ISSN 2278-3091
Volume 9, No.1.3, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse1991.32020.pdf
https://doi.org/10.30534/ijatcse/2020/1991.32020

Optimizing the Effectiveness of Intrusion Detection System by using Pearson Correlation and Tune Model Hyper Parameter on Microsoft Azure Platform

Fuad Mat Isa[1], Alya Geogiana Buja[2], Mohammad Yusof Darus[3], Shahadan Saad[4]
[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia, fuadisa19@gmail.com
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia, geogiana@uitm.edu.my
[3]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia, yusof@tmsk.uitm.edu.my
[4]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia, shahadan@uitm.edu.my

ABSTRACT

This paper studies the effectiveness of implementing classifier algorithm and Pearson correlation for an intrusion detection system. The effectiveness of intrusion detection system is evaluated based on the accuracy, detection rate and false positive rate. This study has been implemented by using simulation in Microsoft Azure platform. The machine learning algorithm is applied together with filter-based feature selection which provides the classifier algorithm to ensure the quality of the NSL-KDD, KDD 99 and CICIDS 2017 dataset. In addition, the tune model hyper parameter has been applied to enhance the performance of the classifier algorithm. The findings show that the implementation of classifier algorithm and Pearson correlation for an intrusion detection system has been able to improve the effectiveness of intrusion detection system in terms of accuracy, detection rate and false positive rate. Results from this study are useful for designing an effective intrusion detection system in the future due to the advancement of network attacks that are growing rapidly. This study can be further extended and improved by investigating the effectiveness of intrusion detection system in real network environment.

Key words: Classifier Algorithm, Feature Selection, Intrusion Detection System, KDD, Machine Learning, Pearson Correlation

1. INTRODUCTION

The usage of Internet is increasing from time to time with the evolution of computing technology; cloud and mobile computing. This enables users to access and retrieve online contents at anywhere, anytime. Due to this current state of Internet usage, cybercrime and misuses of Internet are also increasing [1]. Network security tool or system is the first line of defense. Intrusion detection system (IDS) is a network security tool that typically used for mitigating and minimizing the attack. IDS allow its systems to recognize a strange or anomaly traffic and trigger a caution when an interruption occurs[2].There are two types of IDS; host-based IDS (HIDS) and network-based IDS (NIDS). In order to prevent the attack, IDS detects signature or learn the profile of an attack by analyzing information that has been obtained.

This paper focuses on the optimization of NIDS using machine learning algorithm on Microsoft Azure platform. Network-based IDS analyzes data packets moving through the network, and this analysis can be carried out in two ways namely anomalies and signature-based [3]. The challenge of the study on detecting anomalies based on anomalies is that it has to deal with unprecedented and unknown attacks. The IDS with anomaly detection should be able to study the profile of the attack. This is done by distinguishing between healthy and uninterrupted traffic flow and machine learning techniques that researchers have been exploring for the past few years [4].

Therefore, studies and research works on NIDS with anomaly detection has been able to capture the interest and attention of researcher community[5]. There are several anomaly detection techniques have been proposed including machine learning algorithm Among the studied algorithms are Decision Tree, Genetic Algorithm (GA), Linear Regression, Naive Bayes Classifier, Neural Network, Support Vector Machine (SVM) and Tree Decision Forest [6]–[8] . In previous studies, three datasets have been used namely KDD 99 [9], [10], NSL-KDD [7] and CICIDS 2017 [11]. Many supervised learning models are used to solve classification problems.

Recently, studies of IDS are focusing on the implementation of machine learning algorithm [7], [8], [12]–[16]. Thus, in our works, classifier algorithms with Pearson Correlation have been studied. This paper presents the proposed optimization method in Section 2. The findings and results of comparison with other studies of this study are discussed in Section 3. Section 4 concludes and provides the future works of the study.

2. PROPOSED METHOD

The proposed optimization method has been implemented as in Figure 1 on Microsoft Azure Platform using KDD 99, NSL-

KDD and CICIDS 2017 dataset. The following subsections explain all involved process in the proposed optimization method.

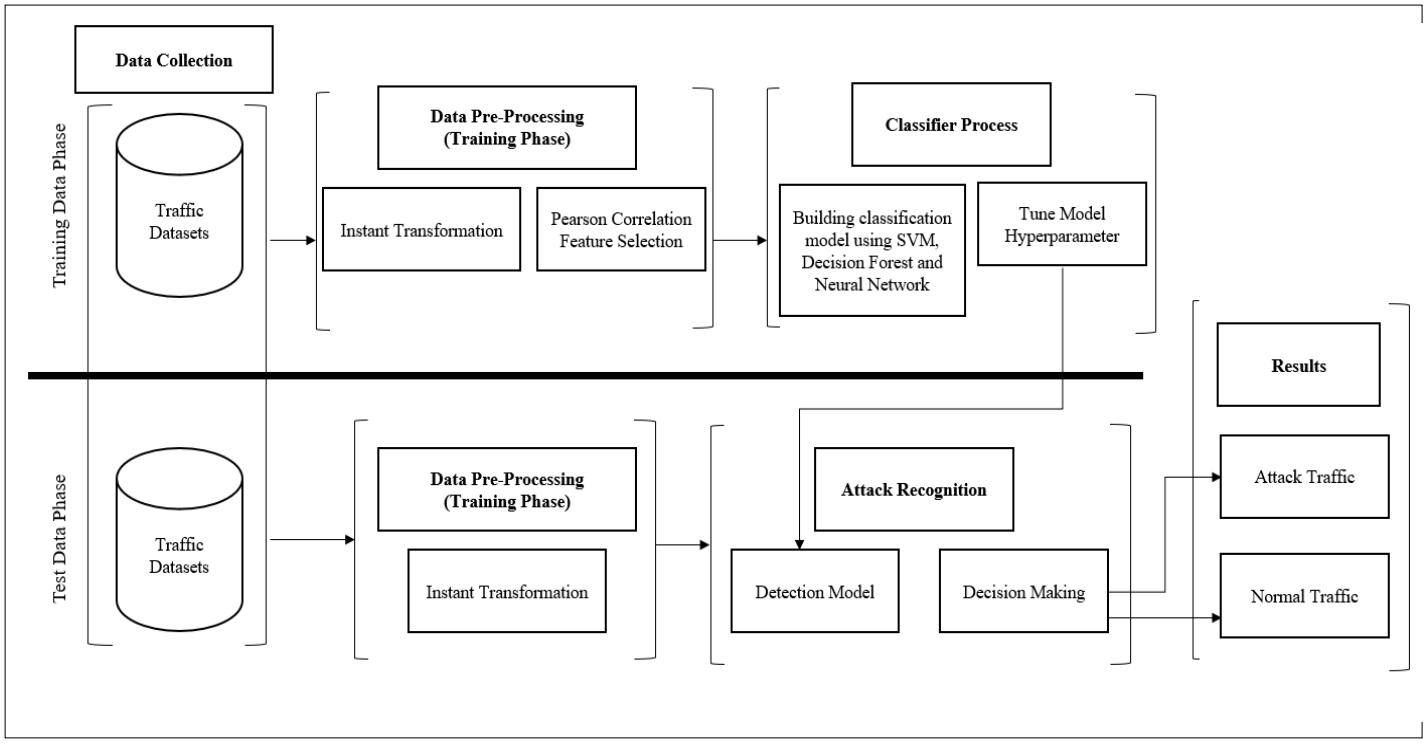

Figure 1: The framework of the IDS based on Pearson Correlation and Tune Model Hyperparameter

2.1 Research Methodology

The overall approach for performing the research in this study is shown in Figure 1. Figure 1 shows four phases namely data collection, data pre-processing, data analytic and performance evaluation. In data pre-processing phase contains sub processes which is instances transformation and normalization and feature selection. In data analytic phase contains classifier algorithm and tune model hyperparameter. The details of each phases and sub-phases will be explain further in the next section.

2.2 Data Collection

Data collection is a crucial process because it will determine effectiveness of an IDS. Data collection is important phase because IDS model detection can be only be as good in terms of accuracy and efficiency as the data on which it bases its decisions. Datasets that is acquired need to be sufficient and optimum in order the classifier to produce a better performance. In this research, three datasets had been used, namely KDD 99, NSL- KDD, and CICIDS 2017. KDD 99 dataset were created by Lee and Stolfo [17] based on DARPA network dataset files. DARPA consist of row TCP/IP dump files and act as a base dataset. Feature extracted from DARPA dataset and pre-processed for machine learning use and thus KDD99 was born. Compared to the DARPA dataset, KDD 99 can be easily used with machine learning research. KDD 99 dataset contains following characteristic:

1. KDD 99 contains 41 attributes with different features and label assigned to determine the traffic as an attack or normal.

2. The last attributes contain data of five categories of network vector. This network vector is grouped in one normal class and four attack classes. The four attack classes are DoS, Probe, R2L and U2R.

3. Contained duplicate records in both training and testing datasets

NSL-KDD is the enhanced version from the previous KDD99 dataset. NSL-KDD solves the following problems; Classifier will produce an un-biased result as redundant records are removed, Train and test data sets have a sufficient number of records which is optimum to execute experiments on the compete set and on each of group level difficulty, the number of selected records is inversely correlative to the percentage of records in the KDD dataset. CICIDS 2017 is new dataset developed by [18]. CICIDS 2017 covers with up-to-date common attacks. The dataset is extracted from network traffic analysis using CICFlowMeter with complete labelled and 80 network traffic features extracted.

2.3 Data Pre-Processing

This phase contains two main processes, instances transformation selection. The details of the process are as follows:

2.3.1 Instances Transformation

The datasets from data collection phase had been processed here. Instance Transformation phase started by changing the label feature into categorical type. Next is convert to indicator values process. Label feature that contained in three datasets was converted into indicator value to determine the type of network traffic. Then all the columns in dataset were selected and ready for feature selection process.

2.3.2 Pearson Correlation

IDS need an informative feature of traffic data so it is important to identify one to achieve a better performance. This study proposed a Pearson Correlation-based feature selection method to determine whether a feature is important. Pearson Correlation-based feature selection is a module that used selected metric to determine unrelated attributes, and exclude out redundant columns from dataset. The columns are returned ranked by their feature scores. The important features can improve the performance of classification.

2.4 Classifier Process

Three supervised learning algorithms are chosen for this sub phases. They are mentioned [10] as the algorithms are suitable for making prediction because the input from the dataset highly correlated to the result produced by classifier algorithm and thus outcome can be predicted based on mathematical function. Three algorithms that are used in this study are Support Vector Machine, Neural Network and Decision Forest.

2.5 Tune Model Hyperparameter

The module based on building and testing multiple models by using different combinations of settings. It also compares metrics over all models to get the combination of settings. All these processes are called tuning, which is the process of finding the optimal configuration. This module run integrated train and tune which based on a set of parameters, the module iterates over multiple combinations, measuring accuracy until it finds a best model.

2.6 Performance Evaluation

According [19], IDS detection technique can be based on performance of two criteria, efficiency and effectiveness. Efficiency handles the resources needed to be allocated to the system including CPU and memory. Effectiveness illustrates the system's ability to differentiate between normal and intrusive activities. In this section, metrics used for evaluation will be further discuss. In order to evaluate the results produced by the model, they can be evaluate using a series of metrics of that can be derived from the confusion matrix. The metrics used for evaluation are accuracy, detection rate and false positive rate.

Accuracy: This metric shows the total number of correctly classified traffics including normal and intrusive. The accuracy is most important parameter detect and differentiate objects [20].

$$Accuracy=(TP+TN)/(TP+TN+FN+FP)$$

Detection Rate: DR is the ratio between the number of correctly detected attacks and the total number of attacks.

$$Detection\ Rate=TP/(TP+FN)$$

False Positive Rate (FPR): FPR is the ratio between the number of misclassified normal connections and the total number of normal connections.

$$False\ Positive\ Rate=FP/(FP+TN)$$

3.RESULTS AND DISCUSSION

The existing method for detection in IDS suffers drawback in terms of accuracy and the features in datasets used is not contributing to the detection results of the IDS. In this research, the combination of Pearson Correlation and Tuning Model Hyperparameter and tested with SVM, Decision Forest and Neural Network was proposed to provide a better performance of IDS. The proposed model has been tested with three different dataset, NSL-KDD, KDD 99, and CICIDS 2017. The results and the discussion were explained in detail in this chapter.

IDS must work efficiently and actively to repel any kind of intrusion. In order to work efficiently, the detection must work in real time and has to performed relatively fast. The proposed method, Pearson Correlation with further enhancement by Tune Model Hyperparameter were designed to improve traditional IDS systems in terms of accuracy, detection rate, and building normal and malicious traffic profile with a minimum amount of data. All the goals mentioned were achieved in this research as shown in the results and discussed here in this section.

Table 1: SVM classification performance with proposed method

	Accuracy	Detection Rate	False Positive Rate
NSL-KDD	94.9	99.7	0.0001
KDD 99	94.9	95.4	0.0725
CICIDS 2017	85.5	64.1	0.02

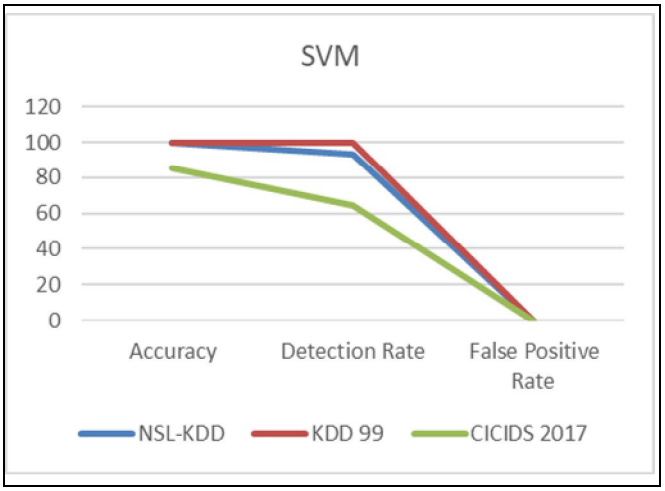

Figure 2: SVM performance graph

Since constructing normal and malicious profile were based on data traffic, hundreds of features were analyzed. Before the data was fed into the classifier, applying feature selection method to synthesize the data proved to be useful. The classifier algorithms were very successful in developed the model that very accurate. With the addition of tuning module, it boots the performance of classifier to achieve greater results. Pearson Correlation filter-based algorithm has been used for feature selection for a long time with proof of effectiveness. The classification performance of the IDS model combined with Pearson Correlation and tune model hyperparameter is shown in Table 1, clearly demonstrate that the IDS's classification performance is enhanced by the Pearson Correlation process and tune model hyperparameter. It shows that the combination of the Pearson Correlation and tuning model has achieved and accuracy rate of 94.9, 94.9 and 85.5 percent based on three datasets on SVM classifier. However, accuracy achieved by CICIDS 2017 dataset suffers a bit drawback in terms of accuracy. This might suggest that SVM classifier might not well train enough with the latest trend of intrusion on network.

Table 2: Decision Forest performance with proposed method.

	Accuracy	Detection Rate	False Positive Rate
NSL-KDD	99.9	93.3	0.03
KDD 99	99.9	99.8	0.0007
CICIDS 2017	99.9	99.9	0.002

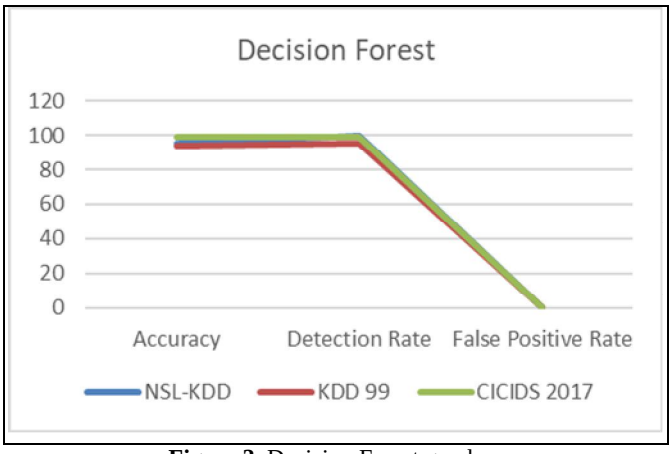

Figure 3: Decision Forest graph

The results of Decision Forest classifier performance without and with proposed model are shown in Table 2. The results show that Decision Forest classifier achieved a quite significant value in terms of accuracy and detection rate. Based on three datasets, Decision Forest classifier combined with Pearson Correlation and tuning model hyperparameter has achieved a constant accuracy rate of 99.9, 99.9, 99.9 percent and 93.3, 99.8, 99.9 percent in terms of detection rate across three datasets.

This shows a high classification result. Decision Forest outperform SVM on CICIDS 2017 datasets by a margin of 14.4 percent on accuracy rate. CICIDS 2017 is a larger dataset compare to the other two datasets. This might indicate that Decision Forest perform better with a bigger dimensionality of data and be able to read a new kind of attacks in network.

Table 3: Neural Network performance with proposed method

	Accuracy	Detection Rate	False Positive Rate
NSL-KDD	99	99	0.9
KDD 99	99	98.9	0.0086
CICIDS 2017	88	67.4	0.01

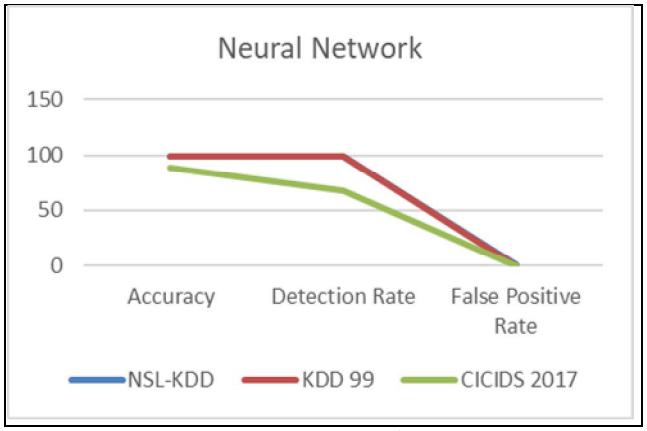

Figure 4: Neural Network performance graph

Table 3 summarize the classification results of Neural Network classifier using the different datasets with regard to accuracy rates, detection rates and false positive rates. It shows clearly that the detection model combined with the Pearson Correlation and tuning model has achieved and accuracy rate of 99, 99 and 88 percent based on three datasets on Neural Network classifier. Based on this result, the performance of Neural Network reacts almost the same as SVM classifier for CICIDS 2017. Neural Network could not perform properly and suffers as same as SVM in terms of accuracy performance. Another finding is that Neural Network algorithms run slowly due to their nature of working with multilayers so features selections will come in handy to speed up classifier process.

Based on the results, Pearson Correlation combine with tune model hyperparameter efficiently improved the detection model with features selection on the datasets. As mentioned by [21], features selection is to improve the accuracy of classification algorithm and this statement can be strengthen by the results obtained by the proposed model.

4.COMPARATIVE STUDY

In order to demonstrate the accuracy performance of the proposed method – Pearson Correlation with Tune Model Hyperparameter – experiments have been conducted to make comparison with some of the contemporary literatures. Namely by Rachburee & Punlumjeak [8] titled Big Data Analytics: Feature Selection and Machine Learning for Intrusion Detection on Microsoft Azure Platform, and Taher, Mohamed Yasin Jisan & Rahman titled Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection [7].

While both literatures experimented with more than one method, Rachburee & Punlumjeak with Mutual Information and Chi-Square, Taher et al. with CFS Correlation Filter and Chi-Square, this comparative study experimented and adopted their best method for accuracy performance as a mean of equal-comparison. Hence this study focuses on the comparative results of Mutual Information by Rachburee & Punlumjeak and Chi-Square by Taher et al.

Table 4: Result of comparison with other approaches in terms of accuracy using NSL-KDD dataset.

	Accuracy
Decision Forest+ Pearson Correlation + Tuning Model (Proposed Method)	99.9
Decision Tree + Chi Square (taher2019)	99.2
Decision Tree + Mutual Information (rachburee2017)	99.9

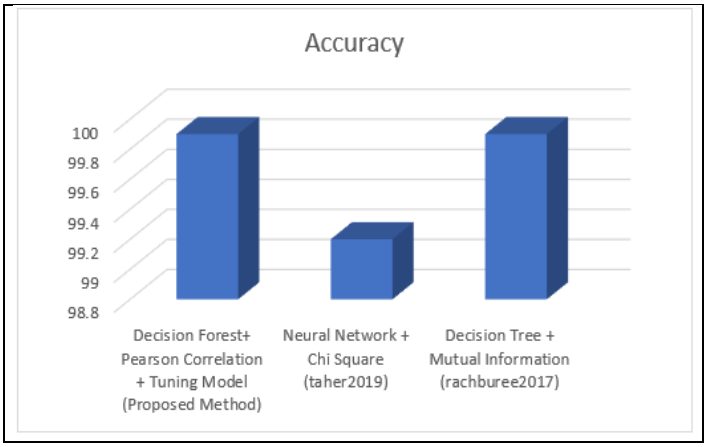

Figure 5: Accuracy performances on NSL-KDD dataset

Table 5: Result of comparison with other approaches in terms of accuracy using KDD 99 dataset

	Accuracy
Decision Forest+ Pearson Correlation + Tuning Model (Proposed Method)	99.9
Decision Tree + Chi Square (taher2019)	99.2
Decision Tree + Mutual Information (rachburee2017)	99.9

Figure 6: Accuracy performances on KDD 99 dataset

Table 6: Result of comparison with other approaches in terms of accuracy using CICIDS 2017 dataset

	Accuracy
Decision Forest+ Pearson Correlation + Tuning Model (Proposed Method)	99.9
Decision Tree + Chi Square (taher2019)	99.8
Decision Tree + Mutual Information (rachburee2017)	99.4

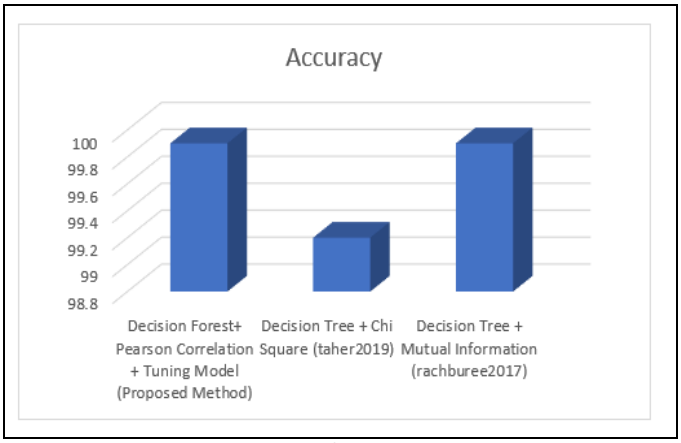

Figure 7: Accuracy performances on CICIDS 2017

Using the same approach with three datasets, Microsoft Azure platform, and the proposed method which is Pearson Correlation with tune model hyperparameter, the recorded results suggest a highly-consistent score in terms of accuracy rate across three datasets. For NSL-KDD, the accuracy scored at 99.9 percent. For other datasets, CICIDS 2017 and KDD 99, they achieved the same value of accuracy, which is 99.9 percent. Considering the limit of our current device and engineering capacity that restricts from reaching 100 percent performance for accuracy, it is implied that the proposed Pearson Correlation method with Tune Model Hyperparameter is the best available feature selection method against other contemporary studies that exist at the date of this research.

5. CONCLUSION AND FUTURE WORK

We proposed an optimization of intrusion detection system using classifier algorithm and Pearson Correlation in improving the accuracy, detection rate and false positive rate. The proposed optimization approach has been applied on training and test data of KDD 99, NSL-KDD and CICIDS 2017. The results show that Support Vector Machine and Pearson Correlation performed well for all evaluation metrics on KDD 99 and NSL-KDD meanwhile Tree Decision Forest excellent on CICIDS 2017. In addition, the proposed method of employing classifier algorithms; Support Vector Machine and Tree Decision Forest together with Pearson Correlation and Tuning Model has achieved higher accuracy compared to Decision Tree + Chi Square and Decision Tree + Mutual Information. For future works, we intend to implement the proposed optimization method in real environment. Besides, the existing classifier algorithms can be improved based on current dataset such as CICIDS 2017.

REFERENCES

1. P. Alaei and F. Noorbehbahani. **Incremental anomaly-based intrusion detection system using limited labeled data**, *2017 3rd Int. Conf. Web Res. ICWR 2017*, pp. 178–184, 2017. https://doi.org/10.1109/ICWR.2017.7959324
2. N. Chandra Sekhar Reddy, P. C. R. Vemuri, and A. Govardhan. **An implementation of novel feature subset selection algorithm for IDS in mobile networks**, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 5, pp. 2132–2141, 2019. https://doi.org/10.30534/ijatcse/2019/43852019
3. T. S. Urmila and R. Balasubramanian. **A novel framework for intrusion detection using distributed collaboration detection scheme in packet header data**, *Int. J. Comput. Networks Commun.*, vol. 9, no. 4, pp. 97–112, 2017.
4. M. Zamani and M. Movahedi. **Machine Learning Techniques for Intrusion Detection**, pp. 1–11, 2013.
5. P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez. **Anomaly-based network intrusion detection: Techniques, systems and challenges**, *Comput. Secur.*, vol. 28, no. 1–2, pp. 18–28, 2009.
6. A. S. Ashoor and S. Gore. **Importance of Intrusion Detection System (IDS)**, *Int. J. Sci. Eng. Res. 2011*, vol. 2, no. 1, pp. 1–4, 2011.
7. K. A. Taher, B. Mohammed Yasin Jisan, and M. M. Rahman. **Network intrusion detection using supervised machine learning technique with feature selection**, *1st Int. Conf. Robot. Electr. Signal Process. Tech. ICREST 2019*, pp. 643–646, 2019.
8. N. Rachburee and W. Punlumjeak. **Big data analytics: Feature selection and machine learning for intrusion detection on Microsoft azure platform**, *J. Telecommun. Electron. Comput. Eng.*, vol. 9, no. 1–4, pp. 107–111, 2017.
9. M. Sabhnani and G. Serpen. **Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context**, *Proc. Int. Conf. Mach. Learn. Model. Technol. Appl.*, pp. 209–215, 2003.
10. A. Abubakar and B. Pranggono. **Machine learning based intrusion detection system for software defined networks**, *2017 Seventh Int. Conf. Emerg. Secur. Technol.*, pp. 138–143, 2017. https://doi.org/10.1109/EST.2017.8090413
11. R. Panigrahi and S. Borah. **A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems**, *Int. J. Eng. Technol.*, vol. 7, no. 3.24 Special Issue 24, pp. 479–482, 2018.
12. M. C. Belavagi and B. Muniyal. **Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection**, *Procedia Comput. Sci.*, vol. 89, pp. 117–123, 2016.
13. J. Zheng, F. Shen, H. Fan, and J. Zhao. **An online incremental learning support vector machine for large-scale data**, *Neural Comput. Appl.*, vol. 22, no. 5, pp. 1023–1035, 2013.

14. F. Gharibian and A. A. Ghorbani. **Comparative study of supervised machine learning techniques for intrusion detection**, *Proc. - CNSR 2007 Fifth Annu. Conf. Commun. Networks Serv. Res.*, pp. 350–355, 2007.
15. J. Schmidhuber. **Deep Learning in neural networks: An overview**, *Neural Networks*, vol. 61, pp. 85–117, 2015. https://doi.org/10.1016/j.neunet.2014.09.003
16. M. Tavallaee, N. Stakhanova, and A. A. Ghorbani. **Towards Credible Evaluation of Anomaly-based Intrusion Detection Methods,** Faculty of Computer Science , University of New Brunswick," *Evaluation*, 2010.
17. O. Atilla and E. Hamit. **A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015**, *PeerJ*, pp. 0–21, 2016.
18. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani. **Toward generating a new intrusion detection dataset and intrusion traffic characterization**, *ICISSP 2018 - Proc. 4th Int. Conf. Inf. Syst. Secur. Priv.*, vol. 2018-Janua, no. Cic, pp. 108–116, 2018.
19. M. Tavallaee. **An Adaptive Intrusion Detection System**, *Univ. New Brunswick*, 2011.
20. B. Dudi and V. Rajesh. **Medicinal plant recognition based on CNN and machine learning**, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 4, pp. 999–1003, 2019. https://doi.org/10.30534/ijatcse/2019/03842019
21. I. Ullah and Q. H. Mahmoud. **A filter-based feature selection model for anomaly-based intrusion detection systems**, in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 2151–2159. https://doi.org/10.1109/BigData.2017.8258163