BIG DATA SECURITY ENHANCEMENT BASED INTRUSION DETECTION SYSTEM USING K-MEAN CLUSTERING OF DECOMPOSITED FEATURES

Virendra Kumar Swarnkar, Dr Asha Ambhaikar and Suman Kumar Swarnkar
swarnkarvirendra@gmail.com, dr.asha.ambhaikar@gmail.com, sumanswarnkar17@gmail.com
[1,3]*Department of Computer Science and engineering, Kalinga University, Naya Raipur, India*
[2]*Department of Computer Science and IT, Kalinga University, New Raipur, Chhattisgarh India*.

Abstract

The protection of networked information systems is a critical problem impacting individuals, companies and governments. The number of attacks against networked networks has risen significantly and the methods used by the attacker are continuing to develop. Intrusion prevention is one method to avoid these threats from happening. A popular approach to creating an IDS system is by machine learning. The efficiency of the IDS is currently increased when discriminative and representative features are taken. AE and PCA are used to minimise dimensionality of features (PCA). The attribute extraction techniques employed are then used to construct an RF classification technique with K-Mean Cluster. This research effort will reduce the features of dataset "CICIDs" from 79 to 45, while retaining a high accuracy of 99.7% in Random Forest classifier with k-means clustering.

Index Terms— Dimensionality Reduction, Intrusion Detection System (IDS), Sparse Auto Encoder (SAE), Principle Component Analysis (PCA), K-Mean Clustering, Random Forest.

I. Index Terms— Dimensionality Reduction, Intrusion Detection System (IDS), Sparse Auto Encoder (SAE), Principle Component Analysis (PCA), K-Mean Clustering, Random Forest.Introduction

A network intrusion detection system (NIDS) programme (or hardware) is used to identify malicious activity in the network (or system) It is categorised into anomaly-based and signature-based intrusion detection. IDS engineers can use different identification strategies. Machine learning is one of these methods that is being used for computational policing. Machine learning (ML) techniques can forecast and avoid possible security incidents (SI). Binary classification involves classifying items into two groups. Multi-class grouping involves categorising instances into more than two categories. We use both classifications in this study. There are 15 classes for the multi-class grouping, where each class represents either regular network flows or a one of 14 types of attacks. In the binary grouping, packets are being categorised as either normal or irregular packets.

This is an artificial neural network of entangled artificial neurons. Various types of ANNs include Deep Convolutional Networks (DCNN), Recurrent Neural Networks (RNN), and Auto-Encoder (AE) neural networks. Deep learning is a sophisticated machine learning approach that has the ability to allow outstanding security systems.

With the high-dimensional features of machine learning problems, the classification process takes a long time. These features may minimise these processes in some situations. Classification of network traffic data with different class distributions may adversely affect the efficiency of classical classification algorithm. The frequency and number of imbalanced class distributions indicates the importance of further study. Previous reviews of intrusion detection systems have not dealt with type of database records with distorted class distributions. Adopting balanced data will increase the measure of the classifier.

Key features of this paper includes the development of a system for machine learning-based network intrusion detection. The implementation of Artificial Neural Network, Random Forest Classifier with K-Mean Clustering for anomaly detection is applied. The key takeaways from this paper are as follows:

1. We also accomplished successful analysis of characteristics using Artificial Intelligence (AI) and Principal Component Analysis (PCA).
2. By using the CICIDS2017 dataset we can compare the utility of dimensionality reduction methods with Random Forest and K-Mean Clustering.

Related work

Dimensionality Reduction Approaches Selection Criteria

This section attempts to review the current applicable literature that refers this issue. The criteria for selection are:

1. It is linked to the CICIDS2017 dataset.
2. The work discussed is highly applicable to dimensionality reduction approaches, so specifically, Auto encoder and PCA.
3. Being important to K-clustering of aggregate feature.
4. This paper is applicable to machine learning-based intrusion detection.

Dimensionality reduction is used for many different purposes including: minimising computing overhead, reducing noise in the data, and for improved data analysis and analysis.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

One common way to minimise dimensionality of data is MVR. This method is useful when the number of data points is significant. The number of missing data in the CICIDS2017 is still limited. Therefore, we excluded the Missing Value Ratio approach. Other techniques include Forward Feature Creation (FFC) and Backward Feature Exclusion (BFE) methods. Both FFC and BFE are too late on big-scale datasets. As a result, we did not share these methods. PCA, on the other hand, is relatively computationally cost effective, can cope with massive databases, and has been commonly used in the past. PCA, is a type of dimensionality reduction technique for feature extraction. Another big difference is that in the AE there is no expectation of linearity in the results. The auto-encoder optimizer figures out the function by the set of parameters of the weights that least encrypt the results, provided the reconstruction error. This is because the PCA is handling reduced data linearly. Moreover, the numerical complexity of an algorithm depends on the dimensionality of the data and the number of weights in the auto-encoder.

A. CICIDS2017

Sharafaldin et al. [4] used Random Forest regression to support the determination on the most necessary features to detect Ebola virus. The authors conducted feature selection by using various models that included K Nearest Neighbor (KNN), AdaBoost, Multi-Layer Perceptron (MLP), Naïve Bayes, Random Forest (RF), Iterative Dichotomiser 3 (ID3) and Quadratic Discriminant Analysis (QDA). The consistency of test results was 99.8%. The model production time was 74.39 seconds. It takes 21.52 seconds for our suggested approach by Random Forest to identify photos to be right. Our suggested systems can detect suspicious intrusions that come from several intrusion families.

Some researchers have used the genetic algorithm (GA) as a feature selection tool and several Support Vector Machines (SVM) for classification in wireless mesh systems. They used a mix of SVM classifiers to tackle new threats. Each classifier was learned to detect a specific attack from the training data by using selected features. Part of the CICIDS2017 dataset is accessed to test their method. In this article, we use all instances of the CICIDS2017 dataset.

Authors in [11] compared and contrasted frequency-based algorithms with an aggregation-based method. Therefore, it concluded that the frequency-based paradigm is preferable to the convolutional LSTM.

Researchers in [12] tested the conference dataset with the aid of digital wavelets. Their approach guarantees the detection and avoidance of denial-of-service attacks of both Slow Loris and HTTP Denial of Service (DoS).

Further, the authors of [13] implemented Multi-Layer Perceptron classifier algorithm and Convolutional Neural Network classifier that use the records from CICIDS2017 to train the deep learning systems. Researchers performed the analysis based on the chosen characteristics of network packet header. We computed the profiles and named inputs for computer

and deep learning purposes.

According to [1], the classification system didn't function. In this way, it showed the ability to identify network threats with an average correct positive rating of 94.5% and an average correct negative rate of 4.68%.

In [14], the authors suggested a Denial of Service (DoS) intrusion detection method using Fisher Score algorithm and the Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Decision Tree (DT) as the learning classifier. Their IDS was averaged at 99.7%, KNN at 57.76% and DT at 99%. Our study proposes an intrusion detection device that can detect all sorts of attacks used by CICIDS2017 competitions and the uncertainty matrix indicates an accuracy of 100% for DDoS attacks using (PC A − RF) Mc−10 with UDBB.

In [15], the dimensionality reduction strategy used is Deep Belief Network. The features that were created were then fed to the multiclass SVM. In 2013, it was successfully used in a real time cluster computing application and can be extended to big data processing [16]. Their plan obtained an F-measure of 0.921.

The authors in [17] suggested a data reduction approach for network intrusion detection called Data Dimensionality Reduction (DDR). They used XGBoost, SVM, CTree and Nnet classifiers. Our model used 36 features and achieved accuracy of 98.93 percent. In comparison, there is zero traffic of weekends on the Monday. Our work achieved a perfect accuracy of 99.6 percent with 10 features. We decided to retain the original scale of the dataset.

B. Auto-Encoder

The researchers of [20] find a method for attribute collection using unsupervised approach. The reverse inference condition was fixed using a softmax activation function.

An intrusion detection system is using Deep Auto-Encoder (DAE) a basic classifier to detect the type of attack. They used across-entropy law and back-propagation protocol to fix the weights [21].

The SAE is used for function learning and dimensionality reduction in this analysis. Writers of this paper used Support Vector Machines (SVM) and achieved 84.96 percent and 99.39 percent classification of five classes. We used CICIDS2017 to uncover the most widely used intrusion attacks and techniques. There are at least 125,923 instances of NSL-KDD in education, and at least 22,544 in empirical science. There are two.8 million network instances of CICIDS2017 which are registered based on actual traffic.

As mentioned in [22], a new approach focusing on the current technique is introduced in [25]. The authors in [25] implemented SVM on the NSL-KDD dataset. [25] reached a precision of 88.39% for binary classifiers and 79.10% for five classifiers. An auto-encoding network intrusion detection framework was introduced in [26] by authors. The system strengthens the standard clustering (or classification)

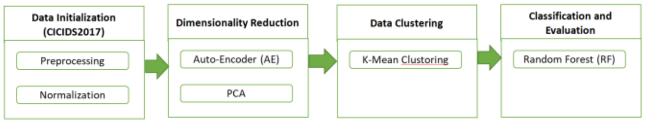


Fig. 1. Proposed Methodology

loss with an auxiliary loss in auto-encoder, thereby offering a more precise analysis. The trials demonstrated the statistical supremacy of our proposed system.

C. *PCA Related Work*

A hybrid feature selection/classification based on machine learning methodology was introduced by the authors in [27].

In the article [27], PCA for feature reduction and decision tree, and Naive Bayes Classifier are used for KDD-99.

The thesis in [28] describes a reduction threshold for intrusion detection. The authors conducted the experiments on KDD-CUP and UNB-ISCX.

II. METHODOLOGY

This thesis acquainted AE and PCA on dimensionality reduction. To check feature reduction model concepts, the paper used the up-to-date CICIDS2017 intrusion detection and prevention dataset [4]. Each file reflects a particular form of attack over a given period of time. The dataset was only compiled over five days, Sunday through Friday. The data flow on Monday contained benevolent network traffic, and some attacks that have been introduced happened in Tuesday, Wednesday, and Thursday and Friday. In this article, the data sets were grouped and stemmed in a compact and lower dimensional form. The idea of Figure 1 is to explain the suggested structure.

A. *CICIDS2017 Dataset*

The CICIDS 2017 dataset contains background traffic of networks created from 25 users' abstraction. The users were identified by the requirements of protocols such as HTTP, HTTPS, FTP, SSH and email protocols The quality indicators to assess the quality of the e-portfolio are:

1. The packet size allocation.
2. The number of pills per hour.
3. The weight of the payload.
4. The order delivery protocols.
5. Moreover.

The attacks covered by CICIDS2017 reflect the regular attack families. The attacks include Brute Force, Heart Bleed, and Botnet, DoS, Distributed DoS, Network Assault, and Infiltration Attack. The dataset is freely accessible by the developers in two formats: 1. The full packet payloads in Packet Capturing (PCAP) format 2. The related profiles and named flows as CSV files for computer and deep learning purposes. There are 2,830,108 documents in this dataset. The friendly traffic is 2,358,036 records (83.3 percent), while the malicious is 471,454 records (16.7 percent) (16.7 percent of the data). It is a special dataset since it lists recent attack patterns. Whereas other datasets such as UNSW-NB15 [30], AWID [32], GPRS [33], and CIDD-001 [34] have features, LUCC has distinct features compared to the other datasets. As a result, CICIDS2017 was chosen as the "bench mark" to test the suggested proposals. The table 2 displays the attacks in the CSV files used in this analysis. CICIDS is a website of 28,307,143 entries, 78 attributes and 1 name.

Table 1 Listed Data Column of network traffic in CICIDS2017

Sn	Column	Dtype	Sn	Column	Dtype
1	Destination Port	int64	41	Packet Length Mean	float64
2	Flow Duration	int64	42	Packet Length Std	float64
3	Total Fwd Packets	int64	43	Packet Length Variance	float64
4	Total Backward Packets	int64	44	FIN Flag Count	int64
5	Total Length of Fwd Packets	int64	45	SYN Flag Count	int64
6	Total Length of Bwd Packets	int64	46	RST Flag Count	int64
7	Fwd Packet Length Max	int64	47	PSH Flag Count	int64
8	Fwd Packet Length Min	int64	48	ACK Flag Count	int64
9	Fwd Packet Length Mean	float64	49	URG Flag Count	int64
10	Fwd Packet Length Std	float64	50	CWE Flag Count	int64
11	Bwd Packet Length Max	int64	51	ECE Flag Count	int64
12	Bwd Packet Length Min	int64	52	Down/Up Ratio	int64
13	Bwd Packet Length Mean	float64	53	Average Packet Size	float64
14	Bwd Packet Length Std	float64	54	Avg Fwd Segment Size	float64
15	Flow Bytes/s	float64	55	Avg Bwd Segment Size	float64
16	Flow Packets/s	float64	56	Fwd Header Length.1	int64
17	Flow IAT Mean	float64	57	Fwd Avg Bytes/Bulk	int64
18	Flow IAT Std	float64	58	Fwd Avg Packets/Bulk	int64
19	Flow IAT Max	int64	59	Fwd Avg Bulk Rate	int64
20	Flow IAT Min	int64	60	Bwd Avg Bytes/Bulk	int64
21	Fwd IAT Total	int64	61	Bwd Avg Packets/Bulk	int64
22	Fwd IAT Mean	float64	62	Bwd Avg Bulk Rate	int64
23	Fwd IAT Std	float64	63	Subflow Fwd Packets	int64
24	Fwd IAT Max	int64	64	Subflow Fwd Bytes	int64
25	Fwd IAT Min	int64	65	Subflow Bwd Packets	int64
26	Bwd IAT Total	int64	66	Subflow Bwd Bytes	int64
27	Bwd IAT Mean	float64	67	Init_Win_bytes_forward	int64
28	Bwd IAT Std	float64	68	Init_Win_bytes_backward	int64
29	Bwd IAT Max	int64	69	act_data_pkt_fwd	int64
30	Bwd IAT Min	int64	70	min_seg_size_forward	int64
31	Fwd PSH Flags	int64	71	Active Mean	float64
32	Bwd PSH Flags	int64	72	Active Std	float64
33	Fwd URG Flags	int64	73	Active Max	int64
34	Bwd URG Flags	int64	74	Active Min	int64
35	Fwd Header Length	int64	75	Idle Mean	float64
36	Bwd Header Length	int64	76	Idle Std	float64
37	Fwd Packets/s	float64	77	Idle Max	int64
38	Bwd Packets/s	float64	78	Idle Min	int64
39	Min Packet Length	int64	79	Label	object
40	Max Packet Length	int64	80		

The data is imbalanced. This dilemma arises in IDS/IPS models because of a large number of false positives and false negatives. My decision will be to optimise precision over model fitting with no overfitting.

Table 2 Counts of all attacks in dataset

LABEL	COUNT
BENIGN	2273097
DoS Hulk	231073
PortScan	158930
DDoS	128027
DoS GoldenEye	10293
FTP-Patator	7938
SSH-Patator	5897
DoS slowloris	5796
DoS Slowhttptest	5499
Bot	1966
Web Attack Brute Force	1507
Web Attack XSS	652
Infiltration	36
Web Attack Sql Injection	21
Heartbleed	11

B. *Preprocessing*

According to this analysis, the IP address is translated to integer representation. The IP address mapping involves the Source IP Address (Src IP) and the Target IP Address (Dest IP) (Dst IP). These two numbers are transformed into their integer representation. This analysis broke the data into both the training and trial sets in an equal proportion of 80:20.

Step for preprocessing, you know.

- First, we can take away all the measured mean,
- standard deviation, minimum and maximum values.
- For the reduced feature space.

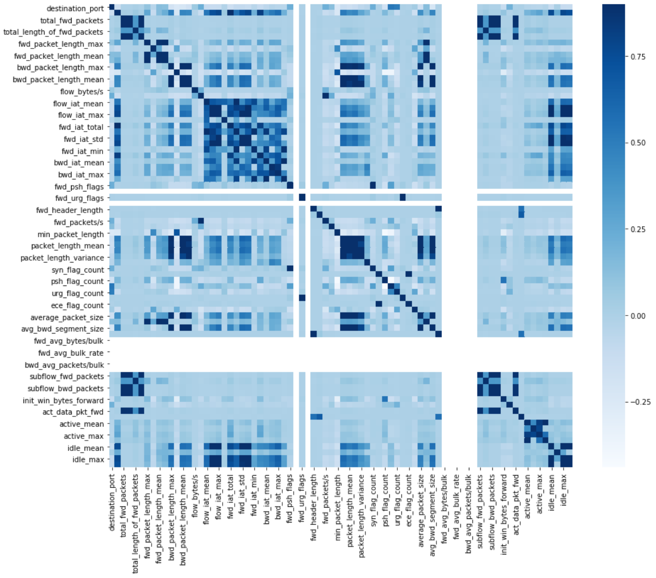


Fig. 2. Correlation between features in CICIDS2017 Dataset.

- Then encrypt the categorical names.
- Fix the missed values.
- Seeing holes in statistics.
- Find all the limitless or finite values.
- All nan values and residual non-finite values should be dropped.
- Still guarantee that the data frame is mounted correctly.

Table 3 Number of Features After and before preprocessing

Pre-processing	Entries	Features
Before	2830743	79
After	2827876	45

C. *Normalization*

In this step, we scaled all variables using Equation (1). Any sections of the initial dataset are between 0 and 1 while other parts are between 0 and ∞. Consequently, we minimised the range of input values between 0 and 1 to be interpreted by an auto-encoder.

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Since this value scale has a spectrum in [0;1], where xi is the value of a particular function, xmin is the minimum value, and xmax is the maximum value.

D. *Features Dimensionality Reduction*

Auto-Encoder (AE) Based Dimensionality Reduction

We address the sparse auto-encoder learning algorithm. Figure 2 displays a block diagram of the method. The lower dimensional hidden representation of the input vector x consists of many or one hidden layers s. (a1,a2, ...,am)

$$a_j^{(l)} = f(z_j^{(l)}) = f(\sum_{i=1}^{n} W_{ji}^{(l-1)}.a_i^{(l-1)} + b_j^{(l-1)})$$

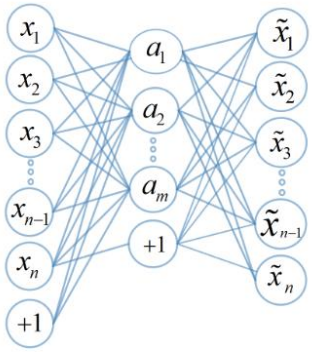


Fig 3.The structure of an AE.

Thus, the underlying secret representation an is then used to obtain the performance ˆx= (ˆx1, ˆx2, ..., ˆxn). Let j be the constant for every neuron in layer l, and I be the constant for every neuron in layer l−1. The output of a neuron in hidden layer can be expressed as this:

Principle Component Analysis (PCA) Based Dimensionality Reduction

Principal component analysis (PCA) is a method that is used to minimise the dimension of the given dataset. PCA is an effective and reliable approach for reducing the dimensions of data. A principal component reduction reduces the dimension of specified attributes into a limited number of dimensions called principal components. This approach accepts all the input as the dataset, which has a lot of attributes because the dimension of the dataset is very high. We approximate our dataset by taking data points on the same axis. The data points are moved to a single axis, and principal components analysis is carried out. The PCA method can be done with the following steps:

- Take the dataset that has all the dimensions, d.
- Calculate implies for each element on each axis.
- Evaluate the covariance matrix for the whole data collection.
- First we also need to determine eigen values of the symmetrical matrix (v1, v2, v3,vd).
- Sort the data into declining order.
- Using M type, you can get a new sample space.
- The primary components are collected.

E. K-means Clustering Algorithm

Clustering, focused on distance measuring of objects, and classifying objects (invasions) into clusters Unlike grouping, there is no accompanying specific details about the learning data. In order to find anomalous events, we can use welding and further research. The value of calculation is to relate findings into homogeneous categories. The Jacquard affinity calculation is critical because the event is to awaken the size to decide whether the meat has healthy bacteria or not. Euclidean distance gives a precise calculation of the distance between two vectors. Euclidean distance can be defined as the square root of the sum of the same vector's normalised squared differences. Both category and classification algorithms need to be efficient, massively, and feasible to manage the dimension of network data and heterogeneity [13].

In this study, we use K-means algorithm to cluster groups. K Means is one of the well known clustering strategies. K-means clusters the data in line with their characteristic values into a number of K user-specified clusters. In this way, data categorised into the same cluster has similar attribute values. K, the positive integer of number of clusters needs to be given. The measures involved in the K-means algorithm are described as below-

K data points to be clustered are inserted into the room. This show the major group centres. The data are allocated based on closeness to the centre of the cluster.

- The

positions of all the K centroids are modified right after any task.

Repeat steps 2 and 3 before it doesn't shift.

This results in the partition of data into groups. The preprocessed dataset partition is performed using the K-means algorithm with K value as 5. Because we have the dataset that contains normal and 4 attack categories such as DoS, Probe, U2R, and R2L.

F. Random Forest Algorithm

One of the most widely employed approaches used by science experts is random forest. It is supervised classification methodology. This software was developed to help deter an unnecessary forest from being created. The more trees there are in the forest, the more precise the outcome would be. However one needs to note that the action of building is not the same as obtaining knowledge or index.

Random Forests populates several classifications of knowledge. Each tree is planted accordingly.

1. Randomly pick N cases from your training data collection. This sample tree will be used to grow the big tree.
2. We randomly pick the m variables at each node, and then choose the best split that used the m chosen variables. The value of m is kept constant in the forest during the growing process.
3. Each leaf tree is grown as much as possible. There is no fall.

There are also wonderful advantages of Random Forest. The economic advantages are as follows:

• Clarity.

Excel functions operate consistently on massive arrays of data.

We manage thousands of input variables without deleting any of them.

•Creates a method of calculating incomplete data.

"Accuracy improves when a large proportion of the data are missing."

III. RESULTS

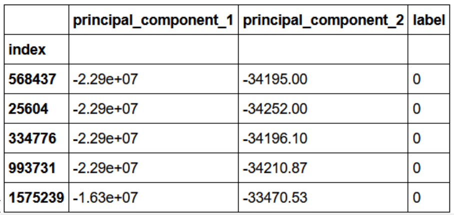
Variance Ratio of Principal Component Analysis after auto encoder.

The kmeans clustering algorithm might obtain several false positives because of the closer proximity of features. I think this is an indication that there is an immense difference between quantitative sociology and ordinary network results.

Using Elbow method to determine k-value and use k=30 since it is the point where the sum of squared error between the points and cluster centers reach the elbow point.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Table 4. Principal component 1 and principal component 2 analysis



	principal_component_1	principal_component_2	label
index			
568437	-2.29e+07	-34195.00	0
25604	-2.29e+07	-34252.00	0
334776	-2.29e+07	-34196.10	0
993731	-2.29e+07	-34210.87	0
1575239	-1.63e+07	-33470.53	0

A co vork and related work is highlighted in Table 5.The authors is reported the accuracy. Our proposed framework outperforms previous studies in terms of accuracy.

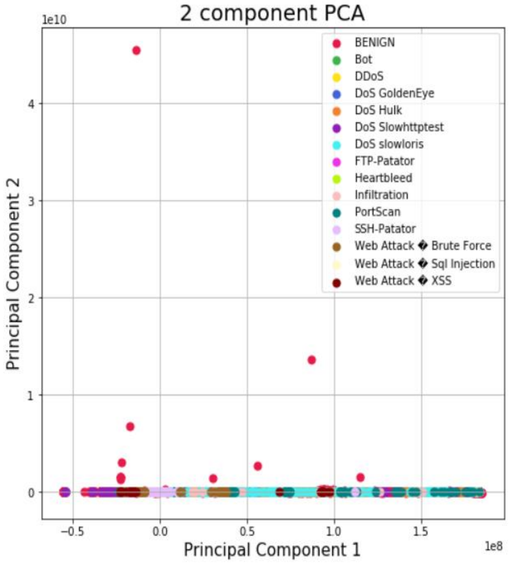


Fig. 4. Principal component 1 and principal component 2

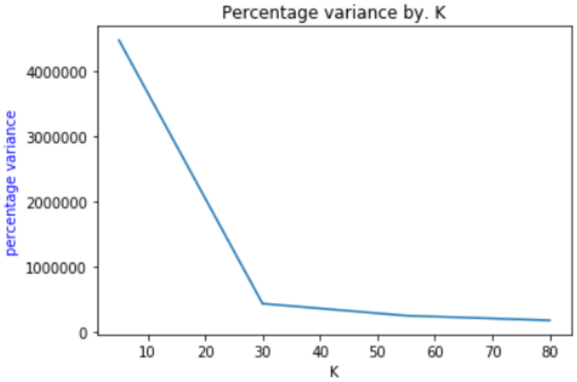


Fig. 5. Principal component 1 and principal component 2

Table 5. comparison of previous result with proposed result

Reference	Classifier name	FPR	TPR	Accuracy
1	PCA with Random forest	0.21	96.77	0.9678
2	AdaBoost classifier using PC	9. 997	0.8183	0.92
3	REP Tree, JRip algorithm and Forest PA	1.145	94.475	0.9666
4	Random Forest	4.10	94.69	0.9469
5	Deep Learning for Port Scan Attacks	3.59	97.85	97.80
6	Random Forest	3.53	97.60	97.90
7	Xgboost, random forest	4.35	91.99	92.00
8	Random Forest	0.002	93.71	0.9377
9	Random Forest	2.13	95.37	0.9586
EIDS	PCA, K-Mean and Random Forest	0.001	99.50	0.996

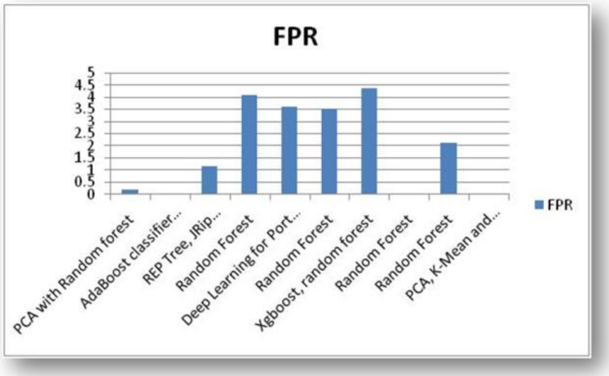


Figure 5.6 : Comparison with EIDS of False positive rate

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

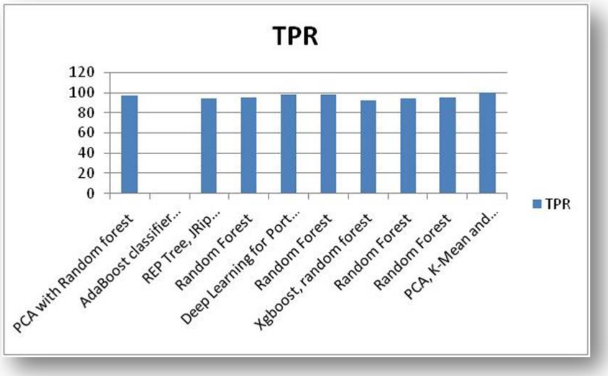


Figure 5.6: Comparison with EIDS of True positive rate

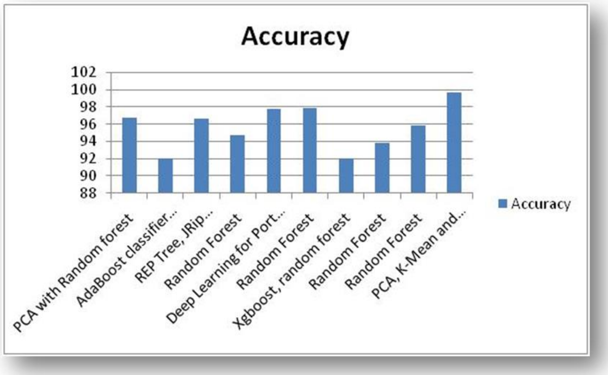


Figure 5.6: Comparison with EIDS of Accuracy

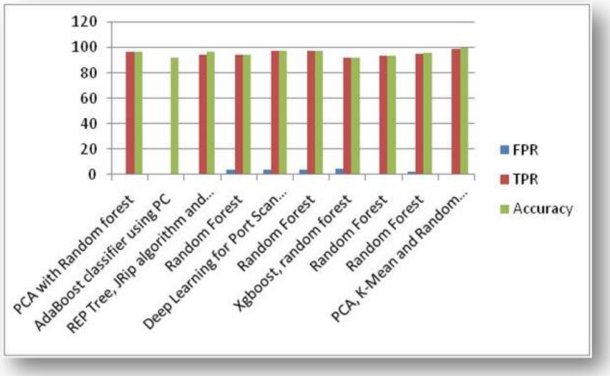


Figure 5.6: Comparison with EIDS of performance

I. Conclusion

The goal of this analysis was to obtain insights into a detection system by considering factors such as dimensionality reduction, PCA and network intrusion detection. Function dimensionality reduction methods contribute to improved efficiency metrics for some issues. This highlights the possible utility of PCA and auto-encoder in the reduction of dimensionality for IDS. From our tests, we find that PCA is easier, less costly, less interpretable and reduces dimensionality of data to two functions. The restricted preparation time and data constraints became obstacles for the methodology. This study indicated that Automatic Ensemble is applicable when a highly non-linear feature representation is required. Random Forest recognises many ideas from many different fields of the sectors.

Random Forest is recommended for defining essential factors in high-dimensional data. These reasons justify why Random Forest offered better classification results in comparison with other classifiers. The PCA dependent method in CICIDS2017 retained valuable information while effectively decreasing dimensions, and substantially portrayed. This analysis showed that PCA was equivalent to AE. Approaches PCA and AE are very special compared with PCA. This study will also act as a basis for potential laboratory assaults upon IDS structures. Such systems may be used to incorporate anomaly detection for online systems.

References

[1] M. Albanese et al., “Recognizing unexplained behavior in network traffic” in Network Sci. Cyber Sec. Berlin, Germany: Springer, pp. 39-62, 2014 doi:10.1007/978-1-4614-7597-2_3.

[2] “Abdul hammed, R; Faezipour, M; Elleithy, K. Intrusion Detection in Self organizing Network: A Survey”; Chapter in Intrusion Detection and Prevention for Mobile Ecosystems, vol. 13, G. Kambourakis, A. Shabtai, C. Kolias and D. Damopoulos, Eds. New York, NY, USA: CRC Press Taylor & Francis Group, 2017, pp. 393-449.

[3] C. H. Lee et al., “Machine learning based network intrusion detection” in Proc. 2017 2nd IEEE Intl. Conf. on Comp. Intell. and Appl. (ICCIA),Beijing, China, Sept. 8-11 2017, pp. 79-83.

[4] I. Sharafaldin et al., “Toward generating a new intrusion detection dataset and intrusion traffic characterization” in Proc. Fourth Intl. Conf. on Information Systems Security and Privacy. Funchal, Madeira, Portugal: ICISSP, Jan. 22-24 2018.

[5] C. O. S. Sorzano et al., A Survey of Dimensionality Reduction Tech-Niques, Available at: arXiv:1403.2877, Ar.Xiv2014.

[6] I. K. A. Fodor, Survey of Dimension Reduction Techniques. Livermore, CA, USA: Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, 2002; Vo-lume 9, pp. 1-18.

[7] S. Rosaria et al., Seven Techniques for Dimensionality Reduction. Zurich Switzerland: KNIME, 2014.

[8] L. Van Der Maaten et al., “Dimensionality reduction: A comparative review,” J. Mach. Learn. Res., vol. 10, pp. 66-71, 2009.

[9] P. Bertens, “Rank ordered,” Autoencoders, Ar.Xiv2016, Ar.Xiv: 1605.01749.

[10] R. Vijayanand et al., “Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature selection,” Comput. Secur., vol. 77, pp. 304-314, 2018 doi:10.1016/j.cose.2018.04.010.

[11] B. J. Radford and B. D. Richardson, “Sequence aggregation rules for anomaly detection in computer,” Network. Available at: Traffic.ar, p. Xiv2018, Available

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

at: arXiv:1805.03735.

[12] D. Lavrova et al., “Wavelet-analysis of network traffic time-series fordetection of attacks on digital production infrastructure.SHS Web,” CrossRef Conf. EDP Sci, vol. 44, no. 00052, 2018.

[13] G. A. Watson, Comparison of Header and Deep Packet Features When Detecting Network Intrusions [Technical report]. MD, USA: University of Maryland – College Park, 2018.

[14] D. Aksu et al., “Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm” in, Communications in Computer and Information Science Intl. Symp. on Comput. and Inf. Sci. Berlin, Germany: Springer, pp. 141-149, 2018 doi:10.1007/978-3-030-00840-6_16.

[15] [15] H. Wang et al., “Distributed Abnormal Behavior Detection Approach based on Deep Belief Network and Ensemble SVM using Spark,” IEEE Access, vol. 15) Marir, N., 2018.

[16] [16] A. Spark, 2018, “PySpark 2.4.0 documentation”. Available at: https://spark.apache.org/docs/latest/api/python/index.html.

[17] [17] A. Bansal, DDR Scheme and LSTM RNN Algorithm for Building an Efficient IDS [Master’s thesis]. PB, India: Thapar Institute of Engineering and Technology, 2018.

[18] T. Chen et al., Xgboost: Extreme Gradient Boosting. R Package version 0.4-2, 2015, pp. 1-4. Available at: http://cran.fhcrc.org/web/packages/xgboost/vignettes/xgboost.pdf.

[19] T. Hothorn and K. Hornik, 2015, “Zeileis, A. C tree: Conditional Inference Trees. The Compre-hensive R Archive Network”. Available at: https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf.

[20] M. E. Aminanto et al., “Deep abstraction and weighted feature selection for Wi-Fi impersonation detection,” IEEE Trans.Inform.Forensic Secur., vol. 13, no. 3, pp. 621-636, 2018 doi:10.1109/TIFS.2017.2762828.

[21] J. Zhu et al., “Mechanism of situation element acquisition based on deep auto-encoder network in wireless sensor networks,” Int. J. Distrib. Sens. Netw., vol. 13, no. 3, 2017 doi:10.1177/1550147717699625.

[22] [22] M. Al-Qatf et al., “Deep learning approach combining sparse Autoen-coder with SVM for network intrusion detection,” IEEE Access, vol. 6, pp. 52843-52856, 2018 doi:10.1109/ACCESS.2018.2869577.

[23] [23] M. Tavallaee et al., 2012, “Nsl-Kdd dataset”. Available at: http://www.unb.ca/research/iscx/dataset/iscx-NSL-KDD-dataset.html.

[24] D. Kibler et al. The UCI KDD archive of large data sets for data mining research and experimentation, “S.D.,” ACM SIGKDD Explor. Newsl., vol. 24) Bay, pp. 81-85, 2000,2.

[25] A. Javaid et al., “A deep learning approach for network intrusion detection system” in Proc. 9th EAI Intl. Conf. on Bio-
Inspired Information and Communications Technologies (Formerly BIONETICS), ICST. Social: Institute for Computer Science-Informatics and Telecommunications Engineering), Cotonou, Benin, May 24 2016, pp. 21-26.

[26] E. Min et al., “A semi-supervised and Unsu-pervised framework for network intrusion detection” in Intl. Conf. on Cloud Comput. and Sec. Cham, Switzerland: Springer, 2018, pp. 322-334.

[27] [27] D. Xia et al., “Intrusion detection system based on principal component analysis and grey neural networks” in Proc. 2010 Second Intl. Conf. on Networks Security, Wireless Communications and Trusted Comput., Wuhan, Hubei, China, vol. 2, Apr. 24-25 2010, pp. 142-145.

[28] [28] K. K. Keerthi Vasan and B. Surendiran, “Dimensionality reduction using principal component analysis for network intrusion detection,” Perspect. Sci., vol. 8, pp. 510-512, 2016 doi:10.1016/j.pisc.2016.05.010.

[29] A. Shiravi et al., “Toward developing a systematic approach to generate benchmark datasets for intrusion detection,” Comput. Secur., vol. 31, no. 3, pp. 357-374, 2012 doi:10.1016/j.cose.2011.12.012.

[30] M. E. Aminanto and K. Kim, “Improving detection of Wi-fi impersonation by fully unsupervised deep learning” in Proc. Information Security Applications: 18th Int. Workshop (WISA 2017), Jeju Isl., Korea, Aug. 24-26 2017.

[31] M. E. Aminanto and K. Kim, “Detecting active attacks in WiFi Network by semi-supervised deep learning” in Proc. Conf. on Information Security and Cryptography 2017 Winter, Sochi, Russian Federation, Sept. 8-10 2017.

[32] C. Kolias et al., “Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset,” IEEE Commun. Surv. Tutorials, vol. 18, no. 1, pp. 184-208, 2016 doi:10.1109/COMST.2015.2402161.

[33] D. W. Vilela, Ed.’Wilson, “T.F.; Shinoda,”, A.A.; de Souza Araujo, N.V.; de Oliveira, R.; Nascimento, V.E.A dataset for evaluating intrusion detection systems in IEEE 802.11 wireless networks. In Proceedings of the 2014 IEEE Colombian Conference on Communications and Computing (COLCOM), Bogota, Colombia ,4–6 June 2014; pp. 1–5.

[34] M. Ring et al., “Flow-based benchmark data sets for intrusion detection” in Proc. 16th Eur. Conf. on Cyber Warfare and Security, Dublin, Ireland, Jun. 29-30 2017, pp. 361-369.

[35] “Canadian Institute of Cyber security, University of New Brunswick. CIC Flow Meter,”, 2017. Available at: https://www.unb.ca/cic/research/applications.html#CICFlowMeter.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731