different datasets is that it makes the proposed system robust.

(ii) Malicious data samples generation: To generate the malicious sample, we will propose models to artificially generate the malicious samples in the cloud systems environments. The Adversarial Auto Encoder (AE) is used to artificially generate specific types of attack samples to get rid of data imbalance issues. Currently, the existing system generates malicious samples for the data that follows the Gaussian distribution. Our system will generate malicious samples for other types of distribution of data. For this purpose, we will use the unsupervised machine learning AutoEncoders technique.

(iii) Merging of data samples: The normal datasets and augmented malicious datasets combine together to form a dataset called augmented datasets which represents the single point of reference. We will use the MapReduce algorithm which allows us to combine both datasets then we will implement Bayesian Network and K2 algorithm to analyze our dataset. After this, the augmented datasets will be treated as the original datasets.

(iv) Data Pre-processing: We will propose a data pre-processing in which the features that are in character in nature are converted to numerical in nature and then numerical normalizations are performed on it to attain the normalization of data.

(v) Features selection: We will propose a feature extraction technique that employs the fuzzy rough set for the specific feature extraction from the datasets. The fuzzy rough set approach reduces the redundancy of the features present in datasets and selects the best features subset to train the models. It extracts the low dimensional features from original datasets.

(vi) Data Separation: The given datasets are separated into the training data and test data to train and test the models. Out of the training dataset and test dataset, some are selected randomly for training and testing purposes. The test and training data have different probability distributions as per the data separations. To make the intrusion detection more robust and realistic the test data have some types of attack that are not part of the training data.

(vii) Attack detection: Once the data imbalance problems are addressed by using the above-mentioned technique, we will use the different ensemble techniques of a classifier to detect different cloud bases attacks like as DOS, distributed denial of service attack (DDoS), Cache base side channel attack (CSCA), Structure query language (SQL) injection and TCP flood attack. The existing system only detects one cloud attack at a time; our proposed system will detect a simultaneous attack. The different classifiers that we will use to train the models are random forest (RF), decision tree(DT), support vector machine (SVM) and K-nearest neighbours (KNN)

(viii) Performance evaluations: We will use the different performance metrics to measure the accuracy and effectiveness of the proposed system. The metrics include F1 score, AUC score, Precision, Recall and false positive rate. Finally, we want to design a system that produces a high detection rate and minimizes the false positive rate.

4.3 Proposed System Architecture

The proposed architecture of the cloud-based IDS is shown in the figure. In the proposed system we are trying to detect simultaneous cloud base attacks like distributed denial of service attacks (DDoS), Cache base side-channel attacks (CSCA), and Structure query language (SQL) injection attacks. In DDoS, three types are there, which are ping of death, TCP flood attack and Slowloris attack. The proposed system designs an efficient algorithm that not only detects the attack but also prevents the same attack.

In the first stage, the datasets are divided into normal datasets and attack datasets. The input datasets consist of the attack as well as normal data samples. In the second stage, the attack datasets are augmented to get extra malicious samples, leading to the balancing of datasets. In the third stage, the normal datasets and augmented malicious datasets combine together to form a dataset called augmented datasets. Now this dataset is treated as the input for further processing which consists of a somewhat balance of malicious and normal data samples.

In the fourth stage, datasets are divided into a number of small datasets where a feature extraction is performed we will propose a fuzzy rough set for the feature extraction mechanism. The fuzzy rough set-based approach is used to extract low-dimensional features from original data sets It does so by reducing the number of redundant features in data sets it also selects the best feature subset. The feature extraction reduced the dimensionality of the datasets, consequently improving the time and space complexity.

In the fifth stage, the subsets of these datasets give to the ensemble of classifiers which combine the result of all classifier and in the sixth stage Once the data imbalance problems are addressed by using the abovementioned technique we will use the different ensemble technique of classifier to detect different cloud bases attack like as DOS, distributed denial of service attack (DDoS), Cache base side channel attack (CSCA), Structure query language (SQL) injection and TCP flood attack. The existing system only detects a one cloud attack at a time, our proposed system will detect simultaneous attack. The different classifiers that we will use are support vector machine (SVM), Decision tree (DT), Random forest (RF), etc.

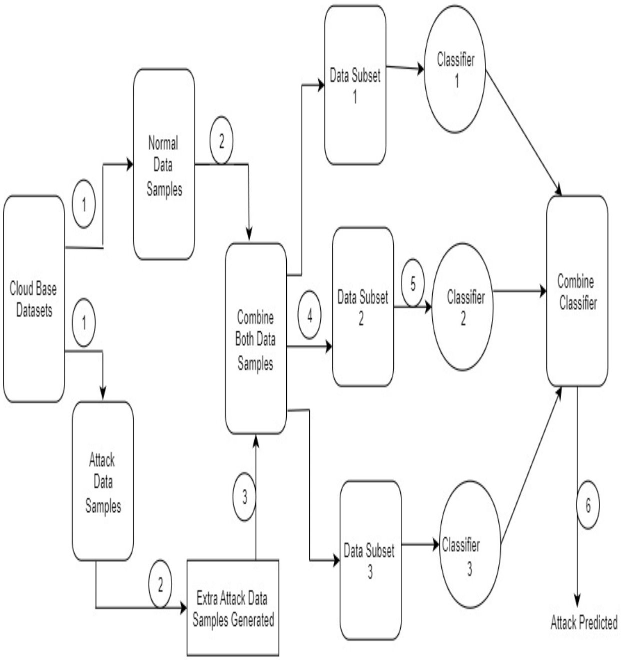

Figure 7. Showing Architecture of Proposed Cloud Base IDS

As shown in Fig. 7 in the proposed cloud-based IDS architecture there are different stages through which optimization and improvements can be achieved by employing the different techniques at each stage.

5. Performance metrics

As our datasets are imbalanced in nature, we use important performance metrics to measure the effectiveness and evaluate the proposed models. The performance metrics we planning to use are the F1 score, AUC score, and GEO score. These metrics are calculated by calculating the recall, precision, FPR and specificity of the proposed model. The Precision (Confidence) (Equ. (1)) measures the ability of a classifier that predict positive. The Recall (Sensitivity) (Equ. (2)) measures the machine learning model's ability to detect the number of actual positive observations. The false positive rate (Equ. (3)) is the number of real negative samples which are predicted incorrectly. While the specificity measures the rate of real negative sample cases that are correctly predicted.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive\ + False\ Positive} \qquad (1)$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$\text{False Positive Rate} = \frac{False\ Positive}{True\ Negative + False\ Positive} \quad (3)$$

$$\text{Specifity} = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (4)$$

The precision and Recall scores result into some drawbacks for that purpose the F1-score is used. F1 score (Equ. (5)) is the mean (harmonic) of Precision and Recall.

$$\text{F1Score} = 2 \times \frac{Precision\ \times Recall}{Precesion + Recall} \quad (5)$$

AUC stands for Area under Receiver Operating Characteristics (ROC) Curve, which plots graph between the True Positive Rate (TPR) and the False Positive Rate (FPR) for threshold values. The FRP and TPR values can vary as per the performance of the classifier. The values of its decide the performance of classifier. For e.g. If scores of FPR = 0 and TPR = 100%, then classifier is termed as the best classifier. While if score FPR = 100% and TPR = 0 then classifier is termed as the worst classifier.

The space which is under the ROC curve is detonated as an area under curve (AUC) score which attempts to counts the quality of the classification model with respect to different thresholds. The values of AUC are between 0.5 and 1.0. The random classifier has an AUC value of 0:5 while value of a perfect classifier is 1:0.

Geometric mean (GEO) is represented as the root of the product of class wise recall. The GEO attempts keep the accuracy balanced. The GEO for binary classification is the squared root of the product of the Recall and Specificity (Equ. (6)). The best value of GEO for a classifier is 1 and the worst value is 0.

$$\text{Geometric Mean} = \sqrt{Sensitivity} \times \sqrt{Specificity} \quad (6)$$

6. Result and Discussion

To demonstrate the working and problems of deep generative methods in cloud computing. This paper compares the existing classifier performance on augmented datasets for cloud-based datasets and network-based datasets. Here the performance of the different classifiers such as Support vector machine, Decision tree and Random Forest is taken into account against the performance metrics like F1, AUC and GUC score.

Table 5. F1, AUC and GEO score of various classifiers on network base IDS datasets

Algorithm	Augmented datasets	NSL-KDD			UNSW-NB15		
		F1	AUC	GEO	F1	AUC	GEO
SVM	ORIGINAL	0.772	0.570	0	0.413	0.129	0
	SMOTE-SVM[10]	0.788	0.688	0.585	0.542	0.218	0.148
	BalanceCascade[11]	0.797	0.620	0.615	0.590	0.304	0.220
	ACGAN[12]	0.821	0.712	0.657	0.602	0.200	0.236
	CVAE[13]	0.822	0.736	0.690	0.634	0.325	0.298
	CAAE[9]	0.828	0.738	0.700	0.642	0.345	0.300
	CDAAE	0.834	0.741	0.748	0.692	0.416	0.406
	CDAAE-KNN	0.842	0.753	0.760	0.722	0.441	0.514
DT	ORIGINAL	0.821	0.430	0	0.426	0.221	0
	SMOTE-SVM[31]	0.824	0.446	0.361	0.461	0.348	0.228
	BalanceCascade[35]	0.798	0.522	0.476	0.497	0.486	0.319
	ACGAN[41]	0.828	0.523	0.588	0.439	0.506	0.420
	CVAE[42]	0.835	0.612	0.656	0.462	0.538	0.460
	CAAE[43]	0.832	0.629	0.652	0.500	0.542	0.470
	CDAAE [44]	0.844	0.650	0.669	0.533	0.522	0.503
	CDAAE-KNN[44]	0.853	0.660	0.672	0.542	0.533	0.533
RF	ORIGINAL	0.742	0.760	0.442	0.413	0.357	0
	SMOTE-SVM[31]	0.749	0.780	0.564	0.709	0.436	0.407
	BalanceCascade[35]	0.752	0.793	0.589	0.702	0.439	0.415
	ACGAN[41]	0.754	0.804	0.628	0.672	0.448	0.519
	CVAE[42]	0.762	0.824	0.692	0.694	0.572	0.594
	CAAE[43]	0.764	0.823	0.708	0.692	0.571	0.602
	CDVAE [44]	0.791	0.835	0.700	0.722	0.602	0.619
	CDVAE-KNN[44]	0.886	0.842	0.715	0.732	0.623	0.633

We are proposing the cloud based IDS deep generative model for the distributions of data other than Gaussians 'distributions of data, which will enhance the F1, AUC and GUC score of the classifier compared to the existing one. Table 5 and table 6 shows the SVM, DT and RF classifier results on augmented datasets for network and cloud based datasets.

Conclusion

As more computing has migrated to the cloud, cloud IDS has become a more important part of research amongst researchers. To protect cloud-based infrastructure and Internet organizations all over the world, spending on cloud security and privacy. This paper survey the contemporary state of cloud based intrusion detection systems, discussing the different techniques employed for cloud IDS to achieve a high detection rate and to minimize the false positive rate. The various cloud-based attacks and their solutions are discussed. The various cloud-based systems are compared to getting a detailed understanding of cloud-based IDS. This IDS has various challenges, like data imbalance. This paper discusses the overview of techniques to address the data imbalance problem. The cloud base and network base datasets are elaborated to use for our proposed system. It also discusses the overview of the proposed architecture of cloud-based IDS.

References

[1] Gao Jun, and Gan Luyun Omni. SCADA intrusion detection using deep learning algorithms. IEEE Internet Things. 2021;8(2): 951–961.

[2] Marteau F P. Random partitioning forest for point-wise and collective anomaly detection application to network intrusion detection. IEEE Trans. Inf. Forensics Security. 2021;16: 2157-2172.

[3] Zhou X, Liang W, Shimizu S, Ma J, and Jin Q Siamese. neural network based few-shot learning for anomaly detection in industrial cyber-physical systems. IEEE Trans. Ind. In form. 2021; 17(8): 5790-5798.

[4] Xu X, Li J, Yang Y, and F. Shen.Toward effective intrusion detection using log-cosh conditional variational autoencoder. IEEE Internet Things Journal. 2021; 8(8): 6187-6196.

[5] Shafiq M, Tian Z, Bashir K A, Du X, and Guizani M. CorrAUC: A malicious Bot-IoT traffic detection method in IoT network using machine-learning techniques. IEEE Internet of Things. 2021; 8(5): 3242-3254.

[6] Injadat M, Moubayed A, Nassif B A, and Shami. Multi-stage optimized machine learning framework for network intrusion detection. IEEE Trans. Netw. Service Manag. 2021;18(2): 1803-1816.

[7] Shin S and Gu, G Cloud Watcher. Network security monitoring using Open Flow in dynamic cloud networks (or: How to provide security monitoring as a service in clouds?). Proc. IEEE Int. Conf. Netw. Protoc. 2012; 1-6.

[8] Chung J. C., Khatkar P., Xing T., Lee J., and Huang D. NICE: Network intrusion detection and countermeasure selection in virtual network systems. IEEE Trans. Depend. Secure Computer. 2013;10(4): 198-211.

[9] Xing T, Xiong Z, Huang D, and Medhi D. SDNIPS: Enabling software-defined networking-based intrusion prevention system in clouds. in Proc. Int. Conf. Netw. Serv. Manage. Workshop. 2014; 308-311.

[10] Viegas E, Santin O A and Abreu V. Machine learning intrusion detection in big dataera: A multi-objective approach for longer model lifespans. IEEE Trans. Netw. Sci. Eng. 2021;8(1): 366-376.

[11] Ning Z. Block chain-enabled intelligent transportation systems A distributed crowdsensing framework. IEEE Trans. Mobile Computing.2021; 21(12): 4201-4217.

[12] Ning Z and Shouming Sun. Intelligent resource allocation in mobile blockchain for privacy and security transactions: A deep reinforcement learning based approach. Sci. China Inf. Sci.2021; 64: 162303.

[13] Du Wenjuan Wang Xuehui, Shan Dibin, Qin Ruoxi, and Wang Na.Cloud Intrusion Detection Method Based on Stacked Contractive Auto-Encoder and Support Vector Machine. IEEE Transaction On Cloud Computing. 2022;10(3): 1634-1646.

[14] Khader S. A., Lai D., Li Y, and Diykh M. An efficient DDoS TCP flood attack detection and prevention system in a cloud environment. IEEE Access. 2017;5: 6036–6048.

[15] Wu Y, Nie L, Wang S, Ning Z and Li S. Intelligent intrusion detection for Internet of things security: a deep convolutional generative adversarial network-enabled approach. IEEE Tran. IEEE Internet of Things Journal. 2023;10(4): 3094-3106.

[16] Chauhan M and Hasbullah H. Adaptive detection technique for cache-based side channel attack using bloom filter for secure cloud. In:3rd International Conference on Computer and Information Sciences (ICCOINS). 2016.293–297.

[17] Wang K and Hou Y. Detection method of SQL injection attack in the cloud computing environment. In: IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC).2016. 487–493.

[18] Nguyen K K, Hoang T D, Niyato D, Wang P, Nguyen N D, and Dutkiewicz E. Cyberattack detection in mobile cloud computing: A deep learning approach. In: IEEE Wireless Communications and Networking Conference, WCNC. 2018. 1–6

[19] Pandeeswari Nand Kumar G. Anomaly detection system in cloud environment using fuzzy clustering based ANN. Mobile Networks and Applications. 2016;21(3): 494- 505.

[20] Dey S, Qiang Y, and Srinivas S. A machine learning-based intrusion detection scheme for data fusion in mobile clouds involving heterogeneous client networks. Information Fusion. 2019;49: 205-215.

[21] Kiranmai B and Damodaram A and Extenuate. DDoS attacks in the cloud. In: 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). 2016. 235–238.

[22] Nadeem M, Arshad A, Riaz S, Band S S and Mosavi 2021 A: Intercept the Cloud Network from Brute Force and DDoS Attacks via Intrusion Detection and Prevention System. In IEEE Access. 2021; 9: 152300-152309.

[23] Zhang Z, Wen J, Zhang J, Cai X and Xie L A Many Objective-Based Feature Selection Model for Anomaly Detection in Cloud Environment. IEEE Access. 2020; 8:60218-60231.

[24] Zhang Y and Wang D. A cost-sensitive ensemble method for class-imbalanced datasets. Abstract and Applied Analysis. 2013;215-225.

[25] Kong X Li, Lu Z, Wenyin L, and Yin J. Boosting weighted ELM for imbalanced learning. Neurocomputing. 2014;128: 15–21.

[26] Pozzolo D A, Caelen O, Waterschoot S, and Bontempi G. Cost awarepertaining for multiclass cost-sensitive deep learning. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016.1411–1417.

[27] Wang S, Liu W, Wu J, Cao L, Meng Q, and Kennedy J P. Training deep neural networks on imbalanced data sets. In;IEEE International Joint Conference on Neural Networks, IJCNN. 2016.4368–4374.

[28] Raj V, Magg S, and Wermter S. Towards the effective classification of imbalanced data with convolutional neural networks. In: Artificial Neural Networks in Pattern Recognition - 7th IAPR TC3Workshop. 2016.150–162.

[29] Pozzolo D A, Caelen O, Waterschoot S, and Bontempi G. Racing for unbalanced methodsselection. In: Proceedings of Intelligent Data Engineering and Automated Learning – IDEAL 2013 - 14th International Conference. 2013. 8206: 24–31.

[30] Drummond C and Holte C R. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. In: Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets. 2003.1–8.

[31] Chawla V N, Bowyer W K, and Hall O L. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16: 321–357.

[32] Y W Han Wang and Mao B H. Borderline-SMOTE: A new oversampling method in imbalanced data sets learning. In: Proc. Int. Conf.Intell. Computer. 2005.878-887.

[33] Nguyen M H, Cooper W E and Kamei K. 2011 Borderline oversampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms. 2011;3(1): 4–21.

[34] Batista E. G., Prati C. R., and Monard C. M. A study of the behaviours of several methods for balancing machine learning training data. ACMSIGKDD Explore. 2004; 6(1): 20-29

[35] Liu X BC, Wu J and Zhou Z. Exploratory under sampling for class imbalance learning. IEEE Transaction Systems, Man, and Cybernetics. 2009; 39(2):539–550.

[36] Namvar A Siami M and Rabhi F. Credit risk prediction in an imbalanced social lending environment. International Journal of Computational Intelligence Systems. 2018; 11(1):925–935.

[37] 37. Kumar R, Lal P S, and Sharma. Detecting denial of service attacks in the cloud. In: 2016 IEEE 14th International Conference on Dependable, Autonomic and Secure Computing. 2016.309–316.

[38] Nsl-kdd dataset [online], http://nsl.cs.unb.ca/NSL-KDD/, accessed:2018-04-10.

[39] Moustafa N and Slay J. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In: Military Communications and Information Systems Conference (MilCIS). IEEE. 2015. 1–6.

[40] 40. M S. Garc´ıa Grill and Stiborek J. An empirical comparison of botnet detection methods. Computers & Security.2014; 45:100–123

[41] Odena A, Olah C, and Shlens J. Conditional image synthesis with auxiliary classifier gains. In: Proceedings of the 34th International Conference on Machine Learning, ICML. 2017.2642–2651.

[42] Sohn K, Lee H and Yan X. Learning structured output representation using deep conditional generative models. Advances in Neural Information Processing Systems. 2015;1: 3483– 3491.

[43] A. Makhzani, J. Shlens, N. Jaitly. Ad- adversarial AutoEncoders," arXiv preprint arXiv: 2015.1511.05644.

[44] Ly Vu, Nguyen Uy Quang, Nguyen N Diep, Hoang Thai Dinh and Dutkiewicz Ery. Deep generative learning models for cloud intrusion detection systems. IEEE Transactions on Cybernetics. 2022; 53(1): 565-577.

[45] Okey O D, Melgarejo D C and Saadi M. Transfer Learning Approach to IDS on Cloud IoT Devices Using Optimized CNN. IEEE Access. 2023;11: 1023-1038.

[46] Mishra P, AggarwalPalak, Vidyarthi Ankit, Singh Pawan, Khan Baseem, Alhelou Hassan Haes et al. VMShield: Memory Introspection-Based Malware Detection to Secure Cloud-Based Services Against Stealthy Attacks. IEEE Transactions on Industrial Informatics. 2021;17(10): 6754-6764.