Training and Classification Techniques in Intrusion Detection Systems Based on Network Anomalies Comparative Study

Johan Mardini-Bovea1, Dixon Salcedo1, Issac Nagles-Pozo1, Yadira Quiñonez2, and Jezreel Mejía3

1 Department of Computer Science and Electronics, University of the Coast, Barranquilla 080020, Colombia

2 Facultad de Informática Mazatlán, Universidad Autónoma de Sinaloa, Mazatlán, 82000, México

2 Centro de Investigación en Matemáticas, Zacatecas Unit, Guanajuato, GJ, Mexico

[jmardini@cuc.edu.co](mailto:jmardini@cuc.edu.co), [dsalcedo2@cuc.edu.co](mailto:dsalcedo2@cuc.edu.co), [inagles1@cuc.edu.co](mailto:inagles1@cuc.edu.co), [yadiraqui@uas.edu.mx](mailto:yadiraqui@uas.edu.mx), jmejia@cimat.mx

**Abstract.** Computer network security is vital due to the large volume of data handled. One of the security tools available to large companies is Intrusion Detection Systems (IDS). However, the increase in information and communication technologies has triggered a growth in intrusive accesses and attacks directed at computer systems. This situation has increased over the years, highlighting the vulnerability of such systems. The primary motivation of this research has been implementing the wrapper method applied to IDSs in different training and classification techniques to identify the best intrusion detection model to improve attack detection rates in computer network systems, using a feature selection procedure and other methods of unsupervised training algorithms. In this research, different metrics that measure the quality of the proposed intrusion detection model are evaluated through simulation processes using the DARPA NSL-KDD dataset and by applying the INFO.GAIN feature selection technique to identify the most relevant features in the classification process. Furthermore, an unsupervised learning algorithm (GHSOM, RANDOM FOREST, BAYESIAN NETWORKS, NAIVE BAYES, C4.5, LOGISTIC, PART, AND NBTREE) was trained to classify the bi-class traffic automatically.

**Keywords:** Dataset DARPA NSL-KDD, IDS, Training and classification techniques, Feature selection techniques.

1. Introduction

Entities must protect the information they store those transits through their computer networks. Considering the above problems, different systems have emerged to detect and protect data against malware that could cause the loss or deterioration of information. However, these systems may not be effective in network attack detection processes when the malware database is not updated with a specific frequency. Given that new attacks are created more frequently, this can generate many vulnerabilities over time.

IDSs that identify malicious traffic on the network for subsequent blocking and documentation to counteract the attacker's actions have been developed to overcome these drawbacks. IDSs can detect attacks with a signature-based methodology (comparing attacks with a signature database or rules) or with an anomaly-based method (using a learning algorithm), the former has been widely implemented in commercial IDSs, and in free software (Snort and Prelude), however, they do not detect new attacks; the latter see further attacks with a certain percentage of accuracy.

To accurately identify the magnitude of the problem and the possible solution alternatives, the following must be addressed in detail: the fundamentals of Intrusion Detection Systems, the inherent characteristics of DARPA datasets, the existing techniques or algorithms related to feature selection and extraction, and training and data classification techniques, such as Artificial Neural Networks (ANN) and Bayesian Networks (BAYES -NET), among others.

As a result of the experiments carried out, the researchers have detected that a variable that directly affects the efficiency of the learning algorithm is the identification of the features to be evaluated during the preprocessing phase since the choice of all the features or some of them that are not appropriate will generate long computational response times, negatively affecting the final evaluation of the learning algorithm.

The proposed model trains an artificial intelligence technique that automatically performs the data flow classification process. Such a neural network can identify the type of traffic, regardless of whether new types of attacks are generated. Several simulation scenarios were implemented to validate the model, comprising three phases (training, classification, and metrics calculation).

Initially, the correlation analysis technique was applied to the KDD DARPA-Train dataset to eliminate the features of less relevance and the INFO.GAIN feature selection technique was applied to identify the optimal number of features to categorize them in order of relevance and then train the neural network with the previously selected features.

In the classification phase, the cross-validation technique was applied to 10 folds using the previously mentioned dataset (KDDD DARPA Train), where the correlation analysis techniques and the INFO.GAIN feature selection techniques are involved, and finally, the data are classified based on the map generated in the training process and on the new subset of data. In the final phase, different performance metrics were calculated (sensitivity, specificity, precision, and accuracy), which allowed us to determine the efficiency of the proposed model.

The structure of this article is as follows; Section 2 describes the literature review, Section 3 describes the simulation process applied to IDS, Section 4 describes the dataset used in intrusion detection systems, Section 5 describes the feature selection technique applied in the proposed IDS system, Section 6 describes data mining techniques as a method for developing network anomaly detection systems, Section 7 describes the simulation results. Section 8 presents conclusions, and Section 9 describes some future work.

1. Literature Review
   1. Computer Security

The first generalized definition of computer security is protection against unexpected behavior based on a set of procedures and technologies to prevent intrusion [1].

According to [2, 3], computer security is the fulfillment of confidentiality, integrity, and availability in a computer system based on a series of conceptual elements that need to be detailed better to understand the topics under study [4].

According to the above concepts, we can infer that computer security consists of ensuring that the resources of the information system (computer hardware or software) of an organization are used in the way it was decided. That access to the information contained therein and its modification is only possible to persons who are accredited and within the limits of their authorization [5 - 7].

* 1. Taxonomy of Computer Attacks and Intrusions

Taxonomy provides prior knowledge for new attacks and a structured way to study these attacks. Many works can be found concerning categorizing and classifying computer attacks and intrusions [8]. In [9] propose a model, the computer attacks contained in the DARPA DATASET are classified into four (4) categories:

### **Denial of Service:**

Also known by its acronym Denial of Service (DoS), it is a set of attacks that lead to stopping the operation of a network, machine, process, or service to authorized users due to the overload of the victim's computational resources.

### **Remote to Local (R2L):**

According to [10], it originates when a computer attacker who does not have access to a machine gains access to that machine either as a standard user or root, using some intrusion method by software.

### **User to Root (U2R):**

It is generated when an attacker, who has an account in a computer system, acquires privileges superior to those initially established without the IT administrator's authorization, executing some intrusion technique based on a specific computer system vulnerability.

### **Probing:**

According to [11], they are a set of attacks that are characterized by probing the victim's network to collect necessary information from the hosts that include it without being detected, providing the attacker with the necessary information to have a list of potential vulnerabilities and carry out a computer attack on the services and machines that run it.

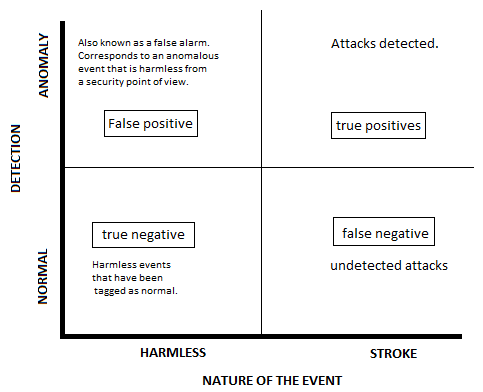
* 1. Intrusion Detection Systems

In [12], IDS is a security measure that helps identify malicious actions that compromise the integrity, confidentiality, and availability of computer resources.

### **Fundamentals Relating to the Evaluation of IDS**

Currently, there is no 100% effective IDS that perfectly classifies regular traffic from malicious traffic because a wide variety of attacks are increasing over time and becoming more novel and unknown to IDSs. In addition to this, there are bad practices in information technologies. This can cause an IDS to make incorrect decisions in network traffic classification processes. For example: identifying attacks as regular traffic.

According to [13], four metrics associated with the nature of the event (harmless traffic or attack) and the detection status (normal or anomalous) have been identified to evaluate the performance of an IDS, as shown in Fig. 1.



**Fig. 1.** Confusion Matrix [13]

An IDS is more efficient when during the data traffic classification process, it presents higher hit rates (i.e., the percentage of true negatives and true positives tends to be 100%) and consequently presents low failure rates (i.e., the percentage of false positives and false negatives tends to 0%). From the above, it is concluded that a perfect IDS detects all traffic correctly, not generating any false alarms. According to [13] above, it can be inferred that:

True Positive (VP): The attack was correctly detected as an anomaly.

False Positive (FP): Harmless traffic incorrectly detected as an anomaly.

True Negative (TN): Harmless traffic is correctly identified as usual.

False Negative (FN): Attack incorrectly identified as regular traffic.

#### Performance metrics

In this research, statistical performance metrics are used to measure the behavior of the IDS concerning the classification process in coherence with that proposed in [14 - 18]; such metrics are defined below.

#### Sensitivity.

Sensitivity is defined as the ability of an IDS to identify "true positive" results:

|  |  |
| --- | --- |
|  | (**1**) |

#### Specificity.

Specificity is defined as the ability of an IDS to measure the proportion of "true negatives" that have been correctly identified.

|  |  |
| --- | --- |
|  | (**2**) |

#### Accuracy.

Accuracy is defined as the degree of closeness of the measurements of a quantity (X) to the value of the true amount (Y), i.e., the proportion of valid results (both true positive and true negative). An accuracy of 100% means that the measured values are the same as the given values:

|  |  |
| --- | --- |
|  | (**3**) |

1. Simulation Process Applied to IDSs

For the effectiveness of a malicious traffic detection process in a computer network by implementing an IDS system that uses feature selection techniques, learning algorithms, and quality measurement of metrics, the evaluation through software simulation in the laboratory is ideal.

Therefore, it requires the execution of several phases, as shown in Fig. 2.

Diagram

Description automatically generated

**Fig. 2.** Simulation phases

* 1. Data Collection Selection Phase (Dataset)

The data collection for the following phases must be selected initially. Different data sources are used in intrusion detection systems: PREDICT, DARPA KDD-NSL, CAIDA, CRAWDAD, DRDC, NIST SAMATE, and Virtual dataset repository.

* 1. Pre-Processing Phase

Preprocessing is a phase before the selection, reduction, or extraction of features in the simulation process. A classification technique based on artificial intelligence techniques is used for unsupervised learning.

This phase allows homogenizing the presentation of the data coming from the dataset and integrating those data in a different format to the simulation tool for data processing purposes. Preprocessing involves the execution of two (2) sub-phases called Parsing and normalization.

Parsing refers to presenting the data coming from the RAW (raw or original) dataset, which is formatted in .txt, to a format that is easy to process by the tool in which the simulation is implemented.

Therefore, the normalization’s purpose is to represent all the attributes as homogeneous as possible so that after being processed in the training process, the data have the same representativeness in the model that will perform the information classification process.

* 1. Characteristics Selection Phase

According to the work of some authors [14, 19, 20], the characteristics selection phase is defined as the optimization process that tries to find the best subset of features from a fixed set of features. It aims to reduce the size of the input data to facilitate processing and analysis, discarding data that do not contribute most to the subsequent classification process. This saves time in data processing with attention to generating optimal results.

* 1. Training Phase

In this phase, the neural network is trained by implementing a learning algorithm such as GHSOM (Growing Hierarchy Self-Organizing Maps), Random Forest, and Bayesian Classifiers, among others, taking as input all the records of the KDDD-Train dataset at 100%, coming from the application of feature selection techniques.

* 1. Classification Phase3

Once the neural network is trained, it proceeds with the classification phase. It is performed autonomously based on implementing a classification algorithm that determines the bi-class traffic (normal and anomalous), presenting the information in a summarized form and based on statistical fundamentals.

* 1. Metrics Evaluation Phase

In this last phase, an evaluation of the quality of the model is performed under the implementation of an algorithm that calculates each of the metrics used for network traffic analysis (sensitivity, specificity, accuracy, and precision) based on the results obtained in the classification stage, to know the main characteristics of the proposed model.

* 1. Cross-Validation Technique Applied in the Classification Phase

The feature selection and training processes have been evaluated using k-fold cross-validation (k = 10) to demonstrate that the system does not overfit and therefore has good generalization performance. This is because k = 10 contains partitions of 90% of the samples randomly selected to fit the model; the rest (10%) were used for testing.

As shown in Fig. 3, these subsets are different and do not share any samples. This process was repeated for all ten (10) folds, ensuring the test data had never been used in feature selection or classifier training. Therefore, the results provided by the selected feature subsets and the classification accuracy are calculated as the average of the ten (10) evaluations across the ten (10) folds.

The main objective of cross-validation is to estimate the generalization error, ensuring that similar results are obtained on new data (i.e., low generalization error). This method calculates the prediction error and avoids double dipping. In addition, it is worth noting that, due to the high number of dataset samples available, both training and testing processes are treated using many samples. This provides a lower generalization error variance estimate.

Tabla

Descripción generada automáticamente

**Fig. 3.** Cross-validation partitioning scheme.

1. Datasets in Intrusion Detection Systems

Researchers continue to evaluate different methodologies and techniques to develop an ever-better IDS solution, requiring an environment that can simulate network traffic as realistically as possible. For this reason, the Massachusetts Institute of Technology - MIT and Defense Advanced Research Projects Agency - DARPA have simulated such a scenario, feeding data collections to provide researchers with a network traffic database, which will serve as input for future research.

The dataset is used to evaluate the efficiency of intrusion detection systems in computer networks. The measurable criteria are the probability of detection and the probability of false alarms of the respective system under analysis.

**Table 1.** IDS-oriented datasets sources

|  |  |
| --- | --- |
| Datasets | Sponsors |
| Dataset DARPA [21] | - IST-LLMIT (Information Systems Technologies Group-Massachusetts Institute of Technology Laboratory).  - DARPA ITO (Defense Advanced Research Projects Agenda-Information Technology Office)  - AFRL/SNHS (Air Force Research Laboratory). |
| Datasets USC/ISI ANT Program PREDICT [22] | - ANT (Network Traffic Analysis Research Group).  - ISI (Information Science Institute).  - USC (University of Southern California).  - Department of Computer Science, Colorado State University.  - USC Department of Electrical Engineering.  - USC Information Technology Services. |
| Datasets CAIDA [23] | - ARIN (American Registry for Internet Members), CISCO, Endance Measurement Systems, U.S Department of Homeland Security, NSF (National Science Foundation). |
| Datasets CRAWDAD [24] | - ACM SIGMOBILE  - Intel Corporation  - National Science Foundation |

Although there is a wide variety of datasets, researchers have commonly opted for using DARPA NSL-KDD in intrusion detection simulation processes, as shown in Table 1, due to its advantages concerning other datasets from the same family and other sources.

Evidence of its implementation is the multiple references to this dataset in articles, conference papers, and papers in press in different indexed databases. Table 2 shows the number of articles per indexed database in the last five (5) years that state that they have used the dataset in cardiovascular disease detection.

**Table 2.** Brief results of the implementation of DARPA KDD

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Indexed database |  | 2018 | 2019 | 2020 | 2021 | 2022 |
| SCOPUS |  | 46 | 59 | 22 | 37 | 86 |
| SPRINGER |  | 31 | 47 | 54 | 31 | 109 |
| SCIENCE DIRECT |  | 36 | 45 | 37 | 49 | 72 |
| IEEE |  | 51 | 71 | 33 | 67 | 95 |

This research has decided to select the NSL-KDD dataset as the input for the subsequent phases of the intrusion detection simulation process. Since the Information Systems Technology Group-IST at the Lincoln Laboratory of the Massachusetts Institute of Technology, LL-MIT has demonstrated considerable improvements in the NSL-KDD dataset over its predecessors, and the worldwide research community (in this field of knowledge) has appropriated and implemented it in their research [25 – 27].

1. IDS Systems Feature Selection Techniques

Feature selection refers to a concept used in data mining to reduce the size of the input data, to facilitate the processing and analysis of such information. Feature selection not only considers the reduction of cardinality, i.e., maintaining a partial or predefined limit on the number of attributes considered when creating a model, but also allows for the appropriate discarding of attributes based on their usefulness for the performance of a good analysis process.

According to Kumar [28], a filter-based feature selection technique (filter) is used to find the best feature subset from the original set; filtering methods seem to be good in selecting a large subset of data, they do not depend on the classification algorithm, and their computational cost is less for large data sets. Wrapper-based feature selection techniques defined in [29] use performance prediction of the learning algorithm for feature selection. It improves the results of the corresponding predictors and achieves better recognition rates, in some cases outperforming filter-based techniques; however, they depend on the classification algorithm, and for a large dataset, the computational cost is higher in the wrapper method.

Finally, embedded methods, also defined in [30], are based on the evaluation of the performance of the metrics calculated directly from the data, without direct reference to the results of the data analysis systems, where there is a union of the feature selection techniques with the learning process for a given learning algorithm. Embedded methods are less prone to overfitting and depend on the classification algorithm.

Using feature selection is essential to perform an effective analysis because the data contains information not necessary for model generation. INFO.GAIN techniques have been selected in this research because, in the preliminary exploration of state of art, it was observed that when implemented in issues related to anomaly detection, their results were promising; however, they have not been implemented in training techniques such as SOM, GHSOM, RED BAYESIANA, NAIVE BAYES, ID3, and RANDOM FOREST to analyze the performance metrics of the proposed model.

The feature selection technique used in the model proposed in this research is presented in detail below.

* 1. Information Gain (INFO.GAIN)

It is a filter-based feature selection technique defined in [31]. It is also known as information gain and is used to identify the relevance or ranking of features in a data collection. The equation that we introduce following defines this relevance level. The attribute with the highest information gain is chosen as the splitting attribute for the N node.

This attribute minimizes the information needed to rank the duplicates in the resulting partition and reflects the lowest randomness or impurity in these partitions.

Implementing the previously documented feature selection technique to the 100% DARPA KDD-Train dataset, the order generated by highest relevance are: src\_bytes, service, dst\_bytes, diff\_srv\_rate, same\_srv\_rate, st\_host\_srv\_count, dst\_host\_diff\_srv\_rate, dst\_host\_serror\_rate, logged\_in, dst\_host\_srv\_serror\_rate, seror\_rate, count, srv\_serror\_rate, dst\_host\_srv\_diff\_host\_rate, dst\_host\_count, dsthost\_same\_src\_port\_rate, srv\_diff\_host\_rate, srv\_count, dst\_host\_srv\_rerror\_rate, protocol\_type, rerror\_rate, dst\_host\_rerror\_rate, srv\_rerror\_rate, duration, hot, wrong\_fragment, num\_compromised, numroot, num\_access\_files, is\_guest\_login, num\_file\_creations, su\_attempted, root\_shell, num\_shells, num\_failed\_logins, land, is\_host\_login, num\_outbound\_cmds, urgent.

1. Data Mining as a Method for the Development of Network Anomaly Detection Systems

According to [32], data mining is the "non-trivial extraction of implicit, previously unknown, and potentially practical knowledge from data. In [33] defines it as "the automatic discovery of interesting and non-obvious patterns or models hidden in a database, which have great potential to contribute to core business issues." After quoting the above definitions and analyzing the concepts, data mining can be defined as integrating patterns that generate helpful knowledge.

* 1. Neural networks SOM (Self-Organizing Map)

It is an efficient neural algorithm (unsupervised) that allows the projection of data from a multidimensional space in a two-dimensional grid called a "map," qualitatively preserving the organization (topology) of the original set [34]. T. Kohonen [35] presented a network model called self-organizing maps or SOM (Self-OrganizingMaps); with this, he wanted to show that an external stimulus (input) can force the formation of maps, assuming a specific structure and a functional description. According to [36], the most preponderant characteristic of the SOM is that it learns to classify data using an unsupervised learning algorithm (a SOM learns to classify training data without any external control).

* 1. Neural networks GHSOM (*Growing Hierarchical Self-Organizing Maps*)

According to [37], GHSOM is a hierarchical and dynamic structure developed to overcome the weaknesses and problems SOM presents. The GHSOM structure consists of multiple layers of several independent SOMs whose number and size are determined during the training phase. The process of adaptive growth is controlled by two parameters that determine the depth of the hierarchy and the breadth of each map.

* 1. Bayesian Network

A Bayesian Network is a directed and annotated acyclic graph that describes the joint probability distribution governing a set of random variables. The topology or structure of the network not only provides information about the probabilistic dependencies between features and the conditional independencies of a variable or set of variables given one or more other variables [38]. Each variable is independent of variables that are not its descendants in the network, given the state of its parent variables. The inclusion of independence relations in the network structure makes Bayesian networks an excellent tool to represent knowledge compactly by reducing the required parameters. In addition, they provide flexible methods of reasoning based on the propagation of probabilities along the network according to the laws of probability theory.

* 1. Naive Bayes

According to [39], this is a descriptive and predictive classification technique based on the probabilistic theory of T. Bayes’ analysis. This theory assumes an asymptotically infinite sample size and statistical independence between the independent variables, referring, in our case, to the attributes, not to the class. With these conditions, the probability distributions of each class can be calculated to establish the relationship between the features (independent variables) and the type (dependent variable).

* 1. ID3

Each internal node of the tree contains a decision on one of the attributes, the value of which will determine the path to classify an example, and each leaf has a class label. Thus, the classification of an example is carried out by traversing the tree from the root to one of the leaves, which will determine the class of the example. Initially, the algorithm takes the entire data set [40].

Despite its simplicity and low computational cost, ID3 has significant drawbacks, some of which are corrected by its successor C4.5. The most obvious is the inability to work with continuous attributes and to deal with missing values.

* 1. Random Forest Algorithm

Random forest, introduced by Breiman in 1999, uses an ensemble (or forest) consisting of many classification trees. To classify a new object, each tree in the ensemble takes it as input and produces an output, its classification. The decision of the set of trees is taken as the class with the most votes in the set [40].

1. Simulation Scenarios and Results

This research's development involved using six (6) experimental test sets. The test sets used the INFO.GAIN feature selection technique hybridizing with the variation of training techniques previously mentioned, as denoted in Figure 2.1.

The different experimentations were executed on a DELL LATITUDE 3470 computer with Intel Core i7 1165g7 processor at 2.5Ghz, 8 GB of Ram DDR 4 at 3200mhz, and NVIDIA GeForce RTX3060 video card with a capacity of 8GB DDR5. Each experiment was performed ten times, thanks to which the values of the metrics that allowed evaluating the quality of the processes were obtained. Table 3 contains each quality metric by the method applied, with their respective standard deviation.

In the classification process, the 10-fold cross-validation technique was used, applied to the "DARPA KDD NSL TRAIN 100%" training dataset, and generated simulations that allow evaluating the tendency to identify traffic and the effectiveness of traffic in a computational network. At the end of each experimental scenario, the proposed functional models were evaluated, calculating the evaluation metrics of sensitivity, specificity, precision, and accuracy.

1. Conclusions

The studies oriented towards intrusion detection systems are of great benefit to ensure the security of different types of computer networks due to the contributions generated by their results with the application of feature selection and classification techniques, which are of great importance for constructing more efficient systems.

When the different simulation scenarios were carried out, it was determined that the INFO.GAIN feature selection method, using the RANDOM FOREST training and classification technique, as shown in table 5.8, generated a precision of 99.77%, an Accuracy of 99.83%, a Sensitivity of 99.89%, and a Specificity of 99.77%, respectively, thus being the most efficient model for the detection of intrusion attacks based on network anomalies, see Table 3.

**Table 3.** Simulation test result

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Technique |  | Precision | Accuracy | Sensibility | Specificity |
| **INFO.GAIN+SOM** |  | 96,20% | 96,36% | 96,52% | 96,20% |
| **INFO.GAIN+ BAYES NET** |  | 93,12% | 94,89% | 96,53% | 93,35% |
| **INFO.GAIN +NAIVE BAYES** |  | 84,61% | 94,43% | 95,76% | 86,22% |
| **INFO.GAIN + RANDOM FOREST** |  | 99,77% | 99,83% | 99,89% | 99,77% |
| **INFO.GAIN+ID3** |  | 99,73% | 99,80% | 99,86% | 99,74% |
| **INFO.GAIN+GHSOM** |  | 99,75% | 99,80% | 99,85% | 99,75% |

About the above, this research considers very important metrics which generate knowledge to be considered in future research.

1. Future Works

Considering the research, methodological, and practical contributions around intrusion detection systems based on network anomalies applying different training and classification techniques (SOM, GHSOM, RED BAYESIANA, NAIVE BAYES, ID3, and RANDOM FOREST), it should be noted that there are several unexplored points or areas of work that may give rise to future research, which are:

Recreate evaluation scenarios where different feature selection techniques (LSA, WRAPPER, GAIN RATIO, RELIEF, COST SENSITIVE, ONE R, FILTERED) are used, implementing the same training and classification techniques addressed in this research to identify a better intrusion detection model.

Develop an embedded system based on free software and hardware where the scripts proposed in this research are implemented to analyze the effectiveness and quality of the results under real attacks in a computer network system.

Analyze alternatives to efficiently implement intrusion detection systems by using optimized modules on computers with multiple processors or programmable network interface cards (e.g., FPGA devices).

References

1. Garfinkel, S., & Spafford, G. (1997). *Web security & commerce* (pp. 349-374). Cambridge, MA: O'reilly.
2. Lehtinen, R., & Gangemi Sr, G. T. (2006). *Computer security basics: computer security*. " O'Reilly Media, Inc.".
3. Ernesto, E, et al. "Neural networks as tool to improve the intrusion detection system." Computer Information Systems and Industrial Management: 20th International Conference, CISIM 2021, Ełk, Poland, September 24–26, 2021, Proceedings 20. Springer International Publishing, 2021.
4. Salcedo, D., et al. "Machine Learning Algorithms Application in COVID-19 Disease: A Systematic Literature Review and Future Directions." Electronics 11.23 (2022): 4015.
5. Henriquez, C., et al. "Predictive Model of Cardiovascular Diseases Implementing Artificial Neural Networks." Computer Information Systems and Industrial Management: 21st International Conference, CISIM 2022, Barranquilla, Colombia, July 15–17, 2022, Proceedings. Cham: Springer International Publishing, 2022.
6. Gaviria, H., & Milton, S. "Detección de cáncer de seno usando imágenes de histopatología y modelos de aprendizaje profundo pre-entrenados Breast Cancer Detection using digital histopathology images and pre-trained deep learning models."
7. Salcedo, D., et al. "Sistema Inteligente para para la gestión automática de un generador eléctrico basado en la arquitectura del IoT." Computer and Electronic Sciences: Theory and Applications 1.1 (2020): 1-10.
8. Liao, H. J., Lin, C. H. R., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, *36*(1), 16-24.
9. Dubey, G. P., & Bhujade, R. K. (2021). Optimal feature selection for machine learning based intrusion detection system by exploiting attribute dependence. *Materials Today: Proceedings*, *47*, 6325-6331.
10. Serpen, G., & Aghaei, E. (2018). Host-based misuse intrusion detection using PCA feature extraction and kNN classification algorithms. *Intelligent Data Analysis*, *22*(5), 1101-1114.
11. De la Hoz, E., De La Hoz, E., Ortiz, A., Ortega, J., & Prieto, B. (2015). PCA filtering and probabilistic SOM for network intrusion detection. *Neurocomputing*, *164*, 71-81.
12. Khamphakdee, N., Benjamas, N., & Saiyod, S. (2015). Improving Intrusion Detection System Based on Snort Rules for Network Probe Attacks Detection with Association Rules Technique of Data Mining. *Journal of ICT Research & Applications*, *8*(3).
13. Mendoza Palechor, F., De la Hoz Manotas, A. K., De-La-Hoz-Franco, E., & Ariza Colpas, P. P. (2015). Feature selection, learning metrics and dimension reduction in training and classification processes in intrusion detection systems.
14. Jaber, A. N., & Rehman, S. U. (2020). FCM–SVM based intrusion detection system for cloud computing environment. *Cluster Computing*, *23*, 3221-3231.
15. Yang, Z., Liu, X., Li, T., Wu, D., Wang, J., Zhao, Y., & Han, H. (2022). A systematic literature review of methods and datasets for anomaly-based network intrusion detection. *Computers & Security*, 102675.
16. Ullah, S., Khan, M. A., Ahmad, J., Jamal, S. S., e Huma, Z., Hassan, M. T., ... & Buchanan, W. J. (2022). HDL-IDS: A hybrid deep learning architecture for intrusion detection in the internet of vehicles. *Sensors*, *22*(4), 1340.
17. Mehmood, M., Javed, T., Nebhen, J., Abbas, S., Abid, R., Bojja, G. R., & Rizwan, M. (2022). A hybrid approach for network intrusion detection. *CMC-Comput. Mater. Contin*, *70*, 91-107.
18. De la Hoz, Emiro, et al. "Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps." Knowledge-Based Systems 71 (2014): 322-338.
19. Liu, G., Zhao, H., Fan, F., Liu, G., Xu, Q., & Nazir, S. (2022). An enhanced intrusion detection model based on improved kNN in WSNs. Sensors, 22(4), 1407.
20. Maldonado, J., Riff, M. C., & Neveu, B. (2022). A review of recent approaches on wrapper feature selection for intrusion detection. Expert Systems with Applications, 116822.
21. Yusof, N. N. M., & Sulaiman, N. S. (2022, August). Cyber attack detection dataset: A review. In Journal of Physics: Conference Series (Vol. 2319, No. 1, p. 012029). IOP Publishing.
22. Mi, Y., Mohaisen, D., & Wang, A. (2022, October). AutoDefense: Reinforcement Learning Based Autoreactive Defense Against Network Attacks. In 2022 IEEE Conference on Communications and Network Security (CNS) (pp. 163-171). IEEE.
23. OZKAN-OKAY, M., AR, Y., & SAMET, R. (2022). A NEW DATASET FOR INTRUSION DETECTION IN WIRELESS LOCAL AREA NETWORK. 9ROXPH, 378.
24. Balakrishna, S. (2022). Multi objective-based incremental clustering by fast search technique for dynamically creating and updating clusters in large data. *Cluster Computing*, *25*(2), 1441-1457.
25. Labrie, M. A., Devereaux, Z., Lecocq, R., & Forrester, B. (2023). Cluster Analysis and Data Exploration Toolset (CADET): Exploring datasets of text documents.
26. Nazim, M. T. B., Faruk, M. J. H., Shahriar, H., Khan, M. A., Masum, M., Sakib, N., & Wu, F. (2022, June). Systematic analysis of deep learning model for vulnerable code detection. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)* (pp. 1768-1773). IEEE.
27. Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, *50*, 102419.
28. Kumar, V., Sinha, D., Das, A. K., Pandey, S. C., & Goswami, R. T. (2020). An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset. *Cluster Computing*, *23*, 1397-1418.
29. Khan, R. U., Zhang, X., Alazab, M., & Kumar, R. (2019, May). An improved convolutional neural network model for intrusion detection in networks. In *2019 Cybersecurity and cyberforensics conference (CCC)* (pp. 74-77). IEEE.
30. Sadikin, F., & Kumar, S. (2020). Zigbee IoT intrusion detection system: A hybrid approach with rule-based and machine learning anomaly detection. In *IoTBDS* (pp. 57-68).
31. Rustam, Z., Arfiani, A., & Pandelaki, J. (2020). Cerebral infarction classification using multiple support vector machine with information gain feature selection. *Bulletin of Electrical Engineering and Informatics*, *9*(4), 1578-1584.
32. Wang, X. (2022, October). Construction of data mining based on multidimensional analysis. In *Third International Conference on Computer Communication and Network Security (CCNS 2022)* (Vol. 12453, pp. 113-118). SPIE.
33. De Silva, K., Mathews, N., Teede, H., Forbes, A., Jönsson, D., Demmer, R. T., & Enticott, J. (2021). Clinical notes as prognostic markers of mortality associated with diabetes mellitus following critical care: a retrospective cohort analysis using machine learning and unstructured big data. *Computers in Biology and Medicine*, *132*, 104305.
34. Tynchenko, V. S., Tynchenko, V. V., Bukhtoyarov, V. V., Kukartsev, V. V., Kukartsev, V. A., & Eremeev, D. V. (2019, May). Application of Kohonen self-organizing maps to the analysis of enterprises’ employees certification results. In *IOP Conference Series: Materials Science and Engineering* (Vol. 537, No. 4, p. 042010). IOP Publishing.
35. Guan, Q., Hong, X., Ke, W., Zhang, L., Sun, G., & Gong, Y. (2021, September). Kohonen Self-Organizing Map based Route Planning: A Revisit. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 7969-7976). IEEE.
36. Mallick, P., Ghosh, O., Seth, P., & Ghosh, A. (2019). Kohonen’s self-organizing map optimizing prediction of gene dependency for cancer mediating biomarkers. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 3* (pp. 863-870). Springer Singapore.
37. Scanagatta, M., Salmerón, A., & Stella, F. (2019). A survey on Bayesian network structure learning from data. *Progress in Artificial Intelligence*, *8*, 425-439.
38. Salmi, N., & Rustam, Z. (2019, June). Naïve Bayes classifier models for predicting the colon cancer. In *IOP conference series: materials science and engineering* (Vol. 546, No. 5, p. 052068). IOP Publishing.
39. Es-Sabery, F., Es-Sabery, K., Qadir, J., Sainz-De-Abajo, B., Hair, A., Garcia-Zapirain, B., & De La Torre-Díez, I. (2021). A MapReduce opinion mining for COVID-19-related tweets classification using enhanced ID3 decision tree classifier. *IEEE Access*, *9*, 58706-58739.
40. Palimkar, P., Shaw, R. N., & Ghosh, A. (2022). Machine learning technique to prognosis diabetes disease: Random forest classifier approach. In *Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2021* (pp. 219-244). Springer Singapore.