مجلة كلية التربية الاساسية
كلية التربية الاساسية – الجامعة المستنصرية
Journal of the College of Basic Education Vol.31 (NO. 130) 2025, pp. 94-114

Machine Learning-Based Intrusion Detection: A Comparative Study

Zinah Sattar Jabbar Aboud[(1)]
The Lebanese University, Lebanon
sattarzeina@gmail.com

Rami Tawil[(2)]
The Lebanese University, Lebanon
rami.tawil@ul.edu.lb

Mustafa Salam Kadhm[(3)]
Computer Department, College of Basic Education, Mustansiriyah University, Iraq
mst.salam@uomustansiriyah.edu.iq

Abstract:
Recently, the internet use is expanded leads for many type of attacks on the network. As a result, a robust and effective network intrusion detection system is needed to strengthen its defense and performance. The main purpose of the intrusion detection system remains to monitor and analyze the system process for potential malicious acts committed by hackers. As a result, researchers have conducted several reviews on such topics, but most of these studies were not comprehensive. In this paper, the authors create a machine learning-based intrusion detection system and use a robust and close set of attribute selection methods with classifiers using a group review, and analyzing of attribute-choosing methods common with functions. The study extracts the important attributes from continuous variables by applying attribute-choosing methods to generate an important variable set and an intrusion detection system. KDD data were double-checked to obtain outcomes from this process. The performance results clearly showed the mathematical k-nearest neighbor's (K-NN) algorithm outperforms the other classifiers. It was also noted that the use of attribute choosing techniques based on the percentage of information gain is preferable compared to other features.

Keywords: Machine learning, intrusion detection, classifier, decision tree, and KNN.
Note: The research is based on a PhD dissertation.

1. Introduction:

Because there is a lot of stuff available online, the Internet currently confronts several issues that pose significant challenges to its development as a dependable and stable network. Additionally, security and stability may be enhanced by upgrading firewalls and other tools. However, the employment of active approaches is the mechanism of the intrusion detection system that permits collaboration amongst system analysts for the system's security and stability [1]. Along with identifying the target. Additionally, it monitors and evaluates network activities during deviations and abnormalities that go against the system's safety regulations [2]. Additionally, it should be watched over and efforts should be made to keep hackers from breaking into the network while it is operational.

Intrusion detection has gained popularity in the research community as a consequence of its efficacy in countering current assaults, as well as the availability of numerous machine learning classification algorithms that are prepared before being applied to unseen data and used to uncover vulnerabilities in system attacks. The attribute reduction approach has also been utilized to improve classification results by speeding up the detection of hacker abuses and abnormalities.

In contrast, intrusion detection systems usually use three techniques. A brief description of these techniques is as follows:

Signature-based intrusion detection: This is a technique that tries to find the attack signature on the detected resources [3]. In addition, it compares log information with new attack methods to determine possible next attacks and try to detect and prevent them from spreading across the network. Hackers that attempt to get into the traffic network leave a form of fingerprint known as an intrusion signature. Although they vary from one to the other, they can be used evidence of illegal access to a folder and file, unsuccessful login attempts, and faulty executable privilege use.

Anomaly-based intrusion detection: This method is a way of understanding behaviour in addition to every deviation from behaviour. It is a method that works in contrast to the signature-based method; it was developed to quickly identify unfamiliar attacks as soon as feasible. By identifying fresh malware that enters the network using machine learning. Where the intrusion prevention system is designed using machine learning by identifying a starting point for a trusted behaviour known as a trust model. This model evaluates each new performance for validation. In the previous network system, in such a case, the authorized person can continue to operate, given

the indication that the warning buzzer was performed wrongly and the action is not malicious.

In 2015, Nutan Farah Haq et al [4] presented a study aimed at proposing a method for aggregate and hybrid classification. Their study also compared the data classification methods used in the experiment settings as well as discussed the variable selection method. In 2016, Chowdhury Nasimuzzaman et al [5] used machine learning methods to classify each abnormal activity of the system. They explained that checking the effectiveness of their presented method was done by evaluating the acceptability of intrusion detection. The cost required to detect intrusion, the nonnegative false ratio, and the non-positive false ratio of ratings of abnormal behaviour detected during network traffic. In 2018, Ravi Kiran Varma, et al [6] proposed a study of an intrusion detection system based on mathematical algorithms and attribute selection methods, which are methods for common machine learning algorithms and data methods classifications.

In 2019, Preeti Mishra et al [7] presented a study on different machine learning methods that were made to identify the causes for challenges connected to these methods to detect intrusion behaviours. In 2020, Hamed Algahtani, et al [8] used a variety of common machine learning methods for detecting interferences. Also, in 2020, Shisrut Rawat et al [9] presented a study on the use of neural networks as well as machine learning methods to design an algorithmic intrusion detection system. Resulting from providing smart services in the field of cyber security. In 2021, Kathryn-Ann Tait et al [10] discussed recent intrusion detection systems and outlined the advantages and disadvantages of each method. They also evaluated various machine learning techniques using certain techniques after they rated them as appropriate in categorizing the attacks.

Also, in 2021, Thirumoothy and Muneeswaran [11] presented a study aiming to compare the Naive Bayes method and the decision tree method. In 2022, Abhishek Raghuvanshi, et al [12] proposed a system to detect and classify the traffic of the Internet of Things systems. They also focused on security as the main worry not just on the Internet of Things systems but also on the uses of the Internet of Things. Noted many types of comparative types of research were conducted on this topic. However, no comprehensive research has been conducted so far on the subject. Their study showed the decision tree is better than Naive Bayes.

Also, in 2022, Emad-ul-Haq Qazi et al [13] presented a study on machine learning classifications of the intrusion detection system. They noted the

decision tree algorithm outperforms the support vector machine algorithm and that Naive Bayes is a type of the Bayes network model.

N. Girubagari et al. (2023) proposed an algorithm for real-time intrusion detection system called PACENIDS using ensemble of Altered Bi-directional Long Short-Term Memory (ABILSTM) and Customized Bi-directional Gated Recurrent Unit (CBIGRU). The proposed algorithm is employed to detect the attacks in the smart cities networks. In order to improve the performance of the proposed algorithm, authors used fuzzy feature selection algorithm. PACENIDS achieved a high classification accuracy 96.59%, 94.47% without feature selection algorithm, and 97.67% classification accuracy with feature selection algorithm using NSL-KDD dataset. However, the applied convolutional architecture takes more time to train the system, which the main limitation of the work [29].

Samer et al. (2023) proposed a hybrid filter-wrapper feature selection method for intrusion detection system called GBA. The proposed method selects a feature subset from the original features to improve the performance of the system. The filter feature selection is based on Information Gain (IG) algorithm, and the wrapper feature selection is based on the Black Hole (BH) algorithm. The aim of GBA to improve the accuracy of the IDs by initialize the features for classification using IG by ignore the zero weighted features. GBA achieved a classification accuracy 96.96% using NSL-WS dataset. However, the work used only one dataset and the obtained accuracy can be improved further [30].

Akindele S. et al (2024) presented combination approach between optimization and machine learning algorithms for network intrusion detection called KOMIC IDS. knapsack optimization algorithm (KO) used first to select the relevant features from the IDs dataset with mutual information gain filter (MIC). After that, a new set of features combine with the selected features. MIC is applied again on the combined features to remove the duplicated features and keep only the highest information gain features. In another hand, several machine learning classifiers are used to evaluate the performance of KOMIC IDS and the best obtained accuracy results was 97.14%, precision 95.53%, recall 99.46%, and F1-score 97.46% using UNSW-NB15 dataset. However, the work used only one dataset, unable to detect the type of attack, and the obtained accuracy can be improved further [31].

In this paper, the authors propose a method for developing an intrusion detection system that contains a combination of classifiers and attribute selection methods. Where it is necessary to create an appropriate intrusion detection system that selects the attribute that can be used in the intrusion detection method along with the classifiers.

2. Selection of feature

An attribute selection is defined as a subset of the primary feature set. By removing redundant and unwanted elements from the core feature set. This set of offerings allows for a faster understanding of the study method, followed by the creation of another comprehensive mathematical algorithm. Also, selecting the attribute helps in understanding the information. Therefore, this paper includes a brief overview of the different popular attribute selection methods.

2.1. Pearson Correlation for Feature Selection: This method is a filtering technique that verifies the given score of the relationship between the calculated and target attributes within the given data sample while preserving the n-sample attributes. This is an equation that evaluates pairs using an appropriate correlation scale with an experimental search approach. It is also an algorithm that assumes appropriate data variables cover attributes that are closely related to the set but distinct from one another [14].

To determine whether there is a relationship between the classification attribute, and the new attribute, the correlation coefficient between the two variables must be used to measure numerically the sample used. However, there is a problem, which is the inability of the relationship between two variables to produce valuable results for the sample due to a large number of data. Therefore, if there is a significant relationship between attributes, it should be discovered after the whole sample has been collected. This is verified according to the following equation:

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} \quad (1)$$

Where r the relationship coefficient, x is the value of the x-attribute in the example data, $\bar{x}$ is the mean calculated value of the x-attribute, y the value of the y-attribute in the example data, and $\bar{y}$ represents the mean calculated value of the y-attribute [15].

2.2. Information Gain Ratio-based attribute: A attribute is selected after obtaining information about variables containing many values. The goal of the information gain ratio-based attribute is to be able to find the feature that returns the most information. That is, who returns the most homogenous divisions? Choosing an attribute based on the information gain ratio eliminates any difficulties by calculating the variable hash Entropy to allocate patterns into sections by dividing the variable data. As the partition value increases, the gain of the variants decreases [16].

2.3. Minimum Redundancy Maximum Relevance: This technique attempts to adjust a variable correlation using redundancy. The mean value of the total exchanged values between the variable and class determines the suitability of the variable set of classes c [17]. While using the MRMR method, one has to choose one option and also needs to decide how many attributes he would like to have. The integer y represents the number of those attributes he wants to keep. Assuming $y=3$ in the starting before entering the loop.

But in practical implementations, one can select the variable y under specific restraints, for example, pattern volume, device memory, or offered time. Note that the MRMR method an operates iterative loop and in every iteration, one has to select the attribute that gets the highest score in terms of the policy it adopts. After selecting the best feature, it adds that selected attribute to a container of allotted attributes. After the attribute is placed in the container, it is impossible to fire it again. After every iteration loop terminates, the following results are obtained:

Table 1: MRMR Processing Operations (when $y=3$).

Iter.	Best	Chosen	Not-Chosen
0	-	()	[age, position, loan, credit, address, company, gender]
1	loan	(loan)	[age, position, credit, address, company, gender]
2	age	(loan, age)	[position, credit, address, company, gender]
3	position	(loan, age, position)	[credit, address, company, gender]

A policy adopts to select the highest attribute in every iteration is as follows:

1. select the loan
2. select the age
3. select the income

According to the above rule, the MRMR method is selected where the attribute most relevant to the target attribute is selected in each iteration as well as the lowest iteration related to the attributes identified in the previous iterations. In fact, in the implementation of every iterative loop x, the value is calculated for every gauged attribute m. The general idea of the MRMR method in the following equation:

$$s_x(m) := \bar{r}(m|t)/r(m|a);$$

(2)

$$until\ r(m|a)\ reach\ x-1;$$

Where s represents the score, m represents the attribute being measured, t represents the target attribute, x represents the iterative conditional statement, r represents the redundancy, r' represents the relevance, and a represents the chosen attribute. So, a good iteration attribute is the one that gets the best value, and according to this option. The designer of the MRMR method has identified several variations for it when measuring accuracy and computing the time taken together, and among these attributes are RFRQ, FCD, FCQ, and FRQ. The method developers have found that the RFCQ attribute is a better selection compared to other features, especially when the direction of a flow categorization pattern is an arbitrary set.

They also found that the FCQ attribute has lower computing time and high accuracy in various directions where the flow categorization pattern is clear and quick, thus outperforming other features. As a result, the emphasis of researchers is on that feature or variable. This does not prevent improvements to other variables to be more efficient and accurate. In iterative x, both the $F-$test of the attribute as well as the target attribute are used to calculate the significance of the attribute m. However, the redundancy of the attribute with the other attributes was chosen in preceding iterations and is calculated using the mean method of Pearson's correlation coefficient. As a result, the equation that performs the calculations looks like this:

$$s_x(m) := F(m,t)/\sum_{s\in a}|Pcc(m,s)|(x-1);$$

(3)

Where x represents xth iteration, m represents the attribute measured, F represents the F-test, and PCC represents the Pearson correlation coefficient. It should be noted the relationship is measured in absolute value. When two attributes are combined with range $(-1..1)$, it seems from the results of table 2 below that both attributes are redundant. Now, consider again the three iterations:

Table 2: MRMR Scores: (a) F-test, (b) PCC, and (c) MRMR

(a)

F-Test

		Target
		income
Not Chosen	credit	46.50
	address	44.52
	position	38.88
	company	13.89
	gender	08.89

(b)

Pearson Correlation Coefficient

		Chosen			Mean
		Age	loan		
Not Chosen	credit	0.02	0.71	→	0.37
	address	0.01	0.63	→	0.35
	position	0.99	0.10	→	0.58
	company	0.00	0.01	→	0.05
	gender	0.02	0.12	→	0.04

(c)

MRMR Method Score

Not Chosen	credit	46.50/ 0.37 =	131.14
	address	44.52/ 0.35 =	129.61
	position	38.88/ 0.58 =	69.93
	company	13.89/ 0.05 =	3001.09
	gender	08.89/ 0.04 =	588.55

On the first iteration, the loan was chosen;
On the second iteration, the age was determined;
On the third iteration, the income was determined;
The F-test and correlations are required to calculate the value of each attribute:
The second attribute in terms of preference is that which gets the highest score, in terms of value, and in such a case, the rotational iterative process illustrated previously is directly performed. The application of the FCQ variant of the MRMR method is easy to understand, but the way it is implemented does not live up to expectations.

In the example above, all correlations are counted and some of those correlations will never be used. Because of all attributes, pairs are dealt with without their background. This means that the processing is carried out through the following equation:

$F(F-1)/2$;

For example, if $F=1000$ there are 499,500 thousand correlations. This results in faster processing and saves time when only required attribute pairs are on each iteration. However, once each iteration is completed, the full

relationships between the attribute chosen in the preceding iteration as well as all other attributes that were at no time specified are recalculated and kept in a relationship array. To put it another way:

1. First iteration: no relationship. No attribute is identified yet: it is sufficient to identify the attribute that gets the top score or with the best $F-$test value.
2. Second iteration: $F-1$relationships are required.
3. Third iteration: $F-2$relationships are required.
4. Kth iteration: A correlation $F-y+1$is required.

The calculations are now simple. As the process needs to count the correlations that are not more than $F(y-1)$. For example, when $F=50$ and $y=20$, the resulting value is more plausible, then the terminated results are 9,500 relationships, so the enhanced version of the equation is as follows:

3. Classifier

It is a program for classifying information and placing it inside one or more groups of categorizations. The classification method is as follows: First, the machine learning information on which the decision-making system is built is provided. Then, the decision-making system is provided with invisible examples to classify them [19]. For example, sort emails by classification: spam and non-spam. Here is a summary of common categorizations methods that were employed in this paper:

3.1. k-nearest neighbor: This classifier can support most machine learning systems with the results being determined by nearest neighbors [20]. Regression and classification are two of their applications. In these applications, the data entry comprises the sample data by nearest the results generated by k-NN. The method of computing the results differs subject to the job at hand. If an unidentified type classification, the type is initially allocated to the class that acts most often amongst the training type closet.

3.2. Decision Tree: This classifier is used for classification as well as estimation of functions. Also, it can be a decision-making tool that employs a decision-tree structure as well as its potential outcomes, such as event results, resource expenses, and benefits. Also, this classifier can display the algorithm containing merely conditional statement data. It creates a method that guesses the result of the pattern using the results of the method data variables. Decision tree creation does not need any knowledge, only knowledge of the given data. It is made up of three main types of nodes which are the decision node, the branch node, and the terminal node [21].

3.3. Naive Bayes Classifier: This is a method for reducing the likelihood of misclassification. For example, assume two variables (A,B) with numbers in $w^d \times \{1,2,...,y\}$, and B is a target field for A. Suppose a subjunctive distribution and due the target field has a number assigned by $(A \mid B = j) \sim P_j$ $^{for(j:=1\ to\ y)do}$ with $\sim$ represents an average distribution and P_j represents a likelihood distribution. The classifier will calculate the observed target field $A = x$ from the observation $B = j$.

In theory, the function is a calculable program, $S : w^d \to \{1,2,...,y\}$ where S represents set in class $S(x)$. The misclassification likelihood expressed as $\Gamma(S) := \mathrm{P}\{S(A) \neq B\}$ denotes Bayesian $S^b(x) := \max \mathrm{P}(B = j \mid A = x)$ where $j \in \{1,2,...,y\}$. The designing probability theory is linked to problems as well as accuracy. $\Gamma(S) - \Gamma(S^b)$ denotes the surplus danger of a generic classifier, which may be based on specific sample data. As a result, this positive is critical for comparing the outcomes of various classification methods. The classifier can rely upon the crossover line intercepts zero because in such a case the sample data size approaches infinity [22].

3.4. Artificial Neural Network: This classifier processes data by connecting a huge numbers of artificial neurons among themselves [23]. A neural network contains invisible layers whose job is to process information before sending the results to the output layer. For neural network training, a type of sample data is passed as input to that neural network. Then the found results are considered, and when they are true, it passes as input. But if the results are incorrect, the mistakes are sent to the application layer by the backpropagation method. Then the load is corrected to achieve the right results for trained systems. The goal of the training is to provide the latest invisible information for training neural networks as well as for classifying the found results into the categories to which they are owned.

3.5. Support Vector Machine: also known as a support vector network, this classifier manages training methods with connected training methods which evaluate the information of classification as well as regression result. Also, this classifier adopts the concept of reducing obstructed threats to improve performance while increasing haste as well as scalability [24]. The goal of using this classifier eliminates the multidimensional space selection constraints that separate invisible methods from classes.

4. Intrusion Detection System

The attribute selection method specifies the process for picking the attribute classification. Furthermore, each set of vital features contributes more specifically to the efficiency of the suggested system. The classifiers are then trained using data from the key attribute. Once the classifiers are trained, comparable data of important attributes are utilized for the final validation of the set of significant attributes and to verify if each network entry is a single normal entry or a network entry.

One of the tasks of the intrusion detection system is to track the network for unusual behaviour. They are host-based or network-based systems [25], and they use the national identification system, or NID system, which is a unique, dependable, and safe system that checks a singular identity before allowing the ID system to be used by that individual. Where the ID system is linked to the ID system through a special technology that detects any abnormal entry to any person by looking at the traffic within the network and distinguishing pretenders while trying to enter any of the hosts through the use of the sense of smell that characterizes the NID system, which is automatically activated during people entering the LAN.

The NID system also monitors other active sites on the network to detect the vulnerabilities created by the hacker who passed through the host-based intrusion detection system or HID system locations. The inability to identify them by reading the inability of packet headers or its inability to detect new types of aggressive attacks. But it can be said that the NID system can find the majority of IP-based denial-of-service attacks and the reason is due to it can diagnose and identify the headers of packets as they move within the network. Although the NID system does not depend on the host operating system as a source to identify the hacker, the HID system relies on the host operating system as a source to identify the hacker and thus his ability to perform its duties properly.

A hybrid intrusion detection system that integrates client-based intrusion detection methods and network-based intrusion detection method have been built. To determine the type of intrusion detection, that intrusion can be the result of misuse or as a result of anomalies [26]. There are three major methods used in intrusion detection systems, which are as follows: the machine learning method, the data extraction method, and the statistical method. Figure 1 shows the proposed intrusion detection system.

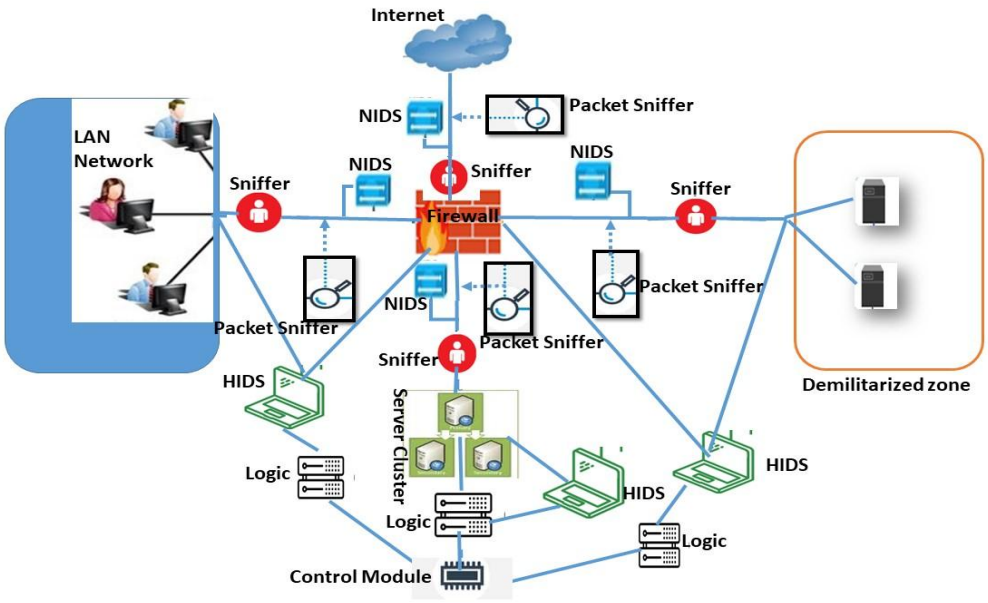


Figure 1: The proposed intrusion detection system

5. Analysis Results with Discussion

Knowledge discovery in a database (KDD) is a multi-stage process to obtain important data from a big database for use in experiments. It should be noted that although some difficulties have emerged in the modified version of the knowledge database. A large number of experiments have been conducted on that data to develop an effective intrusion detection system to provide the cybersecurity public with accessible information that is not available to it. In addition, the need to apply it in intrusion detection systems that occur over the network [27]. It is worth noting that the KDD database is huge and involves about 100,000 rows and 40 columns of very useful data. However, one of the clear drawbacks to dealing with a database of this size is that it grows the expense of complexity. For this reason, the data has been reduced to the point where it is in line with the requirements.

Therefore, only 15,000 samples were sampled from that data for the experiments. Those samples were split 5 times to facilitate the process of checking the validity of the sample extracted from the mother samples. 4 methods for selecting attributes as well as 5 classifiers are also chosen that are compatible intrusion detection [28] and network bypasses. Note that the experiments were performed in the Waikato environment for knowledge analysis or WEKA where the efficiency of the classifiers used to classify the KDD database was checked. Table 3 shows the results of the various groups as well as the methods of selecting attributes and classes to detect standard types of intrusion as well as detecting the other types of aggressive intrusions.

Table 3: Classification Accuracy of Five Classifiers Using Five Attribute Selection Methods.

Attribute Selection Methods	Classifier Type	Accuracy %
Pearson Correlation Coefficient (PCC)	K-nearest neighbor selection	98.29
	Decision Tree	99.51
	Naïve Bayes Classifier	83.49
	Artificial Neural Network	84.36
	Support Vector Machine	80.29
Information Gain Ration-based (IGR)	K-nearest neighbor selection	91.34
	Decision Tree	98.26
	Naïve Bayes Classifier	91.18
	Artificial Neural Network	98.48
	Support Vector Machine	95.37
Minimum Redundancy Maximum Relevance (MRMR)	K-nearest neighbor selection	99.13
	Decision Tree	99.15
	Naïve Bayes Classifier	88.22
	Artificial Neural Network	95.28
	Support Vector Machine	89.77

The outputs of Table 3 illustrate the following observations:

1. k-nearest neighbor selection using the information gain ratio attribute selection technique gives the best results compared to other classifiers. The k-nearest neighbor selection determined using the Pearson correlation coefficient technique yields worse results compared to the k-nearest neighbor selection using other attribute selection techniques.

   Concerning the runtime and storage complexity is that suppose that k a constant and that the runtime complexity of the k-nearest neighbor $O(d*n*\log(n))$ with $(n,d) > 0$ are integers. In addition, the k-nearest neighbors' storage complexity $O(n*d)$. Where n represents the number of data points in the sample and d represents the number of attributes in the sample.

   Concerning the limitations is that The limitation of classifying a k-nearest neighbor does not work with a large data sample, is that not work well with a large number of dimensions, and of being sensitive to outliers and missing values.

2. Decision tree using the Pearson correlation coefficient selection technique achieves better results than a decision tree using other attribute selection

techniques with second order among the other classifiers. A decision tree using the information gain ratio attribute selection technique leads to worse results compared to a decision tree using other attribute selection techniques.

Concerning the runtime and storage complexity is that the number of queries related to the worst-case input and the resulting worst-case determines the decision tree classifier's complexity. The decision tree is used to demonstrate the type of n elements selected. The query for comparison is a comparison of two elements a, b, with two results, presuming no elements are equal to either $(a < b)$ or $(a > b)$. So, the runtime complexity $O(n \log n * d) + O(n * d)$. Where n represents data points, d represents the dimensions and it will be $O(depth)$ because it must traverse the decision tree from root to leaf node. The storage complexity will be $O(nodes)$.

Concerning the limitations is that the limitation of classifying a decision tree is that it is a very unstable tree in comparison to other decision predictions. A minor change in the data can lead to a significant change in the structure of the decision tree, resulting in a different outcome than what users would receive in a typical event. In addition, it is a problem of overfitting and independence between samples, as well as the greedy approach.

3. Naïve Bayes classifier using the information gain ratio attribute selection technique outperforms Naïve Bayes classifier using other attribute selection techniques technique also leads to lower results compared to other attribute selection techniques.

Concerning the runtime and storage complexity is that the naïve Bayes classier method has a runtime complexity of $O(d * c)$ where d is the dimension of the query vector and c is the sum of the classes. The storage complexity is also $O(d * c)$ because it simply stores the probability of each attribute about the classes where d the attribute and c is the class.

Concerning the limitation is that The limitation of a naïve Bayes classier method is that it is more efficient on large data samples and has less space complexity. This method performs poorly as an estimator. As a result, the predictor's probability output should not be underestimated. Furthermore, all attributes are independent. It might sound great in theory but obtaining a set of independent attributes is difficult.

4. Artificial Neural network using the information gain ratio attribute selection technique outperforms an artificial neural network using other attribute selection techniques in terms of performance accuracy as shown in Table 3.

An artificial neural network using the Pearson coefficient technique leads to worse results compared to other attribute selection techniques.

Concerning the runtime and storage complexity is that the runtime and space complexity of a neural network classifier is the same, measured by the number of features in the network. This is the total number of neurons and their connections between neurons. The total runtime complexity of one iteration is $O(t*(xy+yz+zv))$. This time is multiplied by the number of repetitions. Thus, the runtime complexity of a 4-layer neural network with features x, y, z, and v, where t training cases, will be $O(n*t*(xy+yz+zv)$ and n times, if it has L layers, including input and output layers.

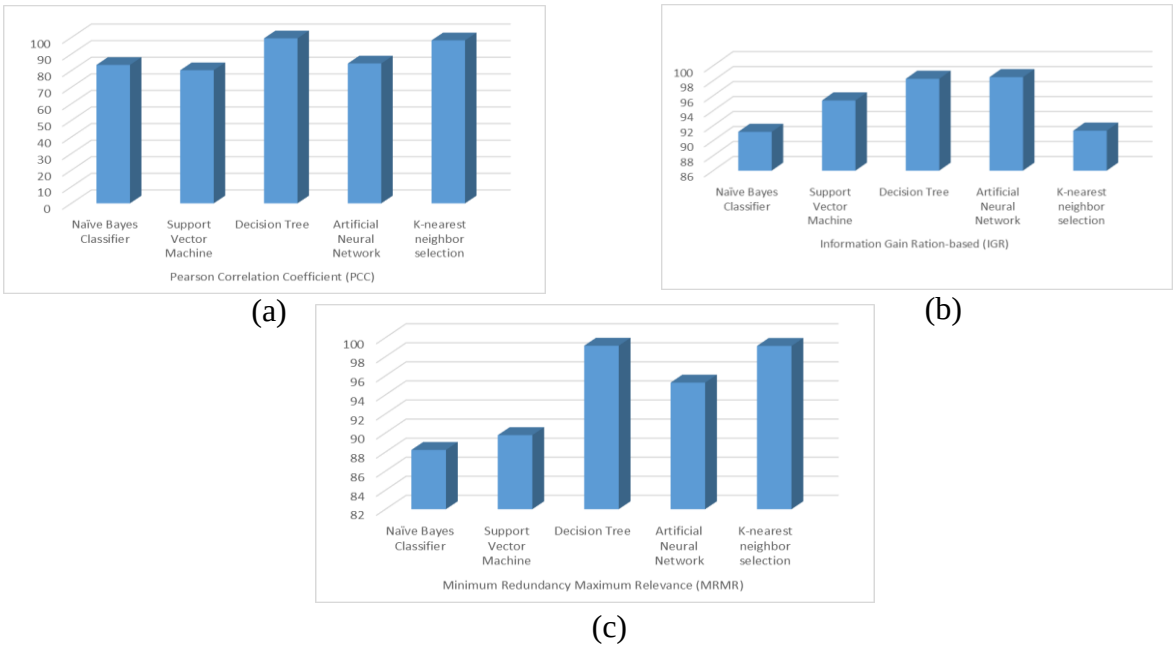
Concerning the limitation is that the limits of neural networks include a fixed number of input layers. It can only accept fixed-size inputs and outputs for any activity. This is a limiting factor for many pattern recognition activities.

5. Support vector machine using the Pearson correlation coefficient technique outperforms support vector machine using other attribute selection techniques in terms of performance accuracy. The Pearson correlation coefficient technique also leads to worse results compared to other attribute selection techniques.

Concerning the runtime and storage complexity is that the support vector machine method is characterized by the runtime complexity of $O(n^3)$ and the storage complexity of $O(n^2)$, since n the data sample is used by the quadratic programming formulation.

Concerning the limitations is that the limitations of the support vector machines are unacceptable for large data sets. It is not well implemented when the data set has more sound i.e. target classes are nested. When the number of characteristics per data point exceeds the number of data samples, the performance of the support vector machine suffers. There is no probabilistic clarification of the classification since the support vector classifier works by placing data points, above and below the hyper-classified plane,

The performance shows that the k-nearest neighbor classifier outperforms the other classifiers, as does the information gain ratio attribute selection and the Pearson correlation technique.



(a) (b)

(c)

Figure 2: The Classification Accuracy: (a) PCC, (b) IGR, and (c) MRMR

6. Conclusion

In this paper, an intrusion detection system that compares the outcomes of several groups has been proposed. Using attribute selection methods, five classifiers are picked based on a set of key qualities. The Pearson correlation coefficient, information gain ratio, minimal redundancy, maximum relevance, and five classifiers, including the k-nearest neighbor method, decision tree, artificial neural network, support vector machine, and Naive Bayes. Both classifiers and attribute selection approaches have advantages and disadvantages. As a result, a range of strategies were adopted for feature selection using classifiers.

When implementing an intrusion detection system, it is difficult to select one. On the other hand, the authors' conclusions demonstrated that machine learning works accurately in intrusion detection since most of the methods used yielded high accuracy results. However, the results also showed that the k-nearest neighbor algorithm outperforms the other methods, as well as between the attribute chosen techniques, the method of selecting the attribute of the information gain ratio outperforms the other methods, while the results showed that, the Pearson correlation coefficient technique was weak in performance compared to the other methods.

The attribute information gain ratio selection using the k-nearest neighbor algorithm resulted in the best accuracy across the groups. Therefore, it may be useful to design an efficient intrusion detection system using an

integration of information gain ratio attribute selection with the k-nearest neighbor algorithm.

Moreover, deep neural networks could be employee as a future research to enhance the performance of the intrusion detection system. Also. metaheuristic algorithms could be used for feature selection task in order to improve the obtained accuracy.

References

1. Ali Ahmadian Ramaki, Abbas Rasoolzadegan, Abbas Ghaemi Bafghi, (2019), "A Systematic Mapping Study on Intrusion Alert Analysis in Intrusion Detection Systems", ACM Computing Surveys, Volume 51, Issue 3, May 2019, Article No., 55, pp. 1- 41.

2. Hurrah, N. N., Parah, S. A., Sheikh, J. A., Al-Turjman, F., & Muhammad, K. (2019). Secure data transmission framework for confidentiality in IoTs. Ad Hoc Networks, Science Direct, Elsevier 95, 101989.

3. Mohammed A. Ambusaidi, Xiangjian He, Priyadarsi, Zhiyuan Tan, (2016), "Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm", IEEE Transaction on Computers, Volume 65, Issue 10, 1 October 2016, pp. 2986 – 2998.

4. Nutan Farah Haq, Abdur Rahman Onik, Avishek Khan Hridoy, Musharrat Rafni, Faisal Muhammad Shah Dewan Md. Farid, (2015), "Application of Machine Learning Approaches in Intrusion Detection System: A Survey", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Volume 4, No.3, pp. 9-18, 2015.

5. Chowdhury Nasimuzzaman, Ferens Ken, Ferens Mike, (2016), "Network Intrusion Detection Using Machine Learning", Proceeding of the International Conference on Security and Management (SAM), Athens, Computer Engineering and Applied Computing, pp. 30-35, 2016.

6. Ravi Kiran Varma, Valli Kumari, Srinivas Kumar, (2018), "A Survey of Feature Selection Techniques in Intrusion Detection System: A Soft Computing Perspective", Progress in Computing, Analytics and Networking, Springer, pp.785-793, 11 April 2018.

7. Preeti Mishra, Vijay Varadharajan, Uday Tupakula, Emmanuel S. Pilli, (2019), "A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection", IEEE Communications Surveys & Tutorials, (Volume: 21, Issue: 1, Firstquarter 2019), pp. 686 – 728.

8. Hammed Algahtani, Iqbal H. Shrker, Asra Kalim, Syed Md. Minhaz Hossain, Sheikh Ikhlaq, Sohrab Hossain, (2020), "Cyber Intrusion Detection Using Machine Learning Classification Techniques", International

Conference on Computing Science, Communication and Security, COMS2: Computing Science, Communication and Security, pp. 121-131.
9. Shisrut Rawat, Aishwarya Srinivasan, Vinayakymar Ravi, Uttam Ghosh, (2020), "Intrusion detection systems using classical machine learning techniques vs integrated unsupervised feature learning and deep neural network", Internet Technology Letter, Volume 5, Issue 1, Special Issue: Deep Learning for Future Smart Cities, January/February 2022, Wiley, pp. 1-5.
10. Kathryn-Ann TaiFahad Sher Khan, Fahad Alqahtani, Awais Aziz Shah, Fadia Ali Khan, Mujeeb ur Rehman, Wadii Boulila, Jawad Ahmad, (2021), "Intrusion Detection using Machine Learning Techniques: An Experimental Comparison", International Congress of Advanced Technology and Engineering (ICOTEN), IEEE Publisher.
11. Thirumoothy K, Muneeswaran K, (2021), "Feature selection using hybrid poor and rich optimization algorithm for text classification", Pattern Recognition Letters, Volume 147, Science Direct, Elsevier, July 2021, pp. 63-70.
12. Abhishek Raghuvanshi, Umesh Kumar Singh, Guna Sekhar Sajja, Harikumar Pallathadka, Evans Asenso, Mustafa Kamal, Abha Singh, and Khongdet Phasinam, (2022), "Intrusion Detection Using Machine Learning for Risk Mitigation in IoT-Enabled Smart Irrigation in Smart Farming", Journal of Food Quality, Volume 2022 |Article ID 3955514.
13. Emad-ul-Haq Qazi, Muhammad Imran, Norman Haider, Muhammad Shoaib, Imran Razzak, (2022), "An intelligent and efficient network intrusion detection system using deep learning", Computers and Electrical Engineering, Volume 99, April 2022.
14. Touraj Sattari Naseri, Farhad Soleimanian Gharehchopogh, (2022), "A Feature Selection Based on the Farmland Fertility Algorithm for Improved Intrusion Detection Systems", Journal of Network and Systems Management, Volume 30, Number 40, Springer, 19 March 2022.
15. Yaqing Liu, Yong Mu, Keyu Chen, Yiming Li & Jinghuan Guo, "Daily Activity Feature Selection in Smart Homes Based on Pearson Correlation Coefficient", (2020), Natural Processing Letters, Springer, 51, 1771-1787, 7 January 2020.
16. M. Nivaashini, R. S. Soundariya, H. Vidhya Shri, P. Thangaraj, (2018), "Comparative Analysis of Feature Selection Methods and Machine Learning Algorithms in Permission based Android Malware Detection", 2018 International Conference on Intelligent Computing and Communication for Smart World (I2C2SW), IEEE, 14-15 Dec. 2018.

17. Yahye Abukar Ahmed, Baris Kocer, Shamsul Huda, Bander Ali Saleh Al-rimy, Mohammad Mehedi Hassan, (2020), "A system calls refinement-based enhanced Minimum Redundancy Maximum Relevance method for ransomware early detection", Journal of Network and Computer Applications, Science-Direct, Elsevier, Volume 187, 1 October, 2020.
18. Ardy Wibowo Haryanto, Edy Kholid Mawardi, Muljono, (2018), "Influence of Word Normalization and Chi-Squared Feature Selection on Support Vector Machine (SVM) Text Classification", 2018 International Seminar on Application for Technology of Information and Communication, IEEE Explore, 21-22 Sept. 2018.
19. Jintao Zhang, Zhuo Wang, Naveen Verma, (2017), In-Memory Computation of a Machine-Learning Classifier in a Standard 6T SRAM Array, IEEE Journal of Solid-State Circuits, Volume 52, Issue 4, pp. 915-924, 10 March 2017.
20. Altman N., (1992), "An introduction to kernel and nearest-neighbor nonparametric regression", The American Statistician, vol. 46, issue 3, pp. 175-185, 1992.
21. László Györf, (2018), Nonparametric Estimations and Predictions, National University of Public Service under the priority project KÖFOP-2.1.2-VEKOP-15-2016-00001 titled "Public Service Development Establishing Good Governance" in Ludovika Research Group, pp. pp. 1-153, August 15, 2018.
22. Hongfang Zhou, Xiqian Wang, Rourou Zhu, (2021), "Feature selection based on mutual information with correlation coefficient", The International Journal of Research on Intelligent Systems for Real Life Complex Problems, 52, Springer, pp. 5457–5474, 2022.
23. Malek Al-Zewairi, Sufyan Almajali, Arafat Awajan, (2017), "Experimental Evaluation of a Multi-Layer Feed-Forward Artificial Neural Network Classifier for Network Intrusion Detection System", International Conference on New Trends in Computing Sciences (ICTCS), IEEE Xplore, Jordan-Amman, 11-13 October 2017.
24. Sandy Victor Amanoul, Adnan Mohsin Abdulazeez, (2022), "Intrusion Detection System Based on Machine Learning Algorithms: A Review", 2022 IEEE 18th International Colloquium on Signal Processing & Applications (CSPA), 12-12 May 2022, Selangor, Malaysia, IEEE.
25. Abdul hammed R, Musafer H, Alessa A, Faezipour M, Abuzneid A, (2019), "feature dim dimensionality reduction approaches for machine learning based network intrusion detection", Electronics 8 2019, 322.

26. Alhakami W, ALharbi A, Bourouis S, Alroobaea R, Bouguila N, (2019), "Network anomaly intrusion detection using a nonparametric bayesian approach and feature selection", IEEE Access 7 2019, 52181–52190.
27. Sarika Choudhary, Nishtha Kesswani, (2020), "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 Datasets using Deep Learning in IoT", Procedia Computer Science, Volume 167, pp. 1561-1573, 2020.
28. Mikel K. Ngueajio, Gloria Washington, Danda B. Rawat, Yolande Ngueabou, (2022), "Intrusion Detection Systems Using Support Vector Machines on the KDDCUP'99 and NSL-KDD Datasets: A Comprehensive Survey", Proceedings of SAI Intelligent Systems Conference, Intelligent Systems and Applications, Proceedings of the 2022 Intelligent Systems Conference (IntelliSys) Volume 2, Springer.
29. N. Girubagari and T. N. Ravi, "Parallel ABILSTM and CBIGRU Ensemble Network Intrusion Detection System," International Journal of Intelligent Engineering and Systems, Nol. 17, No. 1, pp. 93–107, 2024, doi: 10.22266/ijies2024.0229.10.
30. S. S. Issa, S. Q. Salih, Y. D. Salman, and F. H. Taha, "An Efficient Hybrid Filter-Wrapper Feature Selection Approach for Network Intrusion Detection System," International Journal of Intelligent Engineering and Systems, Vol. 16, No. 6, pp. 261–273, 2023, doi: 10.22266/ijies2023.1231.22.
31. A. S. Afolabi and O. A. Akinola, "Network Intrusion Detection Using Knapsack Optimization, Mutual Information Gain, and Machine Learning," Journal of Electrical and Computer Engineering, Vol. 2024, pp. 1–21, 2024, doi: 10.1155/2024/7302909.

مجلة كلية التربية الاساسية
كلية التربية الاساسية – الجامعة المستنصرية
Journal of the College of Basic Education Vol.31 (NO. 130) 2025, pp. 94-114

اكتشاف التطفل في شبكات الانترنيت باستخدام التعلم الآلي: دراسة مقارنة

زينة ستار جبار عبود[1]
الجامعة اللبنانية
sattarzeina@gmail.com

رامي الطويل[2]
الجامعة اللبنانية
rami.tawil@ul.edu.lb

مصطفى سلام كاظم[3]
قسم الحاسبات، كلية التربية الاساسية، الجامعة المستنصرية
mst.salam@uomustansiriyah.edu.iq

مستخلص البحث:
في الاونة الاخيرة زيادة استخدام الإنترنت، ادى الى زيادة العديد من انواع الهجمات على الشبكة. ونتيجة لذلك، هناك حاجة إلى نظام قوي وفعال لكشف التطفل على الشبكة لتعزيز دفاعها وأدائها. ويظل هدف نظام كشف التطفل هو مراقبة وتحليل عملية النظام بحثًا عن أعمال خبيثة محتملة يرتكبها المتسللون. ونتيجة لذلك، تم في هذا البحث اجراء العديد من المراجعات حول مثل هذه الموضوعات، لكن معظم هذه الدراسات لم تكن شاملة. في هذا البحث، تم انشاء نظامًا لكشف التطفل قائمًا على التعلم الآلي وتم استخدام مجموعة قوية ومتقاربة من طرق اختيار السمات مع المصنفات باستخدام مراجعة المجموعة، وتحليل طرق اختيار السمات المشتركة مع الوظائف. تستخرج الدراسة السمات المهمة من المتغيرات المستمرة من خلال تطبيق طرق اختيار السمات لتوليد مجموعة متغيرات مهمة ونظام كشف التطفل. تم التحقق من بيانات KDD مرتين للحصول على نتائج من هذه العملية. أظهرت نتائج الأداء بوضوح أن خوارزمية (K-NN) تتفوق على المصنفات الأخرى. ولوحظ أيضًا أن استخدام تقنيات اختيار السمات بناءً على نسبة اكتساب المعلومات يعد أمرًا مفضلًا مقارنة بالميزات الأخرى.

الكلمات المفتاحية : التعلم الآلي، اكتشاف التطفل، اختيار السمات، التصنيف، شجرة القرار و KNN.
ملاحظة : هل البحث مستل من رسالة ماجستير او اطروحة دكتوراه ؟ نعم