Journal of Naval Sciences and Engineering
2022, Vol. 18, No. 1, pp. 31-59
Computer Engineering/Bilgisayar Mühendisliği

RESEARCH ARTICLE

**An ethical committee approval and/or legal/special permission has not been required within the scope of this study.*

SUPERVISED MACHINE LEARNING-BASED CLASSIFICATION OF NETWORK THREATS/ATTACKS AGAINST COMPUTER SYSTEMS*

Ercan KURU[1]
Tolga ÖNEL[2]
Mehmet Bilge Kağan ÖNAÇAN[3]
Musa MİLLİ[4]

[1]*National Defence University, Barbaros Naval Sciences and Engineering Institute, Department of Cyber Security, Istanbul, Turkey;*
kru.ercn@gmail.com; ORCID: 0000-0001-8359-5684

[2]*HAVELSAN, Department of Command Control and Defence Technologies, Istanbul, Turkey;*
onel.tolga@gmail.com; ORCID: 0000-0002-3256-8122

[3]*Istanbul Okan University, Faculty of Applied Sciences, Department of Information Systems and Technology, Istanbul, Turkey;*
mbko@yahoo.com; ORCID: 0000-0002-7147-0945

[4]*National Defence University, Turkish Naval Academy, Department of Computer Engineering, Istanbul, Turkey;*
mmilli@dho.edu.tr; ORCID: 0000-0001-8323-6366

Received: 28.11.2021 **Accepted: 31.01.2022**

ABSTRACT

With the developing technology, number of people who use computers are increasing nowadays. This increase in computer usage causes an increase in the variety of attacks and the number of attacks against computer systems. This situation reveals the importance of the protection of data processed on the computers and the concept of information security. Thanks to the intrusion detection systems, which have an important place in the protection of computer systems, attacks against computers and computer networks can be detected before they affect systems. Considering the increasing variety of attacks, the development of machine learning-based attack detection systems has been the subject of many studies recently. Although supervised and unsupervised machine learning have separate features, they make different contributions to the areas in which they are used. Within the scope of this study, NSL KDD data set, one of the most frequently used data sets in previous studies to simulate network traffic, was applied to a number of supervised and unsupervised learning algorithms in the WEKA application. When the results are evaluated under certain criteria, it has been determined that supervised learning algorithms give more accurate results, where unsupervised learning algorithms give faster results in the detection of attacks.

Keywords: *Intrusion Detection System, Supervised Learning, Unsupervised Learning, Information Security, Dimensionality Reduction.*

Supervised Machine Learning-based Classification of Network Threats/Attacks Against Computer Systems

BİLGİSAYAR SİSTEMLERİNE YÖNELİK AĞ TABANLI TEHDİTLERİN/SALDIRILARIN DENETİMLİ YAPAY ÖĞRENME İLE SINIFLANDIRILMASI

ÖZ

Gelişen teknoloji ile birlikte günümüzde bilgisayar kullananların sayısı artmaktadır. Bilgisayar kullanımındaki bu artış, bilgisayar sistemlerine yönelik saldırıların çeşitliliğinin ve sayılarının artmasına neden olmaktadır. Bu durum, bilgisayarlarda işlenen verilerin korunmasının ve bilgi güvenliği kavramının önemini ortaya koymaktadır. Bilgisayar sistemlerinin korunmasında önemli bir yere sahip olan saldırı tespit sistemlerinin çalışma prensibi sayesinde bilgisayarlara ve bilgisayar ağlarına yönelik saldırılar sistemleri etkilemeden tespit edilebilmektedir. Artan saldırı çeşitliliği göz önüne alındığında, yapay öğrenme ile saldırı tespit sistemlerinin geliştirilmesi son zamanlarda birçok araştırmaya konu olmuştur. Denetimli ve denetimsiz yapay öğrenme ayrı özelliklere sahip olsa da kullanıldıkları alanlara farklı katkılar sağlamaktadırlar. Bu çalışma kapsamında, WEKA uygulaması kullanılarak bir takım denetimli ve denetimsiz öğrenme algoritmaları, ağ trafiğini simüle etmek için önceki çalışmalarda en sık kullanılan veri setlerinden biri olan NSL KDD veri setine uygulanmıştır. Sonuçlar değerlendirildiğinde, saldırı tespitinde denetimli öğrenme algoritmalarının daha doğru, denetimsiz öğrenme algoritmalarının ise daha hızlı sonuç verdiği tespit edilmiştir.

Anahtar Kelimeler: *Saldırı Tespit Sistemi, Denetimli Öğrenme, Denetimsiz Öğrenme, Bilgi Güvenliği, Boyut Azaltma.*

1. INTRODUCTION

Internet and computer usage is getting more widespread nowadays and we encounter these two definitions in almost every area of our lives. According to the data of March 2021, 66% of nearly eight billion people living on earth use the internet. Depending on the increasing internet usage, the number of malicious software is increasing and diversifying every day. Thus, the information stored/processed on the computer or computer networks and the security of this information appear as a very important issue.

Many institutions/organizations around the world try to find some solutions by developing software and methods to ensure information security. The reliability and the performance of these solutions are the main reasons why users prefer these solutions. Malicious software causes other software running on the computer to behave differently than they should be, or damages the software that it affects (Kramer & Bradfield, 2009). Software using for ensuring information security varies depending on the type of malicious software. For example; while antivirus programs are used against malicious software such as viruses and trojans, anti-spyware software can be used against spyware.

Security of computer networks consisting of more than one computer is provided by intrusion detection systems. Intrusion detection systems examine the behavior of network traffic and determine whether the incoming data is malicious or not. Network traffic behavior is classified with developed algorithms. At this stage, machine learning comes into play. Machine learning algorithms form the basis of intelligent systems used in many areas of our lives. They can be expressed by analyzing the problem encountered by the software programmed in the computer system based on a specific data set or previous experiences (Alpaydın, 2010).

Machine learning is generally examined under three headings as Supervised Machine Learning, Unsupervised Machine Learning and Reinforcement Learning (Simeone, 2018). The data are processed by using the learning type according to the problem and the results are evaluated. There are many articles about classification of threats/attacks against computer systems but, in our study, we use both supervised and unsupervised machine learning algorithms and also dimensionality reduction to classify threats/attacks against computer systems. In this article, the machine learning algorithm that should be used in order to develop a better and effective intrusion detection system was tried to be determined by using NSL KDD

dataset derived from KDD CUP-99 dataset and WEKA application. The remainder of this paper is organized as follows. Section 2 gives information about the concept of knowledge, specifically, the definition and applications of information security. Section 3 and 4 examine the intrusion detection systems and machine learning respectively. Section 5 gives information about the algorithms and test setup. In Section 6, test results are evaluated. Finally, in Section 7, conclusion and discussion of this study are presented.

2. KNOWLEDGE AND INFORMATION SECURITY

Before knowledge, we should explain "data" and "information" first. "Data" are values collected by sensors, consisting of various symbols, letters, numbers and signs and that do not make sense in itself. "Information" is a collection of processed and meaningful data. Thus, "knowledge" is the inferences from information that has become conceptual (Avcı & Avcı, 2004; Bellinger et al., 2004; Kocabıyık, 2005; Kurgun, 2006; İlter, 2011; Önaçan, 2015).

Information is the processed form of data. Data and information show what anyone knows. Knowledge is the conceptual state of information. It shows how anyone knows. The most valuable asset today is the information. There are institutions and organizations operating in many fields from storing, processing and ensuring the security of information. Before the age of technology, the information was transferred from generation to generation, either verbally or in writings, can now be transferred from society to society very easily with computers and the internet.

Although easy access to information is an advantage, reaching accurate and reliable information and ensuring the security of information is currently the focus of many studies. Information that used to be kept in cabinets is nowadays stored in computers and even huge servers thanks to cloud computing. This situation reveals how important the security of information processed on computers and computer networks is. As a matter of fact, today many institutions and organizations take various measures to ensure information security.

Information security is the protection of information and information systems against unauthorized access, unauthorized use, unauthorized alteration and removal (Andress, 2011). As can be understood from the definition of information security, any activity aimed at changing and eliminating the real form of information covers the issue of information security.

In order to ensure information security, confidentiality must be provided, integrity must be maintained, and information must be available where information is processed. These three elements that constitute the basis of information security, form CIA (Confidentiality-Integrity-Availability) triangle, which is shown in Figure 1 (Solomon & Chapple, 2005).

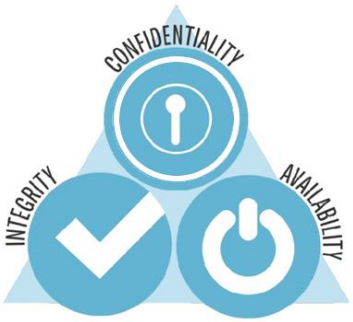

Figure 1. CIA triangle.

The purpose of confidentiality is to prevent information from falling into unauthorized hands. Integrity deals with detection of and prevention from the unauthorized change of information. The purpose of integrity is to keep information as it should be. Availability means that information is always available. The purpose of availability is that users can access the data they want to access whenever they want within their authority.

Carnegie Mellon University, with its study (Figure 2), reveals that despite the increasing difficulty in techniques used in attacks against information security, the attacker's knowledge level has decreased (Allen et al., 2000).

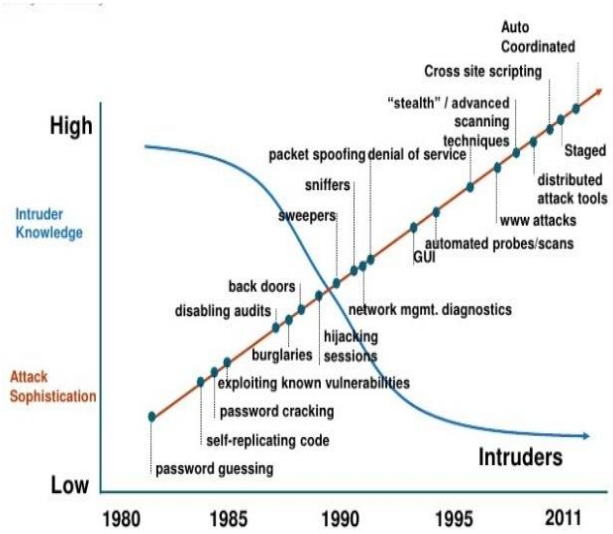

Figure 2. Attack sophistication vs. intruder technical knowledge.

Variety of attacks increases day by day as in Figure 2. Moreover, the level of knowledge required to carry out these increasing attacks tends to decrease. This situation forces computer users to take various measures. It is of great importance to ensure information security, especially in institutions and organizations. In this context, a world standard was established for the first time in 2005 by the International Standards Organization (ISO) and the International Electrotechnical Commission (IEC). The ISO/IEC 27001 standard specifies the requirements for the establishment, implementation, maintenance and continuous improvement of the systems that ensure information security (Ersoy, 2012). The standard was last updated in 2013, and it is a document used in activities carried out to ensure information security today.

3. INTRUSION DETECTION SYSTEMS

Intrusion detection refers to the detection of any attacks on computer systems and information security. Protection from attack, on the other hand, refers to the response to the attack in addition to the detection of the attack. Intrusion detection is the first step to protect against vulnerabilities. Intrusion detection systems detect attacks by collecting information from various systems and network sources and analyzing the data they collect (Taher et al., 2019). Considering the history of

intrusion detection systems, it is seen that the early studies in this subject coincides with the 1960s (Yost, 2016).

Intrusion detection systems have started on the basis of a single computer and have become to protect the whole network system that contains many devices today. Some sources of motivation in the development of attack detection systems are in the following:

- New network systems are complex and as a result they are prone to failure. These errors can also be used by malicious people.

- The network systems have some important defensive deficiencies, which makes the network systems the target of attackers. Although these deficiencies are tried to be covered with some tools and methods, it is not possible to completely eliminate the deficiencies.

- Although there are systems for protection from attack in network systems, full protection may not be possible. As a result, the need for intrusion detection systems is increasing.

- New types of attacks are constantly being developed for protection and detection systems. Thus, a dynamic structure that constantly learns and renews itself is needed for security solutions (Karataş et al., 2018).

Intrusion detection systems analyze and predict users' behavior to determine whether the behavior is an attack or a normal behavior. Intrusion detection systems are generally examined in two sections as Network Based Intrusion Detection Systems and Host Based Intrusion Detection Systems.

Network based intrusion detection systems examine network traffic using basic network packets to detect suspicious situations. Network packets are classified in three ways. In the string signatures method, the data related to the event, which may occur in the packet data, are examined, while in the port signatures method, it is checked whether there is any network traffic different from the relevant gates. In the header signatures method, the headers of the incoming network packets are examined and it is checked whether there is an unreasonable or possibly dangerous request (Liu, 2014).

It is possible to list the advantages and disadvantages of network based intrusion detection systems as follows. Since such systems use network packets for intrusion

detection and network data is a form of internet protocol (IP) packets, they can operate independently of the platform and operating system. The use of network packets in the detection of attacks brings early and rapid detection in case of possible attack. In addition, it does not affect the computer performance of the system it is in. Network based intrusion detection systems do not work effectively in case of ordinary excessive network traffic and may have difficulties in detecting network traffic consisting of encrypted data packets.

On the other hand, host based intrusion detection systems control data on a single computer. Examples of audited data include operating system calls, events, resource usage, and system logs. Any incompatibility or unusual behavior that may occur in these data is tried to be determined (Liu, 2014).

When we examine the advantages and disadvantages of host based intrusion detection systems, these systems enable us to have an idea of whether the attacks are successful or not, and to control the access activities of the user/files and the changes that may occur in the system files. However, host based intrusion detection systems are weaker in real-time response to attacks and against large-scale attacks.

4. MACHINE LEARNING

Machine learning can simply be defined as solving a problem faced by a computer program using previous experiences and data sets defined in that program. To decide whether a problem can be solved by machine learning methods or not, below criteria must be examined:

- The need for functions that lead from well-defined inputs to specific results;

- The need for very large data sets to solve the problem;

- The need for feedback containing clearly defined goals and data;

- Requirement of a detailed explanation of how the decision was made in order to reach the result;

- The solution to the problem does not need the most appropriate solution that is tolerant and provable for the error;

- Special hand skills, physical skills or no need for mobility to solve the problem (Brynjolfsson & Mitchell, 2017).

According to the learning method, machine learning is examined under three sections. Namely, Supervised Machine Learning, Unsupervised Machine Learning and Reinforcement Learning. In supervised machine learning, a function that correlates labeled input values with desired output values is learned. In unsupervised machine learning, a function is learned by using unlabeled data. In reinforcement learning, the learner tries to find the style of action that maximize the output according to the feedback it receives by interacting with the environment.

Supervised machine learning can be explained by the example of a student that learns a subject he/she does not know with the help of his/her teacher. The teacher knows the subject to be taught and what his/her student will learn. By feeding the labeled data set to the learning algorithm, it is "taught" to establish the relationship between the input and output of the algorithm. The trained algorithm performs its subsequent operations in the light of the learned function.

In supervised machine learning, the training set with known inputs and outputs can be thought as a teacher in the learning process. Hence, learning with a teacher is called as the supervised learning. Training process continues until the outputs of the machine algorithm reach to an acceptable level of accuracy (Brownlee, 2017). After this learning stage, the unprecedented data are categorized the algorithm according to the what has learned before.

Supervised learning algorithms are categorized into main classes. Namely the classification algorithms and regression algorithms. Classification algorithms decide which class or category the data sample belongs to. Regression algorithms, establish a relationship between the input data sample and the related output.

Unlike the supervised machine learning, in unsupervised machine learning, there is no “taught” and “labeled” data and the algorithm performs its own learning. In unsupervised machine learning, the algorithm creates a learning pattern for itself by using the features in the input data set. In the learning process unsupervised learning does not need a teacher as in the supervised learning.

Unsupervised learning algorithms are generally classified into two subclasses. These are clustering and association algorithms. Clustering algorithms cluster the data set given as input according to their characteristics, while association algorithms separate the features in the data set by establishing relationships. In Table 1, supervised and unsupervised learning algorithms are compared in terms of their definitions, applications and results.

Table 1. Supervised vs. unsupervised learning algorithms.

Parameter/ Benchmark Feature	**Supervised Learning Algorithm**	**Unsupervised Learning Algorithm**
Input Data	Labeled	Unlabeled
Purpose	To obtain a function that can predict the output of the given data different from the training set	Finding possible structures and hidden models in the input data set
Computational Complexity	Simple	Complex
Data Usage	Connects inputs and outputs	Does not use output data
Accuracy of Results	High reliability and accuracy	Low reliability and accuracy
Number of Classes	The number of classes used is determined	The number of classes used is uncertain
Usage Areas	Pattern recognition in picture and sound files, financial analysis, training of neural networks	Pre-training of raw data processing, data analysis, supervised learning algorithms

Considering the developing technology and the need for machine learning, both supervised learning algorithms and unsupervised learning algorithms are evolving and differentiating day by day. When compared to the unsupervised learning algorithms, supervised learning algorithms produce more accurate results with the light of labeled datasets. Unsupervised learning algorithms are good at investigating the correlations in the input data set.

5. ALGORITHMS AND TEST SETUP

In order to develop a better and more effective intrusion detection system by classifying threats and attacks against network-based computer systems, performances of the supervised and unsupervised machine learning are investigated using the NSL KDD dataset and WEKA application.

5.1. Review of Dataset

KDD CUP-99 data set is the version of the dataset developed by DARPA in 1998 (Ferrag et al., 2020). NSL KDD dataset is the compiled version of KDD CUP-99 dataset. NSL KDD data set is frequently used by researchers today and consists of "KDDTest", "KDDTest-21", "KDDTrain_20Percent", "KDDTrain" sub-sets (Dhanabal & Shantharajah, 2015).

There are three main features that distinguish the NSL KDD dataset from the KDD CUP-99 dataset and cause users to choose it. The first of these features is that the data in the KDD CUP-99 dataset, that mislead the classification algorithms, are reduced in the NSL KDD dataset. Thus, the margin of error is reduced while the classification algorithms are run. Secondly, the data in the NSL KDD dataset with different difficulty levels in terms of attack detection is inversely proportional to the data in the KDD CUP-99 dataset. This feature causes the classification rates of different machine learning algorithms applied with the NSL KDD dataset to spread over a wide range and this situation is beneficial for the users in terms of correctly evaluating the results of different algorithms. The third and last feature is that the number of training and test data in the NSL KDD data set is reduced compared to the KDD CUP-99 data set. This feature allows users to work on the entire data set without having to select any part of the data set (Chae et al., 2013).

The traffic data labeled as attack in the NSL KDD dataset consists of thirty-nine attack types evaluated in four classes in total. The first of the traffic data tagged as attack in the NSL KDD dataset is the “Denial of Service (DoS)” attack, the second is the “User to Root (U2R)” attack, the third is the “Remote to Local (R2L)” attack and the fourth is the “Probing” attack.

Denial of Service attack aims to use computer resources more than normal, and computers exposed to this attack type become unable to respond to users' demands. In the User to Root attack, attackers aim to be a privileged user (root, administrator, etc.) in the system. In the Remote to Local attack, they aim to use vulnerabilities in the local machine with the data they send over the network. In probing attacks, network traffic is examined, data is collected about computers and an attack is developed according to the detected weak points (Thomas & Pavithran, 2018).

5.2. Introduction of WEKA Application

WEKA (Waikato Environment for Knowledge Analysis) is a Java-based data processing and analysis program developed by Waikato University in New Zealand. The program began to be developed in 1993 with the support of the New Zealand government and was first available worldwide in 1999. The modular and extensible structure of the WEKA application allows users to quickly experiment and compare different machine learning methods with different data sets (Witten et al., 2009).

Data can be uploaded to the WEKA application from the database, over the internet (URL) and from the file. The program supports many file formats such as CSV and LibSVM with the ARFF format produced for it. In addition, thanks to the visual interface it offers, users can display their operations with graphics.

5.3. Data to Be Used in the Evaluation of Application Results

The evaluation of machine learning algorithms is made using variables in the confusion matrix. Four variables, called True Positive, True Negative, False Positive and False Negative, form the basis of the calculations to decide which supervised learning algorithm is better. The evaluation criteria of confusion matrix are shown in Table 2 (Nguyen & Armitage, 2008). Diagonal cells in the confusion matrix show the number of correctly detected data, while the other cells show the number of false detections (Deshmukh et al., 2015).

Table 2. Confusion matrix evaluation criteria.

Confusion Matrix		**Predicted Class**	
		A	**Ā**
Real Class	**A**	True Positive	False Negative
	Ā	False Positive	True Negative

- True Positive (TP) is the number of data that actually belong to class A and are predicted to belong to class A.

- True Negative (TN) is the number of data that do not actually belong to Class A and are predicted to not belong to Class A.

- False Positive (FP) is the number of data that do not actually belong to Class A but are predicted to belong to Class A.

- False Negative (FN) is the number of data that actually belong to Class A but are predicted not to belong to Class A.

Using these four variables, data such as Accuracy, Precision, Recall and F-measure are calculated. Thus the performance evaluation of supervised learning algorithms is performed (Kaya, 2016; Yiğidim, 2012). Accuracy is expressed as the ratio of correctly estimated data to total data. It is an important criteria that reveals the performance of the classification algorithm. Precision is the ratio of correctly predicted data to the total number of predicted data. Recall is the ratio of correctly predicted data to the actual number of data belonging to that class. It shows at what rate the algorithm correctly predicts the data. F-measure is calculated by taking the harmonic average of the precision and recall data. Therefore, instead of using both data separately, a comparison of supervised learning algorithms can be made by using this data.

5.4. Application of Supervised and Unsupervised Learning Algorithms

In order to use supervised and unsupervised learning algorithms with the WEKA application, we perform some pre-processes on the data set. These processes are explained below.

As a matter of fact, the NSL KDD data set contains three types of data (Nominal, Numerical and Binary) and two values as "Normal" and "Anomaly" as data label. In this study, in order to obtain results that close to the real situation, these label data were converted to five values as "Normal", "DoS", "U2R", "R2L" and "Probing" for both training and test data. Likewise, data types in the data set have been converted into suitable types for the algorithm used.

To examine the effect of algorithm performances on the entire NSL KDD dataset and on the reduced data, dimensionality reduction is applied to data set attributes and reduced to six attributes. The variance of the aforementioned six attributes is

0.90. That variance means that 125973 data in the data set can be expressed with six attributes with the rate of 0.90.

WEKA application uses four different methods for model creation and testing processes: "use training set", "supplied test set", "k-fold cross-validation" and "percentage split". The most preferred method among these four methods is the "k-fold cross-validation" method (Kohavi, 1995).

With "k-fold cross-validation" method, the training data set given to the WEKA application is divided into k parts, one part is used for testing and the other parts are used to create a model, and the process is repeated k times. During this study, the data set with forty-one attributes and the data set with six attributes, as a result of the dimensionality reduction process, were used with "k-fold cross-validation" and "supplied test set" methods while testing the below-mentioned machine learning algorithms.

5.4.1. k-Nearest Neighbor (k-NN) Algorithm

In the k-nearest neighbor algorithm, the data are classified by calculating the distance by taking into account k number of close neighbors. The most frequently used functions in distance calculation are Euclidean and Manhattan functions (Zhang, 2016). The default distance function of the k-nearest neighbor algorithm in the WEKA application is the Euclidean function.

Our tests for the performance evaluation of the k-nearest neighbor algorithm was carried out by selecting two different values as k = 1 and k = 5 (1 and 5 close neighbors).

5.4.2. Decision Tree Algorithm

In the decision tree algorithm, the class of the data is determined using the decision tree created from the training data set. While creating the decision tree, the root node is determined first. When determining the root node, the feature that best separates the samples is selected. Then, the structure of the tree is determined by repeating this process in leaf nodes (Aksu & Doğan, 2019).

J48 (C4.5) decision tree algorithm in the WEKA application is used in the our tests for performance evaluation.

5.4.3. Artificial Neural Networks (ANN) Algorithm

Artificial neural networks algorithm has a structure consisting of at least three layers: input layer, middle layer (hidden layer) and exit layer. The intermediate layer can be at least one layer or it can consist of more than one layer. In the artificial neural network, learning is provided by back propagation and the threshold function. The algorithm also includes the momentum coefficient and learning rate variables used in updating the weights (Arı & Berberler, 2017).

For the performance evaluation of the artificial neural networks algorithm, the default values of the chosen multi-layer perceptron algorithm in the WEKA application (hidden layer number 23, sigmoid number 67, momentum 0.2 and learning ratio 0.3) are used.

5.4.4. Logistic Regression Algorithm

The relationship between variables is expressed as a nonlinear "S" shaped curve of the logit model. The curve in question is drawn by calculating the distances of the variables from the curve logarithmically (Ürük, 2007). In our tests, the logit model is used in the calculation of the logistic regression algorithm in the WEKA application.

5.4.5. k-means Clustering Algorithm

In the k-means clustering algorithm, first k objects that form the center of the clusters are selected. Then the distances of other objects to central objects are calculated by using a distance metric like the Euclidean distance function. As a result of the calculation, clusters are formed and the new centers of the formed clusters are determined. This process continues iteratively until the center update process of the clusters ends (Na et al., 2010).

During the implementation of the k-means clustering algorithm in our tests, k = 2 was chosen considering the "Normal" and "Anomaly" network traffic.

5.4.6. Apriori Algorithm

The Apriori algorithm is an association algorithm that works with inductive logic. The algorithm first determines the usage frequency of the data in the data set and makes associations between the most frequently used data. In order for the association rule to be formed, the minimum support and minimum trust criteria must be met (Al-Maolegi & Arkok, 2014). The biggest feature that distinguishes

association algorithms from other learning algorithms is that it works successfully with categorical data as well as numerical data.

In our tests, the "KDDTrain" data set was pre-processed by using the feature selection and filtering capabilities of the WEKA application.

6. RESULTS AND EVALUATIONS

The results of our tests are presented by comparing the working time of the algorithms, accuracy and F-measure values.

We first apply the "supplied test set" method of the WEKA application. In the "supplied test set" method, model training is done with the "KDDTrain" data set. The classification performance evaluation of the trained model is carried out with the "KDDTest" data set. We observe that the algorithms classify at a very close accuracy rate with each other as in Table 3. However, the working time of some algorithms is too long when compared to the others. On the other hand, dimension reduction adversely affects the classification performance whereas the processing speed improves with respect to the data with forty-one features.

Table 3. "SuppliedTest Set" method application results evaluation.

Algorithm		**Time (s)**		**Accuracy (%)**	
		41 features	**6 features**	**41 features**	**6 features**
k-Nearest Neighbor	1-NN	676,71	410,16	77,09	39,66
	5-NN	677,51	390,13	76,90	41,37
Decision Tree	J48 (C4.5)	37,33	5,97	75,26	45,58
ANN	MLP	10318,14	91,92	75,54	43,08
Regression	Logistic	58,69	4,6	75,61	44,22

In this method, it is seen that the decision tree algorithm has the shortest processing time when both forty-one features and six features are used.

When the k-nearest neighbor algorithms are examined, there is not much difference in terms of time and accuracy. But, when six features used, it is understood that time is shortened. In addition, when six features are used, it has been determined that the multi-layer perceptron algorithm makes classification in a much faster time than forty-one features.

Since the "k-fold cross-validation" method splits the "KDDTrain" data set into "k" folds and use that folds to learn, it makes a very high rate of correct classification. It

is seen that the processing times are longer for both forty-one features and for six features compared to the "supplied test set" method.

However, when the results obtained by using six features are examined, it is seen that the correct classification rates are close to the results obtained by using forty-one features, unlike the "supplied test set" method. When Table 4 is examined, it is seen that the algorithm with the shortest processing time and the highest accuracy data is the decision tree algorithm.

Table 4. "10-fold cross-validation" method application results evaluation.

Algorithm		**Time (s)**		**Accuracy (%)**	
		41 features	**6 features**	**41 features**	**6 features**
k-Nearest Neighbor	1-NN	2998,7	2670,38	99,72	99,39
	5-NN	2267,13	1840,23	99,57	99,18
Decision Tree	J48 (C4.5)	457,15	79,02	99,76	99,09
ANN	MLP	82045,11	993,96	99,02	96,39
Regression	Logistic	664,63	47,51	97,50	92,35

It has been determined that the F-measure achieved a high rate of success in the classification of data with "Normal", "DoS" and "Probing" labeled data. When compared the "supplied test set" method (Figure 3 and Figure 6) with the "10-fold cross-validation" method (Figure 4 and Figure 7), it was observed that the F-measure data has lower results in the "supplied test set" method, as in the accuracy data.

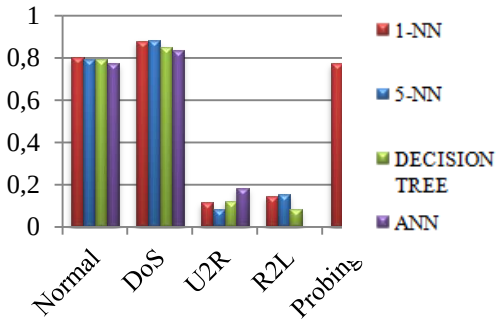

Figure 3. "Supplied Test Set" method F-measure data (41 features).

As can be seen in Figure 3, k-NN, Decision Tree and ANN algorithms used in the "supplied test set" method have shown low success in classifying data with "U2R" and "R2L" labeled data. On the other hand, in the "10-fold cross-validation" method (Figure 4), the accuracy rate in the classification of data with "U2R" and "R2L" labeled data is higher, but the rates were not as high as in the data with the other three labeled data.

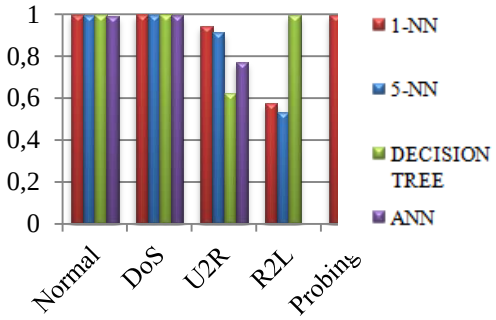

Figure 4. "10-fold cross-validation" method F-measure data (41 Features).

In both figures, it is seen that there is no "R2L" labeled F-measure data for the artificial neural network algorithm. The reason for this is that the classifier cannot classify the data correctly with "R2L" label.

When the F-measure values of the Logistic Regression algorithm were examined (Figure 5), it was observed that the data obtained in the "supplied test set" method was lower than the other supervised learning algorithms, and it was classified at almost the same rates in the "10-fold cross-validation" method.

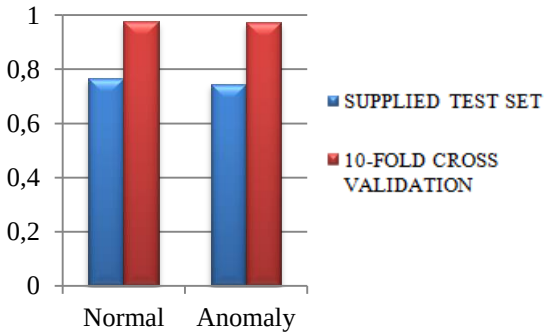

Figure 5. Logistic regression algorithm F-measure data (41 features).

The F-measure data, from the classification using six features obtained as a result of dimension reduction is as in Figure 6, Figure 7 and Figure 8. Algorithms run with "supplied test set" method by using six features showed very low results in F-measure data as well as in accuracy data. When Figure 6 is examined, it is seen that the F-measure data give partial results in the data with "Normal" labeled data.

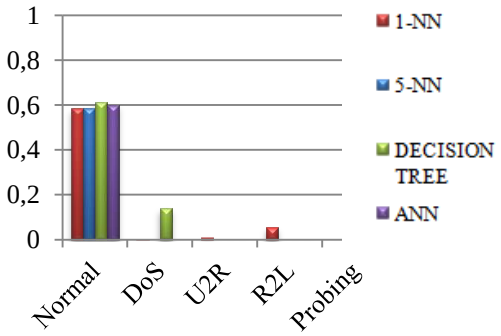

Figure 6. "Supplied Test Set" method F-measure data (6 features).

When Figure 7 is examined, it is seen that the results obtained in the "10-fold cross-validation" method using six features are high, but lower than the results obtained using forty-one features.

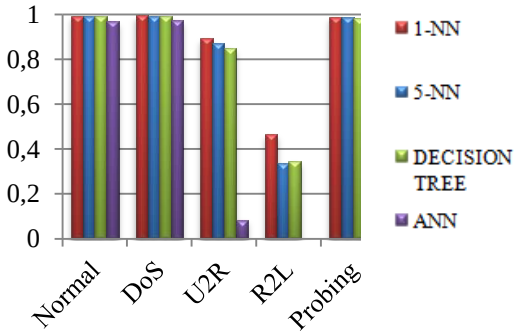

Figure 7. "10-fold cross-validation" method F-measure data (6 features).

When the F-measure values of the Logistic Regression algorithm, which is run using six features, are examined (Figure 8), it is seen that the data obtained in the "supplied test set" method is lower than the other supervised learning algorithms, and the classification in the "10-fold cross-validation" method is almost the same. In the Logistic Regression algorithm, which is run using six features, as in other algorithms, it has been found that a lower rate of success is achieved compared to the results obtained by using forty-one features.

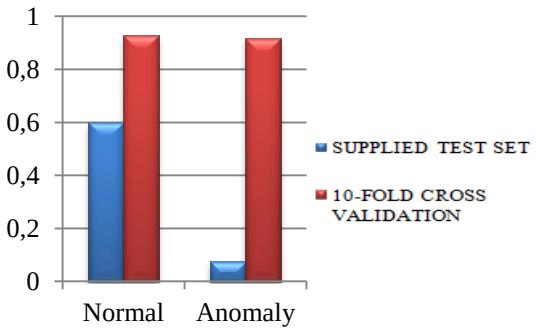

Figure 8. Logistic regression algorithm F-measure data (6 features).

When unsupervised learning algorithms are examined, as a result of the application results of the k-means clustering algorithm using forty-one features, it is seen that it performs a clustering process performed with the "KDDTrain" data set, is 50.3% in 6.05 seconds, and which is run using six features, is 52.1% in 4.99 seconds. Although the k-means clustering algorithm, which is run using forty-one features, provides an advantage in terms of time compared to supervised learning algorithms, it is seen that its accuracy rate is very low. On the other hand, it was determined that the correct classification rate obtained with the k-means clustering algorithm using six features is higher than the correct classification rates obtained in supervised learning algorithms with "supplied test set" method using six features.

The Apriori algorithm, on the other hand, differed from other learning algorithms because it was supported with a feature selection algorithm and subjected to filtering before the application, and it made associations at a reliability level ranging from 70% to 90% in 600.38 seconds. Unlike other algorithms, it can be ensured that the reliability ratio between the data to be correlated is higher in the Apriori algorithm. In other words, the reliability ratio, which is between 70% and

90% in the exemplary application, can be determined to be lower or higher. However, this situation corresponds to more time for a higher reliability rate.

As a result, when the results of all learning algorithms examined within the scope of this study; The fastest working learning algorithm with the lowest classification rate is the clustering algorithm. The Apriori algorithm, which is another unsupervised learning algorithm, can make associations at the desired accuracy rates, but it must be subjected to some pre-processes before this process. When evaluated in terms of time, it has been observed that the process is close to the supervised learning algorithms. It was determined that the 5-Near Neighbor algorithm, one of the supervised learning algorithms, performs better classification in the "supplied test set" method compared to the other algorithms, while the decision tree algorithm performs better in the "10-fold cross-validation method. In general, it has been observed that supervised learning algorithms classify at close accuracy rates, but the algorithms mentioned above are faster than the others in terms of time.

As a result of the classification process performed with the shape of the NSL KDD data set containing forty-one features and six features subjected to the dimensionality reduction process (Principal Component Analysis), it was observed that similar results emerged, and consistent data were obtained considering the applied test methods. It is possible to explain, why the data subjected to the dimensionality reduction process with using "supplied test set" method cannot be classified at desired level, with developing different attack techniques. As a matter of fact, the types of attacks that the learning algorithm "learns" appear in different types day by day. Moreover, the results obtained with the 10-fold cross-validation method are also a positive inference in terms of the saving in time as a result of the dimensionality reduction process.

Apparently, there is a limited amount of data that can be used to compare supervised and unsupervised learning algorithms. Although this situation makes it difficult to choose between algorithms, it is evaluated that the learning algorithm to be used for detecting threats/attacks against network-based computer systems should be preferred among the supervised learning algorithms and pre-processing of the data (Dimensionality reduction, feature selection, etc.) to be used will show high performance in terms of time and correct classification.

7. CONCLUSIONS AND DISCUSSIONS

Today, with the developing technology and increasing internet usage, the security need of computer systems is constantly increasing. It is of great importance to use machine learning in intrusion detection systems developed to meet the increasing security need.

As a matter of fact, attacks on computer systems occur because of the reasons (Zhang et al., 2012):

- Attackers who want unauthorized access to the system;

- Users who are authorized in the system, want to gain additional privileges in matters that they are not authorized, and;

- Misuse of privileges granted to authorized users.

Considering the aforementioned reasons for the attacks, it becomes inevitable to be faced with new types of attacks every day. For this reason, the use of machine learning in intrusion detection systems increases the functionality of attack detection systems and enables new types of attacks to be detected as soon as possible.

Within the scope of this study, the performance evaluation of supervised and unsupervised learning algorithms has been performed by examining intrusion detection systems and machine learning. In order to run supervised and unsupervised learning algorithms, the WEKA application and the NSL KDD data set derived from the KDD CUP-99 data set, which is the most frequently used data set in the literature, were used (Kaya & Yıldız, 2014).

As a result of the measurements carried out, it has been determined that the probability of detecting threats/ attacks against network-based computer systems is higher with supervised learning algorithms. It was observed that the classification rates of the algorithms were close to each other, but the processing times varied, in the two different test methods performed with supervised learning algorithms using the NSL KDD data set.

On the other hand, it has been determined that unsupervised learning algorithms are fast in terms of processing time but have low accuracy rates. Another disadvantage of unsupervised learning algorithms is that they cannot classify. Although they

perform clustering or association process using data, it must be processed once again in order to interpret the outputs of unsupervised learning algorithms.

In the experiments conducted using the NSL KDD data set that was pre-processed (dimensioanlity reduction), it was observed that the "supplied test set" method had a very low classification rate. In the "10-fold cross-validation" method, it was found that it gave similar results to applications with non-preprocessed data, but the algorithm runtime was shorter.

In future studies to be carried out within the scope of detecting threats/attacks against network-based computer systems, it is necessary to examine a combined algorithm in which unsupervised and supervised learning algorithms can work together, as well as to examine the data that will be input to the algorithms such as dimensionality reduction and/or feature selection. It is considered that the implementation of the aforementioned procedures will be beneficial.

As in the example of the Apriori algorithm, it is evaluated that if the results from the unsupervised learning algorithm are applied as input to the supervised learning algorithm, the results of combined algorithm will yield more positive results than the results of the supervised learning algorithm alone, and the threat/attack can be detected in a shorter time depending on the algorithm selection.

Likewise, it is evaluated that applying pre-processes such as dimensionality reduction and/or feature selection to the data set that will be input to the algorithms can increase performance, shorten the processing time of the algorithms and lead to more precise results.

ACKNOWLEDGEMENT

This article is extracted from the Master of Science (M.Sc.) thesis, which is entitled as "Classification of Threats/Attacks Against Network-based Computer Systems with Supervised Machine Learning" at the Barbaros Naval Sciences and Engineering Institute of National Defence University in January, 2022.

REFERENCES

Aksu, G., & Doğan, N. (2019). "Comparison of decision trees used in data mining". *Pegem Eğitim ve Öğretim Dergisi*, 9(4), 1183-1208. doi:10.14527/pegegog.2019.039.

Allen, J., Christie, A., Fithen, W., McHugh, J., Pickel, J., & Stoner, E. (2000). *State of the Practice of Intrusion Detection Technologies*. Carnegie Mellon University Technical Report CMU/SEI-99-TR-028.

Al-Maolegi, M., & Arkok, B. (2014). "An improved apriori algorithm for association rules". *International Journal on Natural Language Computing*, 3(1), 21-29. doi:10.5121/ijnlc.2014.3103.

Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press.

Andress, J. (2011). *The Basics of Information Security: Understanding the Fundamentals of Infosec in Theory and Practice*. Syngress.

Arı, A., & Berberler, M. E. (2017). "Yapay sinir ağları ile tahmin ve sınıflandırma problemlerinin çözümü için arayüz tasarımı". *Acta Infologica*, 1(2), 55-73.

Avcı, U., & Avcı, M. (2004). "Örgütlerde bilginin önemi ve bilgi yönetimi süreci". *Mevzuat Dergisi*. Vol. 7, No. 74. Retrieved from http://www.mevzuatdergisi.com/2004/02a/01.htm

Bellinger, G., Castro, D., & Mills, A. (2004). "Data, Information, Knowledge & Wisdom". Retrieved from http://www.systems-thinking.org/dikw/dikw.htm

Brownlee, J. (2017). *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. Machine Learning Mastery.

Brynjolfsson, E., & Mitchell, T. (2017). "What can machine learning do? Workforce implications". *Science*, 358(6370), 1530-1534. doi:10.1126/science.aap8062.

Chae, H., Jo, B., Choi, S., & Park, T. (2013). "Feature selection for intrusion detection using NSL-KDD". *Recent Advances in Computer Science*, 184-187.

Deshmukh, D. H., Ghorpade, T., & Padiya, P. (2015). "Improving classification using preprocessing and machine learning algorithms on NSL-KDD dataset". *2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, 1-6. doi:10.1109/iccict.2015.7045674.

Dhanabal, L., & Shantharajah, S. P. (2015). "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms". *International Journal of Advanced Research in Computer and Communication Engineering*, 4(6), 446-452.

Ersoy, E. V. (2012). *ISO/IEC 27001 Bilgi Güvenliği Standardı*. ODTÜ Geliştirme Vakfı Yayıncılık ve İletişim.

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. doi:10.1016/j.jisa.2019.102419.

İlter, H. K. (2011). "Bilgeliğe giden yol mideden geçer mi?" *PiVOLKA*, 20(6), 3-7.

Karataş, G., Demir, O., & Şahingöz, O. K. (2018). "Deep learning in intrusion detection systems". *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*. doi:10.1109/ibigdelft.2018.8625278.

Kaya, Ç. (2016). *Saldırı tespit sistemlerinde makine öğrenmesi tekniklerinin kullanılması: Karşılaştırmalı performans analizi* (Master's Thesis). Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Kaya, Ç., & Yıldız, O. (2014). "Makine öğrenmesi teknikleriyle saldırı tespiti: Karşılaştırmalı analiz". *Marmara University Journal of Science*, 26(3), 108. doi:10.7240/mufbed.24684.

Kocabıyık, L. (2005). *Information and knowledge management in the military domain* (Master's Thesis). Vrije Universiteit Brussel, Faculty of Economic, Social and Political Sciences, Brussels.

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". *International Joint Conference on Artificial Intelligence*, 14(12), 1137-1143.

Kramer, S., & Bradfield, J. C. (2009). "A general definition of malware". *Journal in Computer Virology*, 6(2), 105-114. doi:10.1007/s11416-009-0137-1.

Kurgun, O. A. (2006). "Bilgi yönetim sistemlerinin yapılandırılması". *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 8(1), 274-291.

Liu, G. G. (2014). "Intrusion detection systems". *Applied Mechanics and Materials*, 596, 852-855. doi:10.4028/www.scientific.net/amm.596.852.

Na, S., Xumin, L., & Yong, G. (2010). "Research on k-means clustering algorithm: An improved k-means clustering algorithm". *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 63-67. doi:10.1109/iitsi.2010.74.

Nguyen, T. T., & Armitage, G. (2008). "A survey of techniques for internet traffic classification using machine learning". *IEEE Communications Surveys & Tutorials*, 10(4), 56-76. doi:10.1109/surv.2008.080406.

Önaçan, M. B. K. (2015). *Organizasyonlar için bilgi yönetimi çerçevesi ve bilgi yönetim sistemi mimarisi önerisi: Doblyn (Doküman ve bilgi yönetimi)* (Ph.D. Thesis). Ankara Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Simeone, O. (2018). "A brief introduction to machine learning for engineers". *Foundations and Trends in Signal Processing*, 12(3-4), 200-431. doi:10.1561/2000000102.

Solomon, M. G., & Chapple, M. (2005). *Information Security Illuminated*. Jones and Bartlett.

Taher, K. A., Jisan, B. M. Y., & Rahman M. M. (2019). "Network intrusion detection using supervised machine learning technique with feature selection". International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 643-646. doi:10.1109/ICREST.2019.8644161.

Thomas, R., & Pavithran, D. (2018). "A survey of intrusion detection models based on NSL-KDD data set". *2018 Fifth HCT Information Technology Trends (ITT)*, 286-291. doi:10.1109/ctit.2018.8649498.

Ürük, E. (2007). *İstatistiksel uygulamalarda lojistik regresyon analizi* (Master's Thesis). Marmara Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Witten, I. H., Hall, M., Frank, E., Holmes, G., Pfahringer, B., & Reutemann, P. (2009). "The WEKA data mining software". *ACM SIGKDD Explorations Newsletter*, 11(1), 10-18. doi:10.1145/1656274.1656278.

Yiğidim, H. A. (2012). *Makine öğrenme algoritmalarını kullanarak ağ trafiğinin sınıflandırılması* (Master's Thesis). TOBB Ekonomi ve Teknoloji Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Yost, J. R. (2016). "The march of IDES: Early history of Intrusion-Detection expert systems". *IEEE Annals of the History of Computing*, 38(4), 42-54. doi:10.1109/mahc.2015.41.

Zhang, Z. (2016). "Introduction to machine learning: K-nearest neighbors". *Annals of Translational Medicine*, 4(11), 218. doi:10.21037/atm.2016.03.37.

Zhang, X., Jia, L., Shi, H., Tang, Z., & Wang, X. (2012). "The application of machine learning methods to intrusion detection". *2012 Spring Congress on Engineering and Technology*, 1-4. doi:10.1109/scet.2012.6341943.