Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset

To cite this article: Arif Yulianto *et al* 2019 *J. Phys.: Conf. Ser.* **1192** 012018

View the article online for updates and enhancements.

Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset

Arif Yulianto [1], Parman Sukarno [2] and Novian Anggis Suwastika [3]

[1,2,3] School of Computing, Telkom University

E-mail: [1] arifyulianto@student.telkomuniversity.ac.id, [2] parmansukarno@telkomuniversity.ac.id, [3]anggis@telkomuniversity.ac.id

Abstract. This paper considers the use of Synthetic Minority Oversampling Technique (SMOTE), Principal Component Analysis (PCA), and Ensemble Feature Selection (EFS) to improve the performance of AdaBoost-based Intrusion Detection System (IDS) on the latest and challenging CIC IDS 2017 Dataset [1]. Previous research [1] has proposed the use of AdaBoost classifier to cope with the new dataset. However, due to several problems such as imbalance of training data and inappropriate selection of classification methods, the performance is still inferior. In this research, we aim at constructing an improvement performance intrusion detection approach to handle the imbalance of training data, SMOTE is selected to tackle the problem. Moreover, Principal Component Analysis (PCA) and Ensemble Feature Selection (EFS) are applied as the feature selection to select important attributes from the new dataset. The evaluation results show that the proposed AdaBoost classifier using PCA and SMOTE yields Area Under the Receiver Operating Characteristic curve (AUROC) of 92% and the AdaBoost classifier using EFS and SMOTE produces an accuracy, precision, recall, and F1 Score of 81.83 %, 81.83%, 100% , and 90.01% respectively.

1. Introduction

Intrusion Detection Systems (IDS) has turned out to be significant segments in computer and system security. The existing classification methods have their respective advantages and disadvantages, therefore the development of methodologies to overcome the shortcomings of existing classification methods is still very open. There are numerous techniques that have been proposed lately on IDS model and in addition explore on examinations of classification methods. However, most of the works used outdated datasets such as NSL KDD [2] and KDD Cup '99 [3]. For example, in 2015 Debachudamani Prusti [4] proposed several ensemble learning methods on NSL KDD. They concluded that using AdaBoost is a very efficient way to detect false positives and minimize false negatives. The comparison among ensemble AdaBoost, LogitBoost, and Bagging classifications, AdaBoost offers the highest accuracy and F1 Score of 97.44% and 97.9% respectively.

Ployphan Sornsuwit and Saichon Jaiyen [5] in their paper claimed that the AdaBoost ensemble learning method can increase accuracy up to 99.05%. They used a dataset of KDD Cup '99 especially in U2L and R2L classes. However, their method was only applied to U2L and R2L classes only which is not suitable to represent current attacks. Aburomman and Mamun 2016 [6] used Ensemble SVM based on PCA algorithm to evaluate NSL-KDD dataset. They obtained accuracy of 92.162%. In 2013, Abebe and Lalitha [7] conducted a comparison of several IDS methods using Random Forest (RF) as a classification with SMOTE and Feature Reduction for feature selection. They concluded that the RF ensemble classification method improved the detection of minority classes by using SMOTE on NSL-KDD dataset. The results showed precision and detection rate of 96.3% and 99.3% respectively for R2L and 96.2% and 97.2% for U2R.

Recently, research was conducted by Iman, Arash and Ghorbani in 2018 [1] to generate a new IDS dataset, namely Canadian Institute for Cybersecurity (CIC) 2017. The dataset importantly represents current attacks. The paper also evaluated several methods on the new dataset. It shows that although the performance of AdaBoost is superior as reported in [4], [5], and [7], the method produces inferior performance on CIC IDS 2017 Dataset. AdaBoost classification method obtains precision, recall, and F1 Score only of 77%, 84%, and 77% respectively. We think that the poor performance is due to several problems which are new attack patterns in the new dataset, imbalance in the training data, and improper classification method.

In our work we aim at constructing an improvement performance intrusion detection approach, in this correspondence, we apply the AdaBoost with SMOTE and PCA to intrusion detection. To overcome those problem, this paper proposes the use Synthetic Minority Oversampling Technique (SMOTE) [7] to handle the imbalance data. Moreover, Principal Component Analysis (PCA) [6] and Ensemble Feature Selection (EFS) [8] are used as the feature selection for selecting significant attributes in the IDS dataset.

The rest of paper is organised as follows. In section 2, we present the related works. In Section 3, we present the proposed method. In section 4, the experimental setup and results are described. In the last section, the conclusions is presented.

2. Related Work

2.1. Synthetic Minority Oversampling Technique (SMOTE)

A dataset is said to be imbalance if the classification category is not evenly distributed. Network data consists mostly of valid traffic with only a fraction of invalid traffic. For imbalance data, a large portion of the examples originate from the majority class. Accordingly, to limit the general expectation error rate, the order utilizing AdaBoost will build the precision in foreseeing the majority class, where it frequently results in poor prescient exactness for minority classes [9]. There are two strategies of resampling that are utilized for expanding the affectability of a classifier to the minority class: Under-testing of the majority class and Over-testing the minority class. To address the imbalance dataset issue, in this paper we utilized *Synthetic Minority Oversampling Technique* (SMOTE) as the pre-handling.

2.2. AdaBoost Methods

AdaBoost is an Adaptive Boosting, a machine learning algorithm figured by Yoav Freund and Robert Schapire [10, 11]. AdaBoost is a commonplace learning algorithm which can adequately enhance the characterization ability by numerous cycles. In initialisation arrange, all preparation tests are doled out to a similar weight to get some weak learners with some preparation emphasess. After each preparation cycle, the error rate of the weak classification is computed, the weights of the accurately classified examples is expanded, and the weights of the inaccurately grouped examples is decreased. At last, this weak learner turns into a strong learner to finish the order assignment. The objective of the AdaBoost is to advance the general precision. The calculation is outlined as pursues:

```
AdaBoost algorithm [10]:
1. Initiate weight coefficient data $\{W_n\}$with set $W_n(1) = 1/n$ , for n = 1, . . . , n.
2. Choose base classifier Y (it can be CART, Logit, probit, ect.) and train the data with the weight given.
3  For m = 1, . . . , m:
    a. Fit classifier ym (x) with training data by minimizing the weight of the error function.
           $Error_n = \sum_{n=1}^{n} W_n^{(m)} I\left(Y_m\left(X_n\right) \neq t_n\right)$
       Where (rumus) as an indicator function and will be worth 1 when (rumus) and 0 otherwise.
    b. Quantity evaluation
           $\varepsilon_m = \frac{\sum_{n=1}^{n} W_n^{(m)} I(Y_m(X_n) \neq t_n)}{\sum_{n=1}^{n} W_n^{(m)}}$
       Then use $a_m$ to evaluate
           $a_m = \ln\{\frac{1-\varepsilon_m}{\varepsilon_m}\}$
    c. Update the weight coefficients
           $W_n^{(m+1)} = W_n^{(m)} exp\{a_m I\left(Y_m\left(X_n\right) \neq t_n\right)\}$
4. Make predictions using the final model, i.e.
           $Y_M\left(x\right) = sign\left(\sum_{m=1}^{M} a_m Y_m\left(x\right)\right)$
```

2.3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is generally a statistical technique for data analysis and pre-processing that has been widely applied in various fields of research [12]. PCA is intended to change over information in diminished shape and keep the vast majority of the first variations present toward the start of the information. In numerical terms PCA, is utilized to change over n factors that relate into non-associated factors called Principal Component (PC) [13]. In this paper, PCA is used to select features that affect a data based on the highest correlation weight value.

2.4. Ensemble Feature Selection (EFS) Package on R Studio

The utilization of effective features to design classifier will not just diminish the extent of information yet in addition enhance classifier performance and understanding or visualization of data [14]. EFS Package is a R Studio package tool that provides a function to check the importance of features based on dependent classification variables [15]. An ensemble of feature selection methods is used to determine the normalized importance values of all features. Combining this method in one function (building a cumulative value) provides a stable feature selection tool. We used EFS package on R Studio to reduce selection feature.

2.5. Metric Performance

IDS execution examination is assessed as far as the quantity of features selected by the feature selection algorthm and the characterization precision of the machine learning algorithm. Several metrics have been designed to measure IDS effectiveness. This metric is separated into three classes: threshold, ranking and probability metrics [16]. To assess the consequences of our proposed method, we utilize execution measurements, for example, accuracy, recall, precision, and F1 Score. The confusion matrix is utilized to separate the prescient execution of the classification in the test data [17].

2.5.1. Confusion Matrix. Confusion matrix according to [17] can be interpreted as a tool that has a function to perform analysis whether the classifier is good in recognising the tuples of different classes. The calculation of the matrix confusion is showed in Table 1. The values of True-Positive and True-Negative provide information when the classifier in classifying the data is true, while False-Positive and False-Negative provide information when the classifier is wrong in classifying the data.

Table 1: Confusion Matrix.

		Predict Label	
		Intrusion	**Normal**
Host Label	**Intrusion**	TP	TN
	Normal	FP	FN

- TP (*True Positive*): The amount of data with positive actual value and positive predictive value.
- FP (*False Positive*): The amount of data with actual negative value and positive predictive value.
- FN (*False Negative*): The amount of data with the actual value is positive and the negative predictive value.
- TN (*True Negative*): The amount of data with actual negative value and negative predictive value.

The following measurement metrics are used to measure the performance of a dataset:

(1) *Accuracy* is the most intuitive measure of performance, which calculates precisely predicted observation ratios for total observation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

(2) *Recall* is the system's ability to detect all existing attacks. Recall can be calculated from the number of intrusions detected by the system rather than the number of actual intrusions.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

(3) *Precision* is a precisely predicted positive ratio of observed ratios to total positive observation predictions.

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

(4) *F1 Score* is the average of recall and precision values. This value takes positive and negative values. Intuitively this is not an accuracy, but F1 is usually more useful than accuracy, especially if it has an uneven distribution of classes.

$$F1Score = 2 * \frac{precision * recall}{precision + recall} \tag{4}$$

2.5.2. Curve ROC (Receiver Operating Characteristic). The ROC curve shows the visualization of the model accuracy and compares the differences between the classification models. Receiver Operating Characteristic (ROC) expresses confusion matrix [18]. ROC is a two-dimensional graph where false positive as horizontal line while true positives to measure the difference of performing method used. The ROC curve is a technique for visualising the performance of the classification [19]. Better classifier models are those that have larger ROC curves [18].

2.6. Dataset Descriptions

The CIC IDS 2017 dataset [1] is a data developed by the Faculty of Computer Science, University of New Brunswick in 2017. Based on previous research by Shiravi Ali [20] CIC 2017 is a refinement of the ISCX 2012 dataset [21]. The 2017 CIC dataset is generated from the actual generalisation of traffic.

The research conducted by [1] describes the characteristics of the IDS dataset characteristics and the type of approach used in developing the dataset. Faith, Arash, and Ali compare the dataset with other datasets. The evaluation framework of the last dataset published in 2016 [22] includes 11 required criteria for each dataset. These datasets almost meet these criteria: *Complete Network configuration, Labelled Dataset, Complete Traffic, Complete Interaction, Complete Capture, Available Protocols, Attack Diversity, Heterogeneity, Feature Set, Meta Data.* CIC IDS 2017 meets these criteria.

CIC IDS 2017 consist of 5 days of data collection with 225,745 packages with over 80 features and gathered more than seven days of network activity (i.e. normal and intrusion). In the CIC 2017 dataset, the attack simulation is divided into seven categories including Brute Force Attack, Heart Bleed Attack, Botnet, DoS Attack, DDoS Attack, Web Attack, and Infiltration Attack.

In this paper we analyze DDoS attacks on IDS. DDoS attacks usually occur when the system is flooding the bandwidth or resources of the victim. Such attacks are the result of some system compromises (eg, botnets) flooding targeted systems by generating massive network traffic. According to [1] the attribute correlation that affects DDoS attacks is like the time interval of *IAT Flow* related to *min, mean, max, bandwidth packet, total bandwidth, IAT bandwidth* and also *flow duration.* The greater value of the attribute, the higher the chance to be categorised as DDoS attack. For the category of normal correlation, attributes that affect are *B. Len Min server, Subflow F. Bytes, Total Len F.Packets, F.Packet Len Mean.*

3. Proposed Method

The proposed IDS technique is appeared in Figure 1. In our proposed strategy, Synthetic Minority Oversampling Technique (SMOTE) is utilized to enhance the sensitivity of arrangement for minority classes. The feature selection based on obtaining weighted value information is obtained from the ensemble result for feature selection. Another major component of this method is the AdaBoost classification which is used for classification in the training phase.

The SMOTE segment will take a minority class test from the data training to the required level. The subsequent SMOTE training data is normal as more parity data which will be specifically utilized as contribution for the feature selection. The part of feature selection ascertains the procurement of data from all features in the training data produced by SMOTE. The feature will then be ranked based on the value of the information obtained. To select an optimum feature section, the EFS package available on R-Package [8] is used.

The metrics used for evaluation are accuracy, precision, recall and F1 metrics. These metric are also used in [4–6, 23–26].

4. Experimental Setup and Result

Our proposed method is implemented on Intel Core i5-7200U 2.7GHz processor with 16GB RAM Nvidia Geforce 940MX VRAM GPU 2GB. We use R Studio 1.1,423 machine learning tool. R programming language for the development of our IDS model development.

For the proposed, we utilize the CIC 2017 Monday Working Hours dataset. After obtaining the required attributes, cleaning and normalization of the data are used to ensure the dataset is ready to be trained. CIC2017 DDoS *Monday-Working-Hours-DDoS-Attack* was separated into two parts: 70% of training data and 30% of testing data. The training data consist of 158,022 packet (128,737 labelled as normal and 29,285 labelled as DDoS). The testing

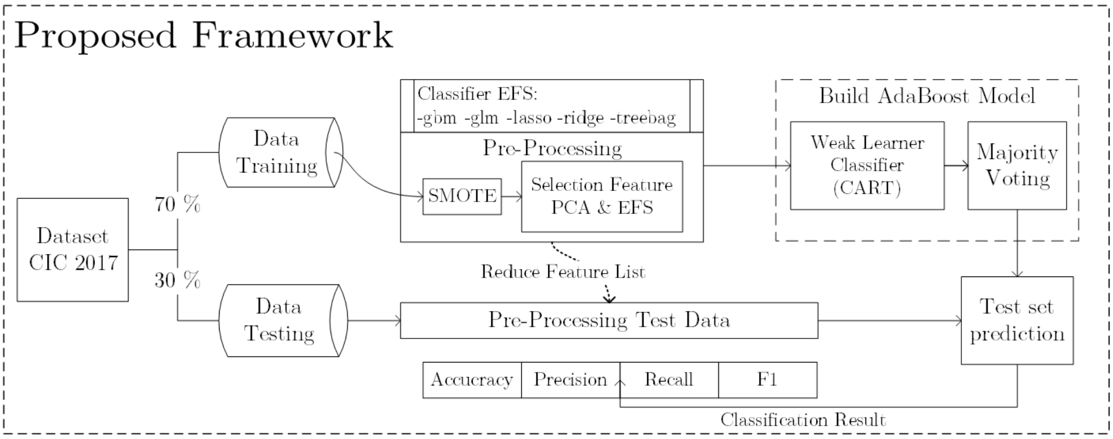


Figure 1: Proposed framework.

data consist of 67,723 packet (55,173 labelled as normal and 12,550 labelled as DDoS). After intensive experiments, to solve unequal training data problems in the CIC 2017 training data, we implemented SMOTE with a minority oversampling class of 200% and for feature selection we used a threshold value of T = 0.9. The number of minority class instances (DDoS) in training data increased from 29285 to 87855.

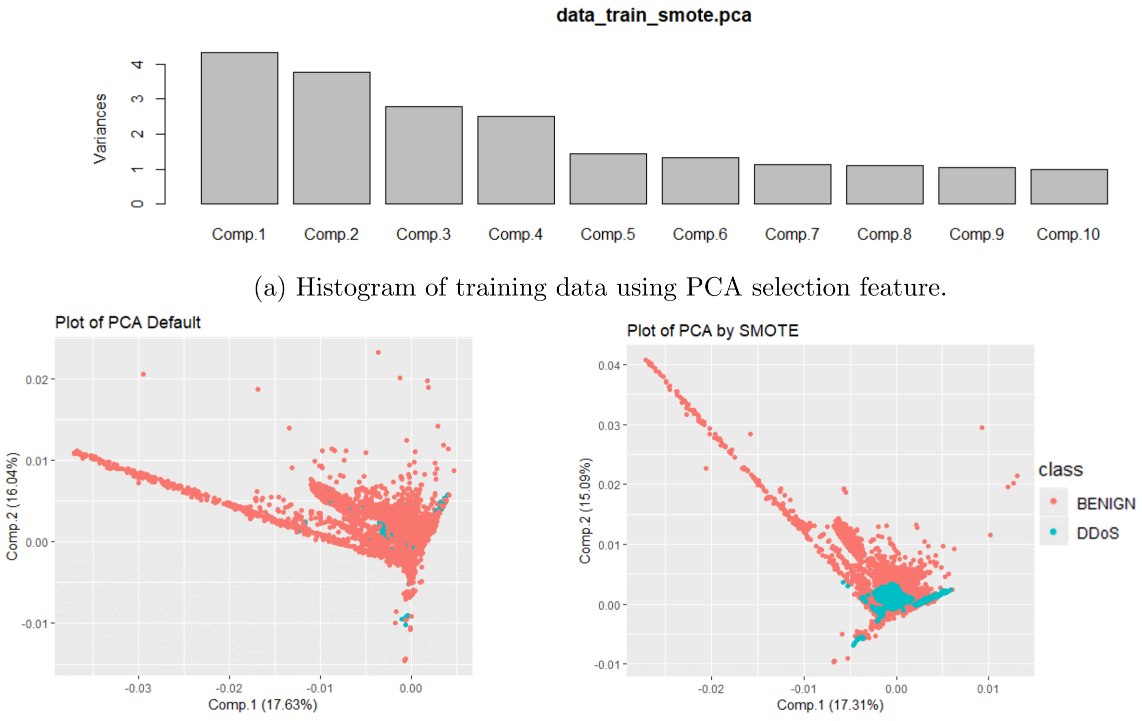
The feature selection uses the ensemble method ”gbm, glm, lasso, ridge and treebag” from the fscaret library [8]. We obtained 25 attributes with the largest weight value. The feature names of the ensemble selection results are shown in Table 2.

Table 2: List feature result using ensemble method.

No.	Name of Feature	No	Name of Feature	No	Name of Feature
1	Total.Length.of.Bwd.Packets	10	Bwd.IAT.Std	19	Average.Packet.Size
2	Fwd.Packet.Length.Min	11	Bwd.IAT.Min	20	Avg.Fwd.Segment.Size
3	Bwd.Packet.Length.Min	12	Fwd.Packets.s	21	Subflow.Fwd.Bytes
4	Bwd.Packet.Length.Std	13	Bwd.Packets.s	22	Init_Win_bytes_forward
5	Flow.IAT.Mean	14	Min.Packet.Length	23	Init_Win_bytes_backward
6	Flow.IAT.Min	15	Packet.Length.Variance	24	Active.Mean
7	Fwd.IAT.Min	16	PSH.Flag.Count	25	Idle.Min
8	Bwd.IAT.Total	17	ACK.Flag.Count		
9	Bwd.IAT.Mean	18	Down.Up. Ratio		

While the selection of features performed using Principal Component Analysis (PCA) taken from the PCA value that has 95% of the highest variance value, so obtained 16 features the highest PCA. Figure 2. shows the data observation using principal component analysis.

The training data was used AdaBoost by 10-Fold and repeated-cross validation 5 times. For the weak learner classification used using the AdaBoostM1 based on decision tree (CART (Classification and Regression Tree)) in R studio with coefficient learning Breiman $alpha = \frac{1}{2}\ln\left(\frac{(1-error)}{error}\right)$, Freund $alpha = \ln\left(\frac{(1-error)}{error}\right)$ and Zhu $alpha = \ln\left(\frac{(1-error)}{error}\right) + \ln(nclasses - 1)$.



(a) Histogram of training data using PCA selection feature.

(b) Plot using PCA feature.

(c) Plot using PCA + SMOTE.

Figure 2: Selection feature.

Table 3: Comparison performance metric.

		Proposed Framework			
metric/methods	**AdaBoost (Iman, et.Al 2018)* (%)**	**AdaBoost + EFS (%)**	**AdaBoost + EFS + SMOTE (%)**	**AdaBoost + PCA Feature (%)**	**AdaBoost + PCA Feature + SMOTE (%)**
Number of Feature	**72**	**25**	**25**	**16**	**16**
Accuracy (%)	-	81,47	***81,83***	81,47	81,47
Precision (%)	77	***85,15***	81,83	81,49	81,69
Recall (%)	84	94,92	***100***	99,93	95,76
F1 Score (%)	77	89,77	***90,01***	89,78	88,17
		TN 102	**TN** 3	**TN** 71	**TN** 8427
		FP 9605	**FP** 12250	**FP** 12512	**FP** 10478
	Confusion Matrix	**FN** 2945	**FN** 0	**FN** 38	**FN** 2072
		TP 55071	**TP** 55170	**TP** 55102	**TP** 46746

From Table 3. it can be seen that the results of the calculation of accuracy, precision, recall, predicted F1 value and testing data amounting to 67,723 packet.

Figure 3. shows the comparison of the AUROC curves of each classifier. The generated results are the value of Area Under the Receiver Operating Characteristic curve (AUROC) which is useful in determining the best classification model. It can be seen It very well may be seen that our proposed technique outflanks different strategies. The utilization of SMOTE has been proven in enhancing the detection rate of the minority classes in imbalance training data.

Moreover, important features of IDS CIC 2017 has been effectively selected by using ensemble feature selection.

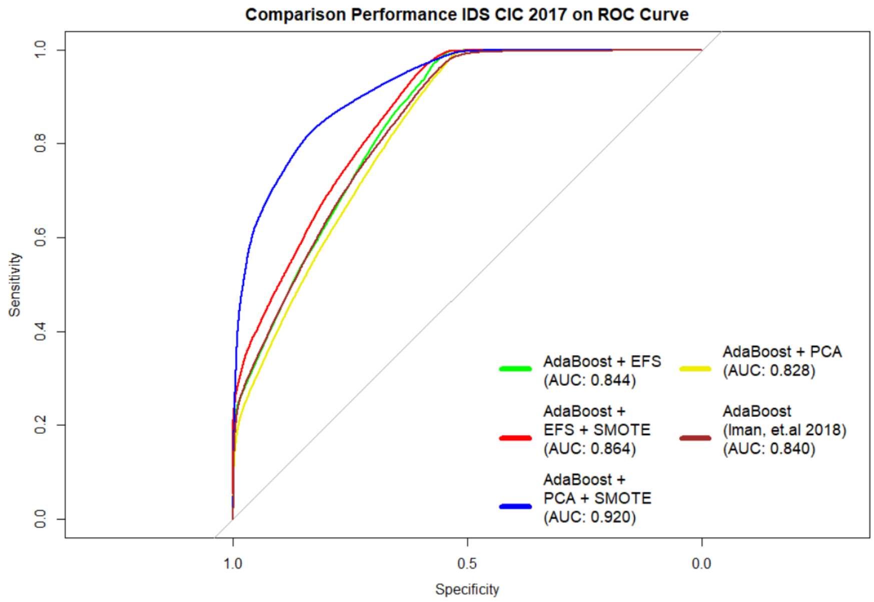


Figure 3: Comparison performance IDS CIC 2017 on AUROC curve.

5. Conclusion

It is the aim of this paper to contribute towards improving Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset. In this paper, we have considered the use of Synthetic Minority Oversampling Technique (SMOTE), Principal Component Analysis (PCA), and Ensemble Feature Selection (EFS) to improve the performance of AdaBoost-based Intrusion Detection System (IDS) on the latest and challenging CIC IDS 2017 Dataset. It has been proven that our proposed method outperforms the performance conducted by [6] with the accuracy of 81,83%, precision of 81,83%, recall of 100%, and F1 Score of 90,01%. It is an irreplaceable piece of the data security framework. Because of the assortment of system practices, it is important to create machine-learning-based IDS model with high performance. In the future, we can implements intrusion detection approach on the latest and challenging dataset using machine-learning-based so it can suitable for use in realtime application.

References

[1] Sharafaldin I, Habibi Lashkari A and Ghorbani A 2018 Toward generating a new intrusion detection dataset and intrusion traffic characterization *4th International Conference on Information Systems Security and Privacy (ICISSP)* pp 108–116

[2] Tavallaee M, Bagheri E, Lu W and Ghorbani A A 2012 *Available at http://www. unb. ca/research/iscx/dataset/iscx-NSL-KDD-dataset.html),[Accessed on 7 June. 2018]*

[3] Cup K 1999 *Available at http://kdd. ics. uci. edu/databases/kddcup99/kddcup99.html,[Accessed on 7 June. 2018]*

[4] Prusti D 2015 *An Efficient Intrusion Detection Model Using Ensemble Methods* (National Institute of Technology Rourkela, Rourkela: Department of Computer Science and Engineering)

[5] Sornsuwit P and Jaiyen S 2015 Intrusion detection model based on ensemble learning for u2r and r2l attacks *2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)* pp 354–359

[6] Aburomman A A and Reaz M B I 2016 Ensemble svm classifiers based on pca and lda for ids *2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEES)* pp 95–99

[7] Tesfahun A and Bhaskari D L 2013 Intrusion detection using random forests classifier with smote and feature reduction *2013 International Conference on Cloud Ubiquitous Computing Emerging Technologies* pp 127–132

[8] Neumann U, Genze N and Heider D 2017 *BioData Mining* **10** 21 ISSN 1756-0381 URL `https://doi.org/10.1186/s13040-017-0142-8`

[9] Li K, Xie P, Zhai J and Liu W 2017 An improved adaboost algorithm for imbalanced data based on weighted knn *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)* pp 30–34

[10] Freund Y and Schapire R E 1997 *Journal of Computer and System Sciences* **55** 119 – 139 ISSN 0022-0000 URL `http://www.sciencedirect.com/science/article/pii/S002200009791504X`

[11] Zhi-Hua Z 2012 *Ensemble Methods: Foundations and Algorithms* (Florida: Chapman & Hall/CRC)

[12] Ringnér M 2008 *Nature Biotechnology* **26** 303 EP – URL `http://dx.doi.org/10.1038/nbt0308-303`

[13] Shlens J 2014 *CoRR* **abs/1404.1100** (*Preprint* 1404.1100) URL `http://arxiv.org/abs/1404.1100`

[14] Mukherjee S and Sharma N 2012 *Procedia Technology* **4** 119 – 128 ISSN 2212-0173 2nd International Conference on Computer, Communication, Control and Information Technology(C3IT-2012) on February 25 - 26, 2012 URL `http://www.sciencedirect.com/science/article/pii/S2212017312002964`

[15] Nikita Genze U N 2017 Tool for ensemble feature selection URL `https://cran.r-project.org/web/packages/EFS/index.html`

[16] Caruana R and Niculescu-Mizil A 2006 An empirical comparison of supervised learning algorithms *Proceedings of the 23rd International Conference on Machine Learning* ICML '06 (New York, NY, USA: ACM) pp 161–168 ISBN 1-59593-383-2 URL `http://doi.acm.org/10.1145/1143844.1143865`

[17] XHEMALI D, HINDE C J and STONE R G 2009 Naive bayes vs. decision trees vs. neural networks in the classification of training web pages URL `http://cogprints.org/6708/`

[18] 2009 *Business Intelligence* (Wiley-Blackwell) chap 1, pp 1–19 ISBN 9780470753866 (*Preprint* `https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470753866.ch1`) URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470753866.ch1`

[19] Gorunescu F 2011 Data mining concepts, models and techniques URL `http://site.ebrary.com/id/10454853`

[20] Shiravi A, Shiravi H, Tavallaee M and Ghorbani A A 2012 *Computers & Security* **31** 357 – 374 ISSN 0167-4048 URL `http://www.sciencedirect.com/science/article/pii/S0167404811001672`

[21] UNB 2012 Intrusion Detection Evaluation Dataset by canadian institute for cybersecurity URL `http://www.unb.ca/cic/datasets/ids.html`

[22] Gharib A, Sharafaldin I, Lashkari A H and Ghorbani A A 2016 An evaluation framework for intrusion detection dataset *2016 International Conference on Information Science and Security (ICISS)* pp 1–6

[23] U N Wisesty A N K and Adiwijaya 2016 *Ind. Symposium on Computing* **Sept 2016** 165–176

[24] Wirawan I N T and 2015 I E 2015 *Jurnal Ilmiah Teknologi Informasi* **13** 182–189

[25] Amudha P, Karthik S and Sivakumari S 2015 Intrusion detection based on core vector machine and ensemble classification methods *2015 International Conference on Soft-Computing and Networks Security (ICSNS)* pp 1–5

[26] W Yassin Nur Izura Udzir Z M and Sulaiman M N 2013 *International Conference on Computing and Informatics, ICOCI 2013* 49