Efficient model intrusion detection system through feature selection based on decision tree algorithm

By

Johani Fauzi
Faculty of Engineering and Information Technology Swiss German University Banten, Indonesia
Email: johani.fauzi@student.sgu.ac.id

Charles Lim
Faculty of Engineering and Information Technology Swiss German University Banten, Indonesia
Email: charles.lim@sgu.ac.id

Eka Budiarto
Faculty of Engineering and Information Technology Swiss German University Banten, Indonesia
Email: eka.budiarto@sgu.ac.id

Abstract

When it comes to data analytics on machine learning, feature selection (FS) is one of the most significant responsibilities of data preparation. This research aimed to combine feature selection to assess significant aspects of massive network traffic, which is further utilized to increase the accuracy of traffic anomaly detection while also decrease the time it takes to complete the analysis. The filter-based and wrapper-based feature selection techniques are the most often utilized method in Intrusion Detection System (IDS) research. This study employed a combination of filter-based and wrapper-based methods by ranking the features according to their minimum weight values to select relevant and significant features, and then implemented the decision tree classifier for learning algorithms in experiments on two datasets of NSL-KDD and CIC-IDS2017 to determine appropriate and powerful features. The experiment investigation demonstrated that the number of relevant and essential features produced were combined by Information Gain and Chi-Square forward selection that has a substantial impact on the improvement of detection accuracy and the production of simpler elements. The feature selection combined filter-based and wrapper-based chi-square forward selection decision tree (IG+FS+DT) algorithm with nine features selected and has the highest accuracy of 98.98%. In this case, the nine features selected were f2, f5, f30, f32, f33, f36, f37, f38, f41 from 41 features on NSL-KDD dataset. In the CIC-IDS2017 dataset, the combination of CS+FS+DT selected best five feature and obtained the highest accuracy of 99.98%. In this case, the five features selected were f2, f17, f23, f65, and f68 from 80 features on the dataset of CIC-IDS2017.

Keywords - Intrusion Detection System, Decision Tree, Feature Selection, filter-based, wrapper-based, NSL-KDD dataset, CIC-IDS2017 dataset

Introduction

Internet access is currently considered one of the most fundamental requirements of most people on the planet. The services provided by the internet are indeed beneficial in our lives today. You can imagine that if the internet services are not accessible, it will have an impact on other life problems. Along with this, the Data Centre has become a business that is quite a concern because it provides various services needed by internet users [1].

Along with the development of business and data centre technology, it is inevitable that there will be problems that can cause services to become unavailable. Therefore, it needs to be anticipated through a preventive first step to keep services available. It is noted that from the services provided by internet providers and other content service providers on the internet, cybercrimes of various forms often occur, starting from identity fraud, data theft, and ransomware attacks. It was recorded that the average cost of the data breach is around 945 million US dollars [2].

Therefore, monitoring a system or network and spotting anomalous activity is the job of an IDS. Host-based IDS or network-based IDS can be used as intrusion detection and prevention [3]. It is possible to have two versions of network intrusion detection systems or NIDS: one that is focused on network events and another one that is focused on individual hosts. Packet sniffers, which are programs that read raw packets from a local network segment, are scanned by an NIDS to identify suspicious activity [4].

In addition to be able to read packet headers and detect specific attack types, it can also follow more network objectives in an effort to detect dangers that may be missed by HIDS. NIDS, for example, can successfully detect most IP-based DoS attacks since they can only inspect packet headers as they go over the network. For example, NIDS do not rely on operating systems to identify hosts but instead require the Operating System (OS) to work. For intrusion detection, there are also hybrid IDS that integrate client and network-based solutions [5].

Detection techniques for intrusions can also be used to detect misuse or anomalies [6]. IDS uses three types of detection mechanisms, those are statistics-based, pattern-based, rule-based, state-based, and heuristic-based. In this case, all types of decision-making are any and all forms of decision-making; in which intrusions are identified primarily through the use of established thresholds, mean and standard deviation, as well as probability and probabilities-based techniques. Meanwhile, pattern-based detection, which employs string matching to identify known assaults, is used to detect these attacks. Rule-based techniques, in addition to being used to detect known intrusions, make use of If-Then and If-Then-Else rules to create a model and profile of known intrusions, which can then be used to detect future intrusions. State-based approaches, in particular, take advantage of finite state machines formed from network activities to distinguish assaults. The latest of them is technique that has been determined by using heuristics strategy as a result of biological principles as well as artificial intelligence concepts heuristic-based [7]. Comparing the four feature selection algorithms on detection attacks, the five algorithms are IG, chi-square, term strength and mutual information. Their results show that IG and chi-square are the most efficient. Furthermore, between the FGLCC-CFA and the decision tree, the results of the study showed that FGLCC-CFA got a good score [8].

From all the research results that have been carried out, the most appropriate Model has not been found to find the highest accuracy with simple feature for the network-based intrusion detection dataset. Therefore, the author will use a decision tree classification learning algorithm to find a feature selection algorithm that combines filter-based: Information Gain (IG) Chi-Square (CS), Correlation (CR) with wrapper-based: Forward Selection (FS), Backward Elimination (BE) and Optimize Selection (OS). The comparison of the combination of feature selection algorithms with the best performance and less feature used intrusion detection datasets based on NSL-KDD and CIC-IDS2017 networks that have been used by various researchers on this topic.

The purpose of this study was to combine data mining classification algorithms and feature selection for network-based intrusion detection datasets. The specific research objectives of this research are:

- To develop a combined feature selection and technique for searching the best combination feature selection for intrusion detection system
- To develop a combined feature selection and technique for improving the classification accuracy of intrusion detection system
- To develop a combined feature selection technique for searching the most influential features of intrusion detection system

Methods

A quantitative methodology has been employed in this study. In this case, the purpose of quantitative methods was to develop models, theories, and hypotheses related to natural phenomena. This research method was carried out in several stages in the research as follows:

Data Preparation

In this process, the data was acquired in detailed. Specifically, this research employed a sampling assault to obtain data from the NSL-KDD dataset for the investigation. This dataset is an upgraded version of the KDD Cup99 dataset. It is available in the following link: https://archive.ics.uci.edu/ml/datasets.php. The inclusion of a substantial number of duplicate records in the KDD Cup99 dataset is the dataset's most major fault and it is the most serious problem in the dataset [9] (Revathi and Malathi, 2013). A network connection between two network hosts is built on the basis of the protocols used by the network. Detailed information about the characteristics is provided in figure 3.2, including the titles of the attributes and the sorts of information that has been gathered. There are five different types of network connections featured in the 41 features. In this case, they were labelled as follows: one class is designated as the normal class (normal flow), and the remaining four classes are designated as intrusion traffic. The major four intrusion classes are as follows: DOS, probe, R2L, and U2R (Denial of Service) [9]. This dataset has been utilized in a number of previous research projects (Sivatha Sindhu, Geetha and Kannan, 2012) (Mirvaziri, Ghazizadeh-Ahsaee and Karimipour, 2019) (Selvakumar and Muneeswaran, 2019) [9-11]. One normal traffic type and 24 different intrusion traffic types are included in this dataset; they are organized in Table 3.2 into four groups, each of which is described in greater detail below.

- Unauthorized users are denied access to a computer as a result of a distributed denial of service (DDoS) attack that, by using resources and memory, floods the network with unnecessary applications or keeps part of the computations or memory resources occupied for a prolonged length of time.
- A "U2R attack" is a kind of abusive attack in which the attacker gets access to a network by using a regular user's account and then tries to acquire access to the network's root.
- An R2L attack is a kind of network attack in which an attacker tries to acquire local access to a network as a machine user who does not have administrative privileges on the system. Generally speaking, probing attacks are focused on acquiring information from a network of computers in order to use the information in subsequent applications.

Table 1: *NSL-KDD assaults should be categorized according to their kind and severity (Revathi and Malathi, 2013)*

Classification Of The Attack	Category Of Attack
DoS	Back, Land, Neptune, Pod, Smurf, Teardrop, Apache2, Udpstorm, Processtable, Worm
U2R	Buffer_overflow, Loadmodule, Rooktik, Pear, Sqlattack, Xterm, Ps
R2L	Guess_Password, Ftp_write, Imap, PhD, Multihop, Warezmaster, Warezclient, Spy, Clock, Xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named
Probes	Satan, Upsweep, Nmap, Portsweep, Mscan, Saint

Initial Data Processing

This stage explained how to process and transform the data to fit the desired form.

In this study, 2 different intrusion detection datasets were used, the first was the NSL-KDD dataset which had been cleaned (Revathi and Malathi, 2013). Since there are no duplicated records in the train set, the classification algorithm will not provide any biased results. Given the fact that there are no duplicate entries in the test set, the reduction rate was much larger than in the control set. When choosing records from each difficult leave group, the percentage of records in the original KDD dataset is inversely proportional to the number of records in each difficult leave group.

This data can be directly processed using the Model that has been determined in this study and further produced a value that can measure of how good the model used in this research is. After getting the best model, the next step was using the intrusion detection dataset, particularly by using the latest data sampling, namely CIC -IDS2017 which was checked again in this latest dataset of how well the Model found in the NSL-KDD dataset and re-applied to the CIC-IDS2017 dataset.

Proposed Model

After the data processing stage has been carried out in the previous stage, the next step was to determine the Model, which was further tested on the previously prepared data. Proposed diagram approach for this research is clearly described step by step in this diagram. Public dataset intrusion from NSL-KDD or another dataset whose criteria have been determined based on the data mining process criteria preprocessing that was done.

The stage was first carried out by selecting the features that have been determined by filter-based with learning algorithm using decision tree ranking feature information gain, correlation, and chi-square selected subset feature. After that, the next feature was selected from the wrapper based forward selection, backward elimination, optimize selection and then result and evaluation using accuracy, recall, precision, and feature selected by combined step one and step two.

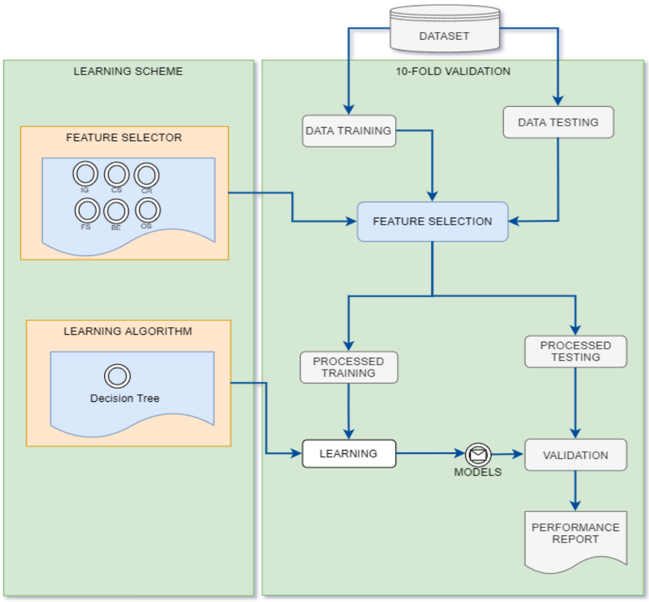

Fig.1: *Proposed Model Framework Approach*

In Figure 3.7, partitioning was carried out on the dataset into ten parts using 10-fold cross-validation, where all parts of the dataset became training data and testing data. After the data were divided into training data and test data, the next step was selecting the features using a filter-based (information gain, chi-square statistics, and correlation) in order to select the most influential feature, The features were than recalculated using a wrapper-based feature selection (forward selection, elimination, backwards, optimization selection). Furthermore, an algorithm was used for learning the feature selection, which was a decision tree classification algorithm. In this case, the results in the performance of this Model was further measured in terms of its Accuracy, Recall, and Precision.

Experiment and Test

At this stage, a test was carried out on the proposed Model to get the performance results of the proposed Model.

Evaluation Model and Result Validation

At this last research stage, an evaluation of the experiment and testing of the proposed model were carried out to find out the performance results of this study.

The validation was carried out using 10-fold Cross-validation where the dataset was divided into ten parts, one part being the setting data and the other part being the training data. The performance of the Model was further compared using decision tree classification without feature selection filter-based or wrapper-based (DT) between feature selection filter-based information gain and decision tree (IG+DT), feature selection filter-based information gain and feature selection wrapper-based forward selection with learning decision tree algorithm (IG+FS+DT). In addition, comparison was also conducted between the feature selection filter based on information gain and feature selection wrapper-based forward selection, as well as backward Elimination and learning decision tree algorithm (IG+FS+BE+DT), and feature selection filter-based information gain with feature selection wrapper-based optimize selection and decision tree learning algorithm (IG+OS+DT).

The comparison results of these features was further employed to assess the accuracy, recall, precision and feature selected so that the results obtained are more accurate.

Result and duscussion

Based on the results of feature selection using various combinations of feature selection between filter-based and wrapper-based on the NSL-KDD dataset, the results are shown in Table 2. In Table 2, information is given between the algorithm used, the number of features, the selected features based on the results of each feature selection combination. Therefore, in an effort to detect IDS using the NSL-KDD dataset, the most influential features to improve detection accuracy are the following four features: f2, f36, f32, and f33.

Table 2: *Feature appear frequently in dataset nsl-kdd*

Method Combined	#OF Feature	Category of Attack	Accuracy
IG+FS+DT	9	f2, f5, f30, f32, f33, f36, f37, f38, f41	98.95%
CR+BE+DT	12	f2, f3, f4, f8, f12, f28, f30, f32, f33, f36, f38, f41	97.25%
CS+FS+DT	9	f2, f3, f5, f26, f32, f33, f35, f36, f37	98.78%

Based on the NSL-KDD dataset, the most influential features to determine attack detection are f2: PROTOCOL_TYPE: During the connection, the protocol was utilized, f32: DST_HOST_COUNT: Number of connections having the same destination host IP address, f33: DST_HOST_SRV_COUNT: Number of connections having the same port number, f36: DST_HOST_SAME_SRC_PORT_RATE: The percentage of connections that were to the same source port, among the connections aggregated in DST_HOST_SRV_COUNT.

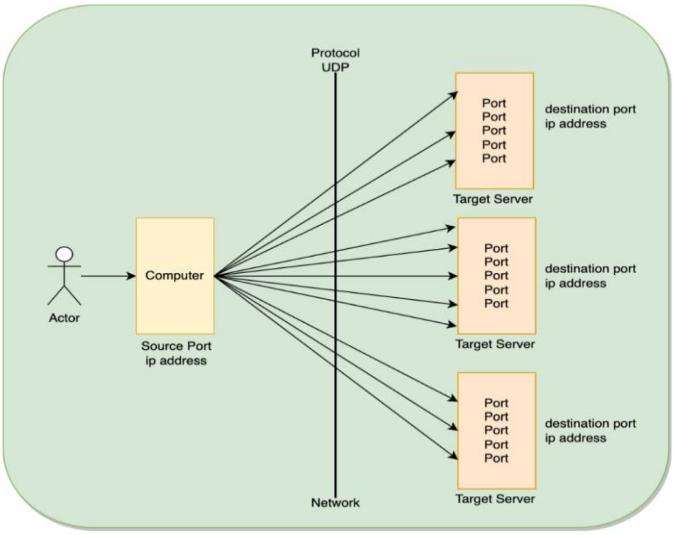

Fig. 3 *Illustration Attack Network*

Intrusion detection can be described in figure 4.6. In this case, the traffic can be determined if the number of protocols feature of f2: PROTOCOL_TYPE or During the connection, utilized high number and only one source is considered an attack. Based on the source port to the destination, the features used were f32: DST_HOST_COUNT and f33: DST_HOST_SRV_COUNT, where the destination of this attacker goes to the target to be attacked. In this feature, the actor's public IP and destination were recorded. Furthermore, the feature f36: DST_HOST_SAME_SRC_PORT_RATE traffic sent by one actor to the destination to the same port continuously is considered an attack.

Feature Selected Analysis Using Combination Srategy Dataset CIC-IDS2017

Table 3: *feature appears frequently in dataset nsl-kdd summary selected feature result using various techniques dataset cic-ids2017*

No	Method	# Of Feature	Selected Feature Number
1	IG+DT	51	f17, f18, f9, f11, f56, f68, f70, f23, f26, f16, f14, f24, f7, f65, f25, f2, f13, f15, f57, f5, f64, f37, f37, f67, f20, f55, f43, f10, f41, f28, f31, f12, f51, f44, f45, f29, f30, f42, f68, f18, f4, f21, f19, f70, f40, f22, f38, f6, f66, f54, f22
2	IG+FS+DT	8	f2, f5, f7, f18, f23, f29, f66, f68
3	IG+BE+DT	44	f4, f5, f6, f7, f37, f9, f10, f11, f12, f13, f14, f15, f16, f17, f18, f19, f20, f21, f22, f23, f24, f26, f27, f28, f29, f31, f38, f39, f40, f41, f42, f43, f44, f45, f51, f54, f55, f56, f57, f37, f64, f65, f66, f67
4	IG+FS+BE+DT	8	f2, f5, f7, f18, f23, f29, f66, f68
5	IG+OS+DT	7	f2, f23, f29, f38, f64, f65, f68
6	IG+OS+FS+DT	4	f2, f23, f64, f68
7	IG+OS+BE+DT	4	f2, f7, f23, f68
8	IG+OS+FS+BE+DT	5	f65, f2, f68, f23, f31
9	CR+DT	52	f15, f57, f13, f1, f2, f51, f43, f55, f44, f41, f42, f45, f70, f14, f11, f56, f9, f7, f65, f28, f54, f12, f31, f30, f49, f33, f47, f29, f10, f20, f39, f18, f19, f68, f26, f21, f25, f50, f78, f54, f40, f75, f77, f38, f72, f6, f66, f32, f27, f79, f17, f46
10	CR+FS+DT	6	f2, f10, f12, f65, f68, f78
11	CR+BE+DT	35	f9, f10, f11, f12, f13, f14, f15, f16, f18, f19, f20, f21, f25, f26, f28, f29, f30, f31, f33, f40, f41, f42, f43, f44, f45, f47, f49, f50, f51, f54, f57, f65, f68, f70, f78
12	CR+FS+BE+DT	6	f7, f2, f68, f78, f20, f18
13	CR+OS+DT	8	f2, f7, f20, f29, f30, f31, f68, f78
14	CR+OS+FS+DT	6	f2, f7, f20, f29, f68, f78
15	CR+OS+BE+DT	8	f2, f7, f10, f20, f29, f65, f68, f78
16	CR+OS+FS+BE+DT	4	f7, f2, f68, f78
17	CS+DT	29	f15, f57, f13, f16, f2, f51, f43, f55, f44, f41, f42, f45, f70, f14, f11, f56, f9, f7, f65, f28, f54, f12, f31, f30, f49, f33, f47, f29, f10
18	CS+FS+DT	5	f2, f17, f23, f65, f68
19	CS+BE+DT	60	f4, f5, f6, f7, f37, f9, f10, f12, f13, f14, f15, f16, f17, f18, f19, f20, f21, f22, f23, f24, f25, f27, f28, f29, f30, f31, f32, f33, f38, f18, f41, f42, f15, f44, f45, f46, f47, f48, f49, f50, f51, f53, f54, f55, f57, f37, f64, f65, f66, f67, f68, f69, f71, f72, f73, f74, f75, f77, f78, f79
20	CS+FS+BE+DT	6	f7, f2, f68, f23, f31, f64
21	CS+OS+DT	6	f2, f7, f20, f63, f68, f78
22	CS+OS+FS+DT	6	f2, f7, f20, f63, f68, f78
23	CS+OS+BE+DT	5	f2, f7, f20, f68, f78
24	CS+OS+FS+BE+DT	5	f7, f2, f68, f23, f32

According to the results of feature selection using various combinations of feature selection between filter-based and wrapper-based in the CIC-IDS2017 dataset, the results are shown in Table 3. Table 3 shows information between the algorithm used, the number of features, and the selected features based on the results of each feature selection combination.

Based on the results of using a combination of feature selection algorithms in the CIC-IDS2017 dataset, we tried to discuss the most influential features so as to produce an accuracy improvement above 97%. Table 3 provides information on the correlation of feature in the CIC-IDS2017 dataset with a combination model of the feature selection algorithm which provides an accuracy above 97%.

The most influential feature is:
F2: Destination Port And F68: Init_Win_Bytes_Forward

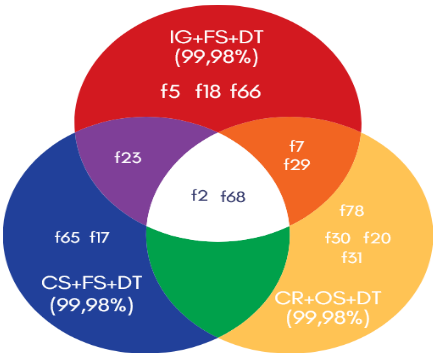

Fig. 4 *Diagram Venn Feature CIC-IDS2017*

Therefore, the most influential features to improve detection accuracy are the following four features in the IDS detection effort using the CIC-IDS2017 dataset: f2 and f68.

Table 4: *Feature appear frequently dataset cic-ids2017*

Method Combine	#OF Feature	Selected Feature Number	Accuracy
IG+FS+DT	8	f2, f5, f7, f18, f23, f29, f66, f68	99.98%
CR+OS+DT	8	f2, f7, f20, f29, f30, f31, f68, f78	99.98%
CS+FS+DT	5	f2, f17, f23, f65, f68	99.98%

Based on the CIC-IDS2017 dataset, the most influential features to determine attack detection are f2: DESTINATION PORT Destination Port to target, and f68: INIT_WIN_BYTES_FORWARD. Figure 4.8 illustrates an actor's attack on several servers with different destinations and the same port.

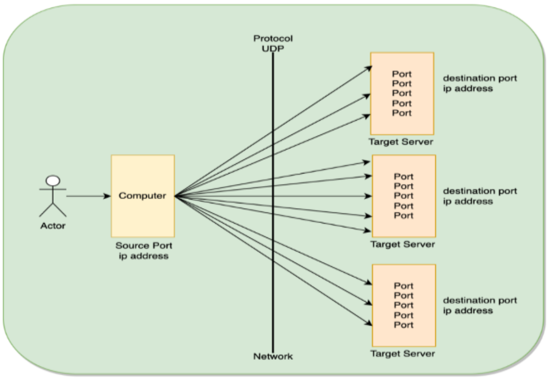

Fig. 5 *Illustation Attack Network CIC-IDS2017*

Conclusion

There are several ways utilized in IDS machine learning technique research, and it is believed that it is vital to continue to develop new machine learning methods in order to get the greatest outcomes in detecting an attack. A large number of irrelevant characteristics may be extracted from the collection of attack types in the NSL-KDD and CIC-IDS2017 datasets, and these features can cause a decline in the performance of the learning method. One of the learning algorithms utilized in this study is the decision tree, which is the best method for identifying network threats.

The experimental results in this study get the simplest feature for the NSL-KDD dataset, where the combination obtained is IG+FS+DT with the selected subset only four features getting an accuracy value of 98.95%. Furthermore, for the CIC-IDS2017 dataset, the combination obtained was the simplest is Information Gain Optimize Selection Forward Selection CS+FS+DT, where the subset obtained from the combination is only four features with a high level of accuracy compared to all features in the intrusion detection dataset.

Acknowledgment

Swiss Germany University for all support to this research

References

Nord JH, Koohang A, Paliszkiewicz J. The Internet of Things: Review and theoretical framework. Expert Systems with Applications. 2019 Nov 1;133:97-108.

Fleisch E. What is the internet of things? An economic perspective. Economics, Management, and financial markets. 2010;5(2):125-57.

Nord JH, Koohang A, Paliszkiewicz J. The Internet of Things: Review and theoretical framework. Expert Systems with Applications. 2019 Nov 1;133:97-108.

Fleisch E. What is the internet of things? An economic perspective. Economics, Management, and financial markets. 2010;5(2):125-57.

Nguyen MT, Kim K. Genetic convolutional neural network for intrusion detection systems. Future Generation Computer Systems. 2020 Dec 1;113:418-27.

Selvakumar B, Muneeswaran K. Firefly algorithm based feature selection for network intrusion detection. Computers & Security. 2019 Mar 1;81:148-55.

Ravale U, Marathe N, Padiya P. Feature selection based hybrid anomaly intrusion detection system using K means and RBF kernel function. Procedia Computer Science. 2015 Jan 1;45:428-35.

Liao HJ, Lin CH, Lin YC, Tung KY. Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications. 2013 Jan 1;36(1):16-24.

Doshi R, Apthorpe N, Feamster N. Machine learning ddos detection for consumer internet of things devices. In2018 IEEE Security and Privacy Workshops (SPW) 2018 May 24 (pp. 29-35). IEEE.

Revathi S, Malathi A. A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection. International Journal of Engineering Research & Technology (IJERT). 2013 Dec;2(12):1848-53

Mohammadi S, Mirvaziri H, Ghazizadeh-Ahsaee M, Karimipour H. Cyber intrusion detection by combined feature selection algorithm. Journal of information security and applications. 2019 Feb 1;44:80-8.

Sindhu SS, Geetha S, Kannan A. Decision tree based light weight intrusion detection using a wrapper approach. Expert Systems with applications. 2012 Jan 1;39(1):129-41.

Mebawondu JO, Alowolodu OD, Mebawondu JO, Adetunmbi AO. Network intrusion detection system using supervised learning paradigm. Scientific African. 2020 Sep 1;9:e00497.

Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.

Kalimuthan C, Renjit JA. Review on intrusion detection using feature selection with machine learning techniques. Materials Today: Proceedings. 2020 Jan 1;33:3794-802.

Kasongo SM, Sun Y. A deep learning method with wrapper based feature extraction for wireless intrusion detection system. Computers & Security. 2020 May 1;92:101752.

Larose DT, Larose CD. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons; 2014 Jul 8.

Quinlan JR. Induction of decision trees. Machine learning. 1986 Mar;1(1):81-106.

Schiller TW, Chen Y, El Naqa I, Deasy JO. Modeling radiation-induced lung injury risk with an ensemble of support vector machines. Neurocomputing. 2010 Jun 1;73(10-12):1861-7.

Subba B, Gupta P. A tfidfvectorizer and singular value decomposition based host intrusion detection system framework for detecting anomalous system processes. Computers & Security. 2021 Jan 1;100:102084.

Wang XY, Zhang HM, Gao HH. Quantum particle swarm optimization based network intrusion feature selection and detection. IFAC Proceedings Volumes. 2008 Jan 1;41(2):12312-7.

Dawson CW. Projects in computing and information systems: a student's guide. Pearson Education; 2005.