# Cyberattack Feature Selection using Correlation-Based Feature Selection Method in an Intrusion Detection System

*By* Deris Stiawan

WORD COUNT 5159
TIME SUBMITTED 31-AUG-2022 07:30PM
PAPER ID 90059828

# Cyberattack Feature Selection using Correlation-Based Feature Selection Method in an Intrusion Detection System


Ahmad Heryanto
Faculty of Engineering
Faculty of Computer Science
Universitas Sriwijaya
Palembang, Indonesia
hery@unsri.ac.id

Deris Stiawan*
Faculty of Computer Science
Universitas Sriwijaya
Palembang, Indonesia
deris@unsri.ac.id

Mohd Yazid Bin Idris
Faculty of Engineering
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
yazid@utm.my

Muhammad Robby Bahari
Faculty of Computer Science
Universitas Sriwijaya
Palembang, Indonesia
robbybahari2467@gmail.com

Agung Al Hafizin
Faculty of Computer Science
Universitas Sriwijaya
Palembang, Indonesia
agungalhafizin1@gmail.com

Rahmat Budiarto
College of Computer Science and Information Technology
Albaha University
Albaha, Saudi Arabia
rahmat@bu.edu.sa



**Abstract—An intrusion detection system (IDS) is software or hardware that works as a monitoring and defense system against cyberattacks. This system monitors computer systems or network activities that have the potential to violate security policies. In general, there are two techniques used by an IDS in its cyberattack detection system: signature-based and anomaly-based. However, these techniques still face some problems, such as false alarm warnings, low accuracy and precision rates, high-dimensional data, complex data structures, and long computational times. IDS performance can be improved by implementing feature selection, which can reduce the amount of data to be processed on the IDS detection engine. This research used correlation-based feature selection (CFS). Experimental results on CIC-IDS2018 dataset show optimal IDS performance. The proposed CFS-based IDS achieves an accuracy of 99.9995%, recall of 100%, specificity of 99.9985%, precision of 99.9992, F1-score of 99.9996%, true positive rate of 99.9992%, and true negative rate of 100%.**



**Keywords: Intrusion detection system, Feature selection, Correlation Based Feature Selection, Cyber Attack, CIC-IDS2018 dataset**


## I. Introduction

A cyberattack is any action that threatens the security of an information system, infrastructure, computer network, or personal computer device, the goal of which is to steal, transform, or destroy predetermined targets. In the related literatures [1]–[3], several groups of cyberattack threats have been found in information systems, including distributed denial of service (DDoS), malware, viruses, botnets/zombies, and scareware. According to [4], and [5], to protect information security from the threat of cyberattacks, network administrators use firewalls, antivirus, anti-malware, intrusion detection and prevention systems, virtual private networks, and more.

Over the past decade, there has been an increase in cyberattacks, and every year, many research institutes prove that cyberattacks pose a threat to information systems globally. According to the Acronis Cyberthreats Report 2022 [6] and the Check Point Cyber Security Report 2022 [7], in 2022, there was a more than 50% increase in cyberattacks against individuals, organizations, and companies when compared to 2020. However, despite numerous cases of cyberattacks occurring, only 20% of cybercrime victims report the incident [6]. Cyberattack patterns continue to evolve, and the various hacker equipment used is increasingly sophisticated and easy to use, making it possible for diverse groups of threat actors to commit cyberattacks. This condition must be quickly anticipated by various parties, so the mechanism for monitoring and detecting cyberattack patterns is a major component in efforts to prevent cyberattacks. Based on [8], there are two techniques used by cyberattack detection systems: signature-based and anomaly-based. However, such techniques still face some problems, such as high false alarm alerts, low accuracy and precision rates, high-dimensional data, complex data structures, and long computational times [9], [10].

Problems in intrusion detection systems (IDSs) can occur because the amount of data processed by the detection engine in IDS is exceptionally large and complex, making the analysis process difficult. Some of the features found in network data may cause more difficult for the analysis process to detect attack patterns. According to [11], and [12], IDS engines' performance can be improved by reducing the size of data to be processed on the detection engine. Feature reduction can be done via data filtering, data grouping, or feature selection [13].

Feature selection in IDSs can affect the performance of the detection system [14]. It can weigh the features with strong or low impact categories, which can then be decisive in data processing of the network to normal levels or attacks. Features not significantly impacted when detecting diverse types of attacks should be removed to obtain good IDS performance parameters. The removal of these features will improve the IDS performance in terms of calculations, dimensionality reductions, and time complexities.

The research in [15] identified that feature selection can reduce data dimensions by selecting the most important attributes, and feature selection algorithms can ignore features that are irrelevant in the IDS computation. Efficient dimensions for data processing will reduce the computation time and prevent the excessive use of computing resources. A wide variety of feature selection algorithms has been widely used in research, including intelligent patterns, fuzzy, deterministic algorithms, artificial neural networks, and swarm intelligence [15]–[18].

Several researchers [15], [19], [20] proposed a feature selection algorithm with a correlation-based feature selection (CFS) algorithm. The CFS algorithm can select the right data features, which are then used for the classification process with intelligent machine algorithms. The results of the feature selection algorithm in [15] reached an accuracy of around 99.9291%. Based on the above

research background, the success of an attack detection system can be determined by selecting the right features.

## II. INTRUSION DETECTION SYSTEM

An IDS is a software or hardware that works as a monitoring and defense system against cyberattacks. It monitors computer systems or network activities that have the potential to violate security policies, and it will generate a report to the management system and stop the violation. The IDS is installed as a sensor to detect any network traffic suspected of malicious activity [21]. This detection system can be grouped into several categories, as follows:

1. Detection system based on its placement.
2. Detection system based on the method.
3. Detection system based on the structure of its system.

### *A.* Detection system based on its placement

IDSs can be classified based on the infrastructure placement, consisting of two parts: host-based IDS (HIDS) and network-based IDS (NIDS) [22].

#### *1.* Host-based Intrusion Detection System

HIDS is a cyberattack detection system installed on computer devices that functions to monitor and analyze data packets from the operating system. The cyberattack detection system will work on a computer independently, and the system will analyze the data packets generated by the observed computer system [23].

#### *2.* Network-based Intrusion Detection System

NIDS is a cyberattack detection system used to monitor and analyze traffic on the network, and it works on computer networks. This system groups normal or abnormal data packets, as well as uses raw data packets that in and out on computer network interface. This system is widely used by computer network managers to monitor cyberattacks originating from outside the computer network owned by an individual, organization, or company. The advantage of this cyberattack detection system is that it can monitor, analyze, and protect computers contained on the owner's computer network [24].

A comparison between host-based and network-based IDS is provided in Table 1.

Table 1. HIDS vs. NIDS [25]

| HIDS | NIDS |
|---|---|
| Single Host | Multi-hosts |
| Useful to detect attacks from the internal network | Used to detect attack from outside the network |
| No understanding of packet headers | Check all packet data that enters the network |
| High false positive rate | Host-independent |
| Does not use other network host bandwidths | Requires good bandwidth |
| Computation load on host | Computation load on network |
| Attack detected when entering end device | Attack detected when entering network |
| Example: OSSEC [26], Samhain [27], SolarWinds Security Event Manager [28], ManageEngine EventLog Analyzer [29] | Example: Snort [30], Cisco Firepower [31] |

### *B.* Detection system based on the method

This IDS classification refers to how the IDS analyzes data packets that go through the network or host. This classification can be divided into two types: signature-based system and anomaly-based system.

#### 1. Signature-based System

Signature-based cyberattack detection system can model attack behavior and save it to an attack signature database. The cyberattack detection system will compare the database of signature attacks owned with the data packets on the network. In this signature technique, the attack detection system cannot detect cyberattacks if the attack is not inputted into the attack signature database [32].

#### 2. Anomaly-based System

Cyberattack detection system can detect attacks using normal or abnormal behavioral methods from log information or data traffic on the network. Abnormal behavior of the network system can be classified as an attack, because its characteristics differ from normal behavior. This technique can detect unknown attacks, but the detection mechanism with this capability might have the disadvantage of imperfect false positive (FP) alerts [33].

A comparison between signature-based and anomaly-based systems can be seen in Table 2.

Table 2. Signature vs. Anomaly based Systems on Detection Method [34]

| Signature-based Detection | Anomaly-based Detection |
|---|---|
| Low false alarm rate | Has better accuracy level for unknown attack behavior |
| Simple method | Low missing pattern rate |
| High detection and accuracy rate for known attack behaviors/patterns | Able to detect new vulnerabilities |
| Need update rule | Need training for attack pattern |
| No separation between detected attack experiments with attack performed by attacker | Low detection rate and high false alarm rate |
| Attack based on data | Able to detect new attacks |

### *C.* IDS Component

With regards to its components, IDS is divided into three parts, i.e.:

#### 1. Event Generator

Event generator is a major component of a cyberattack detection system and source of data, and it comes from the monitoring system. The monitoring bases used in IDSs are host-based, network-based, application-based, and target-based data [32].

#### 2. Analysis Engine

Analysis engine retrieves information from data sources and examines the data to check for possible attacks or other security-related access abuses [24]. The analysis methods used in intrusion detection systems are as follows:

a. Misuse/Signature-based Detection:
The signature-based analysis method uses datasets from known attacks and compares incoming and outcoming packets on the host/network with these databases. The main disadvantage of this method is that it simply looks

for patterns or signatures of known attacks and consequently ignores interference with the latest attacks that occurred in the unknown future. The signature-based detection system does not provide protection against zero-day attacks because it will fail to identify attacks when no attack signature is stored in the database. The following is the signature-based detection architecture described in Figure 1, which works by matching the incoming packets with the database signatures that have been stored. If there are similarities between incoming backets and the database, the system will be alerted [32].

b. Anomaly/Statistical-based Detection

Anomaly-based detection engine will search for something unexpected or unusual. With the help of statistical techniques or artificial intelligence, the cyberattack detection system analyzes the data flow in the system to identify abnormal activity patterns. Anomaly-based detection works by comparing the observed activity with the standard profile, which is normal behavior learned from the system that is monitored and developed while the IDS learns the environment. Anomaly-based methodologies can detect zero-day attacks without updates to the system [32]. In Figure 2, it is explained that anomaly-based detection will update the standard profile database if there is a new anomaly that is close to the standard profile, so this method is quite effective against zero-day attacks.

c. Response Manager

The response manager is a component that responds when unexpected behavior (possible intrusion attacks) is detected in the attack detection system. The response involves alerting a person or system that will overcome the detected attack [32].

## III. RESEARCH METHODOLOGY

Feature selection is performed to obtain key features in the detection process, where feature selection also serves to optimize the computation time when detecting cyberattacks. This research uses the CFS technique, which filters the incoming packet data to obtain important features suitable for the detection process. In the CFS method, important features are selected based on the correlation value of each feature with the class label. According to previous research works, a high correlation value of a characteristic indicates its significance. In this study, feature selection using the CFS method was chosen because this method can perform feature selection quite well and provide high detection efficiency.

The dataset for developing an IDS model is the CSE-CIC-IDS 2018 dataset, a pcap file format dataset that consists of normal traffic and attack traffic. The features of the dataset are shown in Table 3.

Table 3. Features on the CSE-CIC-IDS2018 dataset

| No | Feature Name | Description |
|---|---|---|
| 1 | fl_dur | Flow duration |
| 2 | tot_fw_pk | Total packets in the forward direction |
| 3 | tot_bw_pk | Total packets in the backward direction |
| 4 | tot_l_fw_pkt | Total size of packet in forward direction |
| 5 | fw_pkt_l_max | Maximum size of packet in forward direction |
| 6 | fw_pkt_l_min | Minimum size of packet in forward direction |
| 7 | fw_pkt_l_avg | Average size of packet in forward direction |
| 8 | fw_pkt_l_std | Standard deviation size of packet in forward direction |
| 9 | Bw_pkt_l_max | Maximum size of packet in backward direction |
| 10 | Bw_pkt_l_min | Minimum size of packet in backward direction |
| 11 | Bw_pkt_l_avg | Mean size of packet in backward direction |
| 12 | Bw_pkt_l_std | Standard deviation size of packet in backward direction |
| 13 | fl_byt_s | flow byte rate that is number of packets transferred per second |
| 14 | fl_pkt_s | flow packets rate that is number of packets transferred per second |
| 15 | fl_iat_avg | Average time between two flows |
| 16 | fl_iat_std | Standard deviation time two flows |
| 17 | fl_iat_max | Maximum time between two flows |
| 18 | fl_iat_min | Minimum time between two flows |
| 19 | fw_iat_tot | Total time between two packets sent in the forward direction |
| 20 | fw_iat_avg | Mean time between two packets sent in the forward direction |
| 21 | fw_iat_std | Standard deviation time between two packets sent in the forward direction |
| 22 | fw_iat_max | Maximum time between two packets sent in the forward direction |
| 23 | fw_iat_min | Minimum time between two packets sent in the forward direction |
| 24 | bw_iat_tot | Total time between two packets sent in the backward direction |
| 25 | bw_iat_avg | Mean time between two packets sent in the backward direction |
| 26 | bw_iat_std | Standard deviation time between two packets sent in the backward direction |
| 27 | bw_iat_max | Maximum time between two packets sent in the backward direction |
| 28 | bw_iat_min | Minimum time between two packets sent in the backward direction |
| 29 | fw_psh_flag | Number of times the PSH flag was set in packets travelling in the forward direction (0 for UDP) |
| 30 | bw_psh_flag | Number of times the PSH flag was set in packets travelling in the backward direction (0 for UDP) |
| 31 | fw_urg_flag | Number of times the URG flag was set in packets travelling in the forward direction (0 for UDP) |
| 32 | bw_urg_flag | Number of times the URG flag was set in packets travelling in the backward direction (0 for UDP) |
| 33 | fw_hdr_len | Total bytes used for headers in the forward direction |
| 34 | bw_hdr_len | Total bytes used for headers in the forward direction |
| 35 | fw_pkt_s | Number of forward packets per second |
| 36 | bw_pkt_s | Number of backward packets per second |
| 37 | pkt_len_min | Minimum length of a flow |
| 38 | pkt_len_max | Maximum length of a flow |
| 39 | pkt_len_avg | Mean length of a flow |
| 40 | pkt_len_std | Standard deviation length of a flow |
| 41 | pkt_len_va | Minimum inter-arrival time of packet |
| 42 | fin_cnt | Number of packets with FIN |
| 43 | syn_cnt | Number of packets with SYN |
| 44 | rst_cnt | Number of packets with RST |
| 45 | pst_cnt | Number of packets with PUSH |
| 46 | ack_cnt | Number of packets with ACK |
| 47 | urg_cnt | Number of packets with URG |
| 48 | cwe_cnt | Number of packets with CWE |
| 49 | ece_cnt | Number of packets with ECE |
| 50 | down_up_ratio | Download and upload ratio |
| 51 | pkt_size_avg | Average size of packet |
| 52 | fw_seg_avg | Average size observed in the forward direction |
| 53 | bw_seg_avg | Average size observed in the backward direction |

| | | |
|---|---|---|
| 54 | fw_byt_blk_avg | Average number of bytes bulk rate in forward direction |
| 55 | fw_pkt_blk_avg | Average number of packets bulk rate in the forward direction |
| 56 | fw_blk_rate_avg | Average number of bulk rate in the forward direction |
| 57 | bw_byt_blk_avg | Average number of bytes bulk rate in the backward direction |
| 58 | bw_pkt_blk_avg | Average number of packets bulk rate in the backward direction |
| 59 | bw_blk_rate_avg | Average number of bulk rate in the backward direction |
| 60 | subfl_fw_pk | The average number of packets in a sub flow in the forward direction |
| 61 | subfl_fw_byt | The average number of bytes in a sub flow in the forward direction |
| 62 | subfl_bw_pkt | The average number of packets in a sub flow in the backward direction |
| 63 | subfl_bw_byt | The average number of bytes in a sub flow in the backward direction |
| 64 | fw_win_byt | Number of bytes sent in initial window in the forward direction |
| 65 | bw_win_byt | # of bytes sent in initial window in the backward direction |
| 66 | Fw_act_pkt | # of packets with at least 1 byte of TCP data payload in the forward direction |
| 67 | fw_seg_min | Minimum segment size observed in the forward direction |
| 68 | atv_avg | Mean time a flow was active before becoming idle |
| 69 | atv_std | Standard deviation time a flow was active before becoming idle |
| 70 | atv_max | Maximum time a flow was active before becoming idle |
| 71 | atv_min | Minimum time a flow was active before becoming idle |
| 72 | idl_avg | Mean time a flow was idle before becoming active |
| 73 | idl_std | Standard deviation time a flow was idle before becoming active |
| 74 | idl_max | Maximum time a flow was idle before becoming active |
| 75 | idl_min | Minimum time a flow was idle before becoming active |

When feature selection is performed using CFS, the data is first reduced and then data imbalance processing is performed using the Synthetic Minority Oversampling Tcchnique (SMOTE) algorithm. Using CFS, this research identified the best components to be further classified. A flowchart of CFS can be seen in Figure 3.

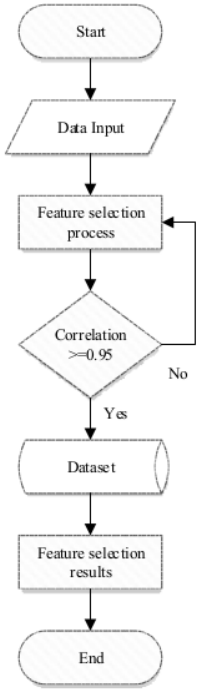


Figure 3. Feature selection process with CFS

## IV. HASIL PENELITIAN

In this research, the CSE-CIC-IDS 2018 dataset used 8,284,196 data rows in total. In the dataset scenario are several groups of d including normal and attack data. The attack data consists of DDoS, denial of service (DoS), Botnet, uteForce, infiltration, and web attacks. Statistics of t ataset based on the total amount of normal data and attack data are shown in Table 5.

Table 5. CSE-CICIDS2018 Dataset

| Dataset | Normal | DDoS | Dos | Botnet | Brute Force | Infiltration | Web Attacks | Total |
|---|---|---|---|---|---|---|---|---|
| CSE-CICIDS 2018 | 6,112,151 (74%) | 687,742 (8%) | 654,301 (8%) | 286,191 (3%) | 380,949 (4%) | 161,934 (2%) | 928 (0.01%) | 8,284,196 |

Table 6. Feature Selection Results

| No | Attribute | No | Attribute | No | Attribute | No | Attribute | No | Attribute |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Dst Port | 11 | Bwd Pkt Len Max | 21 | Bwd IAT Max | 31 | ACK Flag Cnt | 41 | Init Fwd Win Byts |
| 2 | Protocol | 12 | Bwd Pkt Len Min | 22 | Bwd IAT Min | 32 | URG Flag Cnt | 42 | Init Bwd Win Byts |
| 3 | Flow Duration | 13 | Bwd Pkt Len Mean | 23 | Fwd PSH Flags | 33 | CWE Flag Count | 43 | Fwd Act Data Pkts |
| 4 | Tot Fwd Pkts | 14 | Flow Byts/s | 24 | Bwd PSH Flags | 34 | Down/Up Ratio | 44 | Fwd Seg Size Min |
| 5 | Tot Bwd Pkts | 15 | Flow Pkts/s | 25 | Fwd URG Flags | 35 | Fwd Byts/b Avg | 45 | Active Mean |
| 6 | TotLen Fwd Pkts | 16 | Flow IAT Mean | 26 | Bwd URG Flags | 36 | Fwd Pkts/b Avg | 46 | Active Std |
| 7 | Fwd Pkt Len Max | 17 | Flow IAT Max | 27 | Pkt Len Var | 37 | Fwd Blk Rate Avg | 47 | Active Max |
| 8 | Fwd Pkt Len Min | 18 | Flow IAT Tot | 28 | FIN Flag Cnt | 38 | Bwd Byts/b Avg | 48 | Idle Min |
| 9 | Fwk Pkt Len Mean | 19 | Bwd IAT Mean | 29 | RST Flag Cnt | 39 | Bwd Pkts/b Avg | 49 | Label |
| 10 | Fwd Pkt Len Std | 20 | Bwd IAT Std | 30 | PSH Flag Cnt | 40 | Bwd Blk Rate Avg | | |

Table 7. Validation performance

| Ratio | Training | Testing | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | Accuracy | Recall | Sspecificity | Precision | F1-Score | TP | TN |
| 10:90 | 99.99% | 99.9982% | 99.9997% | 99.9951% | 99.9975% | 99.9986% | 99.9975% | 99.9995% |
| 20:80 | 99.99% | 99.9973% | 99.9982% | 99.9955% | 99.9977% | 99.9979% | 99.9977% | 99.9964% |
| 30:70 | 99.99% | 99.9992% | 100% | 99.9978% | 99.9989% | 99.9994% | 99.9989% | 100% |
| 40:60 | 99.99% | 99.9991% | 100% | 99.9975% | 99.9987% | 99.9993% | 99.9987% | 100% |
| 50:50 | 99.99% | 99.9994% | 100% | 99.9982% | 99.9991% | 99.9995% | 99.9991% | 100% |
| 60:40 | 99.99% | 99.9995% | 100% | 99.9985% | 99.9992% | 99.9996% | 99.9992% | 100% |
| 70:30 | 99.99% | 99.9995% | 100% | 99.9985% | 99.9992% | 99.9996% | 99.9992% | 100% |
| 80:20 | 99.99% | 99.9995% | 100% | 99.9985% | 99.9992% | 99.9996% | 99.9992% | 100% |
| 90:10 | 99.99% | 99.9995% | 100% | 99.9985% | 99.9992% | 99.9996% | 99.9992% | 100% |

In a further step, tuning is performed against the correlation value to choose the best feature. In the tuning process, a correlation value of 9.5 is determined. Therefore, after the tuning process is successfully performed, the most relevant features will be obtained in the CIC-IDS2018 dataset, and they will be used in the cyberattack detection process. The number of features that have been successfully obtained is 48. The result of this tuning process is the best feature to develop intelligent machine models by using LSTM. Figure 4 shows a correlation graph of the feature selection process that has been used, and Table 6 identifies the features obtained from the feature selection process using the CFS algorithm.

Research validation uses a confusion matrix based on true positive (TP), true negative (TN), FP, and false negative (FN) parameters. The success rate of the cyberattack detection system experiment was determined based on the values of accuracy, recall, specificity, precision, and F1-score. The experiments performed to obtain the most optimal value are shown in Table 7.

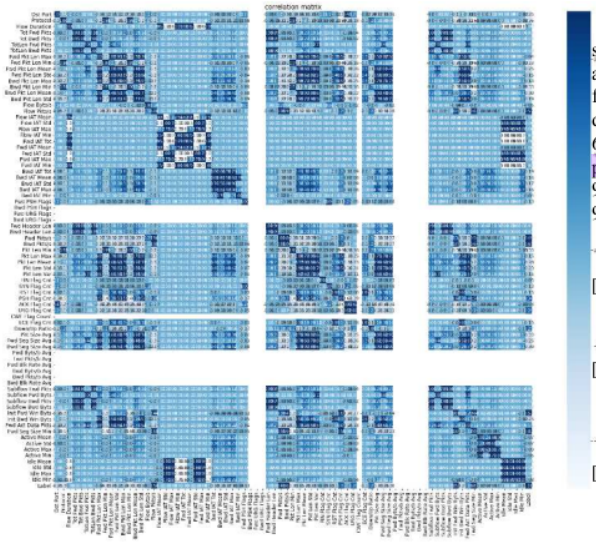

Figure 4. Correlation graph of the dataset

The validation result with the scenario is then visualized in the form of a graph based on the number of data comparisons, as shown in Figure 5.

From Table 7 and Figure 5, it can be concluded that the scenario with 60% training data and 40% testing data up to the scenario with 90% training data and 10% testing data achieved the highest results compared to other scenarios, and scenarios with 20% training data and 80% testing data have the lowest performance results compared to the other scenario. In Table 7 and Figure 6, it can be seen that the scenario with 60% training data and 40% testing data up to 90% training data and 10% testing data achieved the highest accuracy, F1-Score, and recall results compared to other scenarios, while the scenario with 90% training data and 10% testing data achieved the highest sensitivity results compared to the other scenario.

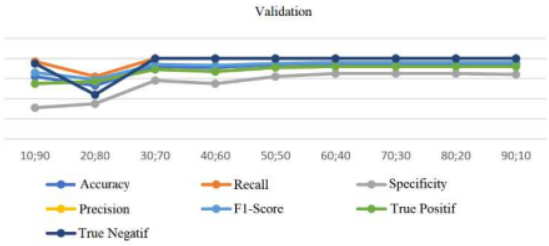


Figure 5. Graph of Validation Results

## V. SUMMARY

From the experimental results carried out in this research, several conclusions were made (i) CFS obtained the best features and can reduce computational time in the detection process. The feature selection results generated 48 best features used for the detection process, (ii) Optimal performance is achieved for 60%:40% testing data and training data ratio. The best performance metrics are 99.9995% accuracy, 100% recall, 99.9985% specificity, 99.9992 precision, 99.9996% F1-score, 99.9992% TP, and 100% TN.

## REFERENCES


[1] W. A. Al-Khater, S. Al-Maadeed, A. A. Ahmed, A. S. Sadiq, and M. K. Khan, "Comprehensive review of cybercrime detection techniques," *IEEE Access*, vol. 8, pp. 137293–137311, 2020, doi: 10.1109/ACCESS.2020.3011259.

[2] V. Kanimozhi and T. P. Jacob, "Artificial Intelligence based Network Intrusion Detection with hyper-parameter optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing," *ICT Express*, vol. 5, no. 3, pp. 211–214, 2019, doi: 10.1016/j.icte.2019.03.003.

[3] X. Clotet, J. Moyano, and G. León, "A real-time anomaly-based IDS for cyber-attack detection at the industrial process level of Critical Infrastructures," *Int. J. Crit. Infrastruct. Prot.*, vol. 23, pp. 11–20, 2018, doi: 10.1016/j.ijcip.2018.08.002.

[4] A. Sawant, "A Comparative Study of Different Intrusion Prevention Systems," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1–5, doi: 10.1109/ICCUBEA.2018.8697500.

[5] Q. R. S. Fitni and K. Ramli, "Implementation of Ensemble Learning and Feature Selection for Performance Improvements in Anomaly-Based Intrusion Detection Systems," in *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 2020, pp. 118–124, doi: 10.1109/IAICT50021.2020.9172014.

[6] AcronisTM, "Acronis Cyberthreats Report 2022 unveils cyberthreat predictions," 2022. https://www.acronis.com/id-id/blog/posts/acronis-cyberthreats-report-2022-unveils-cyberthreat-predictions/ (accessed May 26, 2022).

[7] M. Horowitz and J. Fischbein, "Cyber security report 2022," *Chack Point Research*, 2022.

[8] I. F. Kilincer, F. Ertam, and A. Sengur, "A comprehensive intrusion detection framework using boosting algorithms," *Comput. Electr. Eng.*, vol. 100, p. 107869, May 2022, doi: 10.1016/J.COMPELECENG.2022.107869.

[9] T. Girdler and V. G. Vassilakis, "Implementing an intrusion detection and prevention system using Software-Defined Networking: Defending against ARP spoofing attacks and Blacklisted MAC Addresses," *Comput. Electr. Eng.*, vol. 90, no. July 2020, p. 106990, 2021, doi: 10.1016/j.compeleceng.2021.106990.

[10] A. Patel, M. Taghavi, K. Bakhtiyari, and J. Celestino Júnior, "An intrusion detection and prevention system in cloud computing: A systematic review," *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 25–41, 2013, doi: 10.1016/j.jnca.2012.08.007.

[11] P. Nimbalkar and D. Kshirsagar, "Feature selection for intrusion detection system in Internet-of-Things (IoT)," *ICT Express*, vol. 7, no. 2, pp. 177–181, Jun. 2021, doi: 10.1016/J.ICTE.2021.04.012.

[12] V. Herrera-Semenets, L. Bustio-Martínez, R. Hernández-León, and J. van den Berg, "A multi-measure feature selection algorithm for efficacious intrusion detection," *Knowledge-Based Syst.*, vol. 227, p. 107264, Sep. 2021, doi: 10.1016/J.KNOSYS.2021.107264.

[13] F. Chen, Z. Ye, C. Wang, L. Yan, and R. Wang, "A feature selection approach for network intrusion detection based on tree-seed algorithm and k-nearest neighbor," *Proc. 2018 IEEE 4th Int. Symp. Wirel. Syst. within Int. Conf. Intell. Data Acquis. Adv. Comput. Syst. IDAACS-SWS 2018*, pp. 68–72, Nov. 2018, doi: 10.1109/IDAACS-SWS.2018.8525522.

[14] D. Putra and I. G. A. G. A. Kadnyanana, "Implementation of Feature Selection using Information Gain Algorithm and Discretization with NSL-KDD Intrusion Detection System," *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 9, no. 3, p. 359, 2021, doi: 10.24843/jlk.2021.v09.i03.p06.

[15] C. Kalimuthan and J. Arokia Renjit, "Review on intrusion detection using feature selection with machine learning techniques," *Mater. Today Proc.*, vol. 33, pp. 3794–3802, Jan. 2020, doi: 10.1016/J.MATPR.2020.06.218.

[16] F. Amiri, M. Rezaei Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," *J. Netw. Comput. Appl.*, vol. 34, no. 4, pp. 1184–1199, Jul. 2011, doi: 10.1016/J.JNCA.2011.01.002.

[17] B. Deore and S. Bhosale, "Intrusion Detection System Based on RNN Classifier for Feature Reduction," *SN Comput. Sci. 2021 32*, vol. 3, no. 2, pp. 1–9, Dec. 2021, doi: 10.1007/S42979-021-00991-0.

[18] S. Gavel, A. S. Raghuvanshi, and S. Tiwari, "Maximum correlation based mutual information scheme for intrusion detection in the data networks," *Expert Syst. Appl.*, vol. 189, p. 116089, Mar. 2022, doi: 10.1016/J.ESWA.2021.116089.

[19] T. Ahmad and M. N. Aziz, "Data preprocessing and feature selection for machine learning intrusion detection systems," *ICIC Express Lett.*, vol. 13, no. 2, pp. 93–101, 2019, doi: 10.24507/icicel.13.02.93.

[20] K. O. Akande, T. O. Owolabi, and S. O. Olatunji, "Investigating the effect of correlation-based feature selection on the performance of support vector machines in reservoir characterization," *J. Nat. Gas Sci. Eng.*, vol. 22, pp. 515–522, Jan. 2015, doi: 10.1016/J.JNGSE.2015.01.007.

[21] H. J. Liao, C. H. Richard Lin, Y. C. Lin, and K. Y. Tung, "Intrusion detection system: A comprehensive review," *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 16–24, Jan. 2013, doi: 10.1016/J.JNCA.2012.09.004.

[22] J. Liu, K. Xiao, L. Luo, Y. Li, and L. Chen, "An intrusion detection system integrating network-level intrusion detection and host-level intrusion detection," *Proc. - 2020 IEEE 20th Int. Conf. Softw. Qual. Reliab. Secur. QRS 2020*, pp. 122–129, Dec. 2020, doi: 10.1109/QRS51102.2020.00028.

[23] C. Vargas Martinez and B. Vogel-Heuser, "A Host Intrusion Detection System architecture for embedded industrial devices," *J. Franklin Inst.*, vol. 358, no. 1, pp. 210–236, Jan. 2021, doi: 10.1016/J.JFRANKLIN.2019.03.037.

[24] S. Kumar, S. Gupta, and S. Arora, "Research Trends in Network-Based Intrusion Detection Systems: A Review," *IEEE Access*, vol. 9, pp. 157761–157779, 2021, doi: 10.1109/ACCESS.2021.3129775.

[25] M. Bharati and S. Tamane, "Intrusion detection systems (IDS) & future challenges in cloud based environment," in *Proceedings - 1st International Conference on Intelligent Systems and Information Management, ICISIM 2017*, Nov. 2017, vol. 2017-January, pp. 240–250, doi: 10.1109/ICISIM.2017.8122180.

[26] OSSEC, "OSSEC - World's Most Widely Used Host Intrusion Detection System - HIDS." https://www.ossec.net/ (accessed May 11, 2021).

[27] Samhain Labs, "Samhain Labs | samhain." https://www.la-samhna.de/samhain/ (accessed May 11, 2021).

[28] SolarWinds, "Security Event Manager - View Event Logs Remotely | SolarWinds." https://www.solarwinds.com/security-event-manager (accessed May 11, 2021).

[29] ManageEngine, "EventLog Analyzer - SIEM Log management software." https://www.manageengine.com/products/eventlog/ (accessed May 11, 2021).

[30] Snort, "Snort - Network Intrusion Detection & Prevention System." https://www.snort.org/ (accessed May 11, 2021).

[31] Cisco Firepower, "Cisco Firepower 9300 - Cisco." https://www.cisco.com/c/en/us/products/security/firepower-9000-series/index.html (accessed May 11, 2021).

[32] D. Mudzingwa and R. Agrawal, "A study of methodologies used in intrusion detection and prevention systems (IDPS)," *Conf. Proc. - IEEE SOUTHEASTCON*, 2012, doi: 10.1109/SECon.2012.6197080.

[33] H. Bostani and M. Sheikhan, "Hybrid of anomaly-based and specification-based IDS for Internet of Things using unsupervised OPF based on MapReduce approach," *Comput. Commun.*, vol. 98, pp. 52–71, 2017, doi: 10.1016/j.comcom.2016.12.001.

[34] S. Karde, "Intrusion Detection and Anomaly Detection System Using Sequential Pattern Mining," *Int. J. Res. Eng. Technol.*, vol. 05, no. 08, pp. 154–160, 2016, doi: 10.15623/ijret.2016.0508029.

[35] W. Hu, J. Gao, Y. Wang, O. Wu, and S. Maybank, "Online adaboost-based parameterized methods for dynamic distributed network intrusion detection," *IEEE Trans. Cybern.*, vol. 44, no. 1, pp. 66–82, Jan. 2014, doi: 10.1109/TCYB.2013.2247592.

[36] M. Bijone, "A Survey on Secure Network: Intrusion Detection & Prevention Approaches," *Am. J. Inf. Syst.*, vol. 4, no. 3, pp. 69–88, 2016, doi: 10.12691/ajis-4-3-2.

# Cyberattack Feature Selection using Correlation-Based Feature Selection Method in an Intrusion Detection System

ORIGINALITY REPORT

34%
SIMILARITY INDEX

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | www.warse.org<br>Internet | 474 words — 12% |
| 2 | www.unb.ca<br>Internet | 312 words — 8% |
| 3 | M. Agus Syamsul Arifin, Deris Stiawan, Susanto, Juli Rejito, Mohd. Yazid Idris, Rahmat Budiarto. "Denial of Service Attacks Detection on SCADA Network IEC 60870-5-104 using Machine Learning", 2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2021<br>Crossref | 35 words — 1% |
| 4 | gist.github.com<br>Internet | 29 words — 1% |
| 5 | journalofbigdata.springeropen.com<br>Internet | 27 words — 1% |
| 6 | unbscholar.lib.unb.ca<br>Internet | 25 words — 1% |
| 7 | open.metu.edu.tr<br>Internet | 24 words — 1% |
| 8 | dspace.cuni.cz | |

Internet

23 words — 1%

9 Juan Sebastián Rojas, Álvaro Rendón Gallón, Juan Carlos Corrales. "Chapter 37 Personalized Service Degradation Policies on OTT Applications Based on the Consumption Behavior of Users", Springer Science and Business Media LLC, 2018

Crossref

22 words — 1%

10 Romi Fadillah Rahmat, Tri Ramadhani, Dani Gunawan, Sharfina Faza, Rahmat Budiarto. "Mel-frequency Cepstral Coefficient-Vector Quantization Implementation for Voice Detection of Rice-Eating Birds in The Rice Fields", 2018 Third International Conference on Informatics and Computing (ICIC), 2018

Crossref

19 words — < 1%

11 "Deployable Machine Learning for Security Defense", Springer Science and Business Media LLC, 2020

Crossref

18 words — < 1%

12 "ACIT 2021 Conference Proceedings", 2021 22nd International Arab Conference on Information Technology (ACIT), 2021

Crossref

16 words — < 1%

13 Saparudin, Ade Kurniawan. "Harmony search algorithm with dynamic pitch adjustment rate and fret width for image compression", 2016 Asia Pacific Conference on Multimedia and Broadcasting (APMediaCast), 2016

Crossref

16 words — < 1%

14 Napsiah Amelia Putri, Deris Stiawan, Ahmad Heryanto, Tri Wanda Septian, Lelyzar Siregar,

14 words — < 1%

Rahmat Budiarto. "Denial of service attack visualization with clustering using K-means algorithm", 2017 International Conference on Electrical Engineering and Computer Science (ICECOS), 2017
Crossref

15 Liang Fu Lu, Zheng-Hai Huang, Mohammed A. Ambusaidi, Kui-Xiang Gou. "A Large-Scale Network Data Analysis via Sparse and Low Rank Reconstruction", Discrete Dynamics in Nature and Society, 2014 — 13 words — < 1%
Crossref

16 www.coursehero.com — 13 words — < 1%
Internet

17 www.mdpi.com — 11 words — < 1%
Internet

18 www.nature.com — 11 words — < 1%
Internet

19 "Security, Privacy, and Anonymity in Computation, Communication, and Storage", Springer Science and Business Media LLC, 2019 — 10 words — < 1%
Crossref

20 Mehdi Barati, Azizol Abdullah, Nur Izura Udzir, Ramlan Mahmod, Norwati Mustapha. "Distributed Denial of Service detection using hybrid machine learning technique", 2014 International Symposium on Biometrics and Security Technologies (ISBAST), 2014 — 10 words — < 1%
Crossref

21 jfas.info — 10 words — < 1%
Internet

22 strathprints.strath.ac.uk
Internet

10 words — < 1%

23 webthesis.biblio.polito.it
Internet
10 words — < 1%

24 Arijit Chandra, Sunil Kumar Khatri, Rajbala Simon. "Filter-based Attribute Selection Approach for Intrusion Detection using k-Means Clustering and Sequential Minimal Optimization Techniq", 2019 Amity International Conference on Artificial Intelligence (AICAI), 2019
Crossref
9 words — < 1%

25 Mohammad Al-Fawa'reh, Mustafa Al-Fayoumi, Shadi Nashwan, Salam Fraihat. "Cyber threat intelligence using PCA-DNN model to detect abnormal network behavior", Egyptian Informatics Journal, 2021
Crossref
9 words — < 1%

26 pdfs.semanticscholar.org
Internet
9 words — < 1%

27 scholar.afit.edu
Internet
9 words — < 1%

28 www.ncbi.nlm.nih.gov
Internet
9 words — < 1%

29 www.researchgate.net
Internet
9 words — < 1%

30 Ambusaidi, Mohammed A., Zhiyuan Tan, Xiangjian He, Priyadarsi Nanda, Liang Fu Lu, and Aruna Jamdagni. "Intrusion detection method based on nonlinear correlation measure", International Journal of Internet Protocol Technology, 2014.
Crossref
8 words — < 1%

31 Erdem Tuncer, Emine Doğru Bolat. "Channel based epilepsy seizure type detection from electroencephalography (EEG) signals with machine learning techniques", Biocybernetics and Biomedical Engineering, 2022
Crossref
8 words — < 1%

32 Fargana J. Abdullayeva. "Distributed denial of service attack detection in E-government cloud via data clustering", Array, 2022
Crossref
8 words — < 1%

33 Meng, Yuxin, Wenjuan Li, and Lam-For Kwok. "Towards adaptive false alarm reduction using Cloud as a Service", 2013 8th International Conference on Communications and Networking in China (CHINACOM), 2013.
Crossref
8 words — < 1%

34 Muraleedharan N., Janet B.. "SCAFFY", International Journal of Information Security and Privacy, 2021
Crossref
8 words — < 1%

35 Ola Salman, Louma Chaddad, Imad H. Elhajj, Ali Chehab, Ayman Kayssi. "Pushing intelligence to the network edge", 2018 Fifth International Conference on Software Defined Systems (SDS), 2018
Crossref
8 words — < 1%

36 Ruizhe Zhao, Yingxue Mu, Long Zou, Xiumei Wen. "A Hybrid Intrusion Detection System Based on Feature Selection and Weighted Stacking Classifier", IEEE Access, 2022
Crossref
8 words — < 1%

37 Sugandh Seth, Kuljit Kaur, Gurwinder Singh. "A Novel Ensemble Framework for an Intelligent Intrusion Detection System", IEEE Access, 2021
Crossref
8 words — < 1%

| | | |
|---|---|---|
| 38 | fcc08321-8158-469b-b54d-f591e0bd3df4.filesusr.com<br>Internet | 8 words — < 1% |
| 39 | ia600108.us.archive.org<br>Internet | 8 words — < 1% |
| 40 | section.iaesonline.com<br>Internet | 8 words — < 1% |
| 41 | www.testmagzine.biz<br>Internet | 8 words — < 1% |
| 42 | "Data Mining and Big Data", Springer Science and Business Media LLC, 2021<br>Crossref | 7 words — < 1% |
| 43 | Ramin Atefinia, Mahmood Ahmadi. "Network intrusion detection using multi-architectural modular deep neural network", The Journal of Supercomputing, 2020<br>Crossref | 7 words — < 1% |
| 44 | Merve Ozkan-Okay, Refik Samet, Omer Aslan, Deepti Gupta. "A Comprehensive Systematic Literature Review on Intrusion Detection Systems", IEEE Access, 2021<br>Crossref | 6 words — < 1% |
| 45 | Sugandh Seth, Gurvinder Singh, Kuljit Kaur Chahal. "A novel time efficient learning-based approach for smart intrusion detection system", Journal of Big Data, 2021<br>Crossref | 6 words — < 1% |

| | | | |
|---|---|---|---|
| EXCLUDE QUOTES | ON | EXCLUDE SOURCES | OFF |
| EXCLUDE BIBLIOGRAPHY | ON | EXCLUDE MATCHES | OFF |