Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 29, S. 19-25, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 29, pp. 19-25, December 2021

Research Article

Detecting Internet of Things Attacks by Using Hybrid Learning and Feature Selection Method

Gozde Karatas[1*]

[1] Biruni Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0003-2303-9410)

(International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2021 – 21-23 October 2021)

(**DOI:** 10.31590/ejosat.1017433)

ATIF/REFERENCE: Karatas, g. (2020). Detecting Internet of Things Attacks by Using Hybrid Learning and Feature Selection Method. *European Journal of Science and Technology*, (29), 19-25.

Abstract

Internet of Things (IoT) produces an enormous amount of data, which is used in all areas of our lives and increases the number of data on the Internet with each passing day. Smart watches, robot vacuum cleaners, refrigerators with cameras, and more can all be considered IoT devices. Ease of access to the Internet provides people with advantages as well as disadvantages. Malware and intruders have easier access to the devices we use and our information via the internet. At this point, data security gains great importance especially in IoT devices because accessing our personal data via smart watches or refrigerators we use can pose a great threat to individuals and their families. This study focus the importance of data preprocessing and developing a hybrid machine learning-based intrusion detection system (IDS) for IoT. Decision Tree, which is a popular machine learning algorithm, and n_Balot dataset were preferred for investigations. Accordingly, it is aimed to create a hybrid model by applying K-means and Decision Tree algorithms to the n_Balot dataset with under sampling and feature selection. In the data preprocessing, feature selection was performed with Chi-Square method and under sampling performed with RandomOverSampling method. Then, clustering was done by applying K-means to the processed dataset, and the results obtained with the clustering were classified with the Decision tree algorithm. As a result of the study, while the error rate was 0.39% in the predictions made only with the decision tree, the error rate was reduced to 0.01% with the developed hybrid model.

Keywords: Intrusion detection, n_Balot dataset, Decision tree, K-means, Feature selection.

Nesnelerin İnterneti Saldırılarının Hibrit Öğrenme ve Özellik Seçimi Yöntemi Kullanılarak Tespiti

Öz

Nesnelerin interneti (IoT), hayatımızın her alanında kullanılan ve her geçen gün internetteki veri sayısını artıran muazzam miktarda veri üretmektedir. Akıllı saatler, robot süpürgeler, kameralı buzdolapları ve daha kullanılan birçok cihaz IoT cihazları olarak kabul edilebilir. Ayrıca gelişen teknoloji ile birlikte hayatımızın her alanında olan internete erişim kolaylığı insanlara avantajlar sağladığı gibi dezavantajlar da sağlamaktadır. Kötü amaçlı yazılımlar ve saldırganlar, yoğun olarak kullandığımız cihazlara ve önemli bilgilerimize internet üzerinden daha kolay erişebilmektedir. Bu noktada özellikle IoT cihazlarında veri gizliliği ve güvenliği büyük önem kazanmaktadır çünkü kullandığımız akıllı saatler veya kullandığımız buzdolapları aracılığıyla kişisel verilerimize erişim bireyler ve aileleri için büyük bir tehdit oluşturabilmektedir. Tüm bu durumlar göz önüne alındığında bu çalışma, veri ön işlemenin önemine ve IoT cihazları için hibrit bir makine öğrenmesi tabanlı saldırı tespit sistemi (IDS) geliştirmeye odaklanmaktadır. Çalışmada yapılacak araştırmalar için popüler bir makine öğrenme algoritması olan Karar Ağacı ve n_Balot veri kümesi tercih edilmiştir. Buna göre veri azaltma işlemi ve özellik seçimi ile n_Balot veri kümesine K-means ve Karar Ağacı algoritmaları uygulanarak saldırı tespiti yapan hibrit bir model oluşturulması amaçlanmıştır. Veri ön işlemede, Ki-Kare seçim yöntemi ile özellik seçimi ve RandomOverSampling yöntemi ile veri azaltma işlemleri yapılmıştır. Daha sonra veri sayısı azaltılmış ve özellik seçimi gerçeklenerek işlenmiş veri kümesine K-Means algoritması uygulanarak kümeleme yapılmış ve kümeleme ile elde edilen sonuçlar Karar ağacı algoritması ile sınıflandırılmıştır. Yapılan tüm incelemeler sonucunda hiçbir işlem yapılmadan yani veri ön işleme ve özellik seçimi gerçekleştirmeden sadece Karar Ağacı ile yapılan tahminlerde hata oranı %0,39 iken, geliştirilen hibrit model ile hata oranı %0,01'e düşürülmüştür.

Anahtar Kelimeler: Intrusion detection, n_Balot dataset, Decision tree, K-means, Feature selection.

[*] Corresponding Author: Biruni Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye, ORCID: 0000-0003-2303-9410, gbaydogmus@biruni.edu.tr

1. Introduction

The concept of the Internet of Things (IoT) is the name given to all systems that can transfer data over the network without the need for human beings with digital devices that are related to each other. IoT is a concept that was first introduced by Kevin Ashton in 1991 and was used to explain the communication of countless electronic devices such as mirrors, wrist watches, refrigerators, televisions with the Internet (Ashton, 2019). Increasingly, organizations in different industries are using IoT technology to work more efficiently, improve business quality, and improve decision-making processes. Figure 1 shows popular IoT devices.

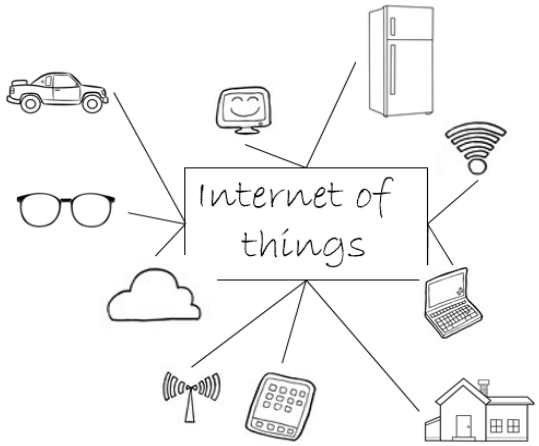

Figure 1. IoT Devices

IoT devices take advantage of internet-connected hardware that uses sensors and communication hardware to collect data and process it (Syms et al., 2020; Anthi et al., 2019). In other words, IoT is a network of devices connected through communication technologies in order to create systems that can provide information and analyze this data that enable researchers to make research or decisions on any subject much faster. After the computer entered human life, many concepts began to gain importance. The most notable of these is computer security. Every device connected to the internet is included in this concept when it comes to computer security, so ensuring the security of IoT devices has become an important work area which have become popular in recent years.

There are not only good people who will do useful things on the internet, but there are also people who aim to misuse people's information and even want to do worse things. These are generally referred to as attackers/hackers/intruders. These people threaten personal security by using the benefits of the internet and the existing security vulnerabilities. Therefore, security should be the most important area in every developed system. However, companies do not prioritize data and computer security, especially in terms of cost. Security area in computer sciences is commonly referred to as the Intrusion Detection System (IDS). Systems that detect attacks on IoT devices (or computers) in advance and intervene if necessary take this name. By using IDSs, the security of computers and IoT devices can be ensured.

In this study, n_Balot, a popular IoT dataset was used and the effect of feature selection and clustering on detection rate in attacks on IoT devices was examined. In this context, firstly K-means algorithm was applied to the preprocessed dataset and then K-means algorithm was applied to the subset of the dataset which is created by feature selection. The number of features was determined intuitively and it was observed that there was no change in the accuracy rate until the number of features decreased to 5. On top of that, after applying K-means to the 5-featured subset, the resulting clustered data was trained with the Decision Tree machine learning algorithm and it was seen that the accuracy rate was higher than the system trained with Decision Tree without any processing. In addition, while performing all these operations, under sampling was performed on the dataset and the results were obtained. The reason for this is that even if the number of data decreases, there is no change in the performance rate. In this way, the detection speed has also been increased.

Rest of the paper is organized as follows: Section II covers the related work on Intrusion Detection on IoT devices. Information about the methods used in the study is given in Section III. Section IV presents the new clustering based IDS and experimental results. Finally, the conclusion and discussion is given Section V.

2. Material and Method

In this section, related work that are similar to the study and the proposed system is given. The dataset, feature selection, data reduction, data preprocessing and machine learning algorithms used in the proposed system are detailed.

2.1. Related Work

In this section, previous studies on Intrusion detection in different electronic databases such as Elsevier, Springer and Ieee Explorer are examined. 10 of the studies related to the subject were selected and information was given about the techniques and architectures they contain.

IoT device security is one of the most important issues. Aware of this, researchers designed a hybrid intrusion detection system for IoT devices in 2020 (Syms et al., 2020). The focus of the study is to classify the request to the device as attack or not, at this point, they focused on machine learning and deep learning concepts, which are rapidly increasing in popularity today. As a result, they proposed a model in the form of hybrid convolutional neural network. They used the UNSW-NB15 dataset for development and compared their proposed model with another existing learning algorithm, RNN. As a result of the examinations, they have seen that the proposed hybrid model is much better in terms of both time and accuracy.

Anthi and others proposed a three-layer intrusion detection system (IDS) to detect popular attacks on IoT in 2019 (Anthi et al., 2019). The proposed system consists of three main components: classifying the IoT devices and their profile connected to the Internet, identifying harmful requests on the network when an attack occurs, and classifying the type of incoming attack. In order to develop and examine the system, a real test environment consisting of various IoT devices was created and dataset was collected with different attacks. As a result of the study, it has been seen that the proposed system can automatically detect attacks on IoT devices and detect which attack has been applied to a device on the network. The F-

measurement of the developed system varies between 90% and 98%.

In 2016, researchers aimed to develop a model for intrusion detection on IoT devices, the core component of which is a core component (Sforzin et al., 2016). Their main focus has been examine the performance of Raspberry Pi computer working with open source Snort and make recommendations accordingly. For development and testing, they used network traffic recorded in trace files. As a result of the investigations, they have seen that the Raspberry Pi computer can be used effectively to detect attacks on IoT devices.

Raza et al. developed a model called SVELTE to detect attacks on IoT devices (Raza et al., 2013). To develop this approach, they studied the IPv6 and 6LoWPAN packet structures and transmitted on them. Their main goal is to detect information attacks on the network device. As a result of the studies, it was seen that the developed system successfully detected the test attacks. They aim to completely zero the misunderstanding rate for future studies.

The concept of IoT, which has entered our lives in the last few years, is both a very popular and a new concept, so it is very vulnerable to attacks. In 2016, researchers examined the attack analysis of IoT devices and proposed a new model (Hodo et al., 2016). For the developed model, it is aimed to detect and prevent DoS attacks by using artificial neural networks. Therefore, they have classified incoming requests as attack and normal. They used a dataset with 2313 samples for the study and reached a success rate of 99.4% as a result of the study.

Yang et al. have examined the main challenges in wireless intrusion detection, an active learning approach that builds on its core concepts (Yand et al., 2018). They compared traditional learning methods with active learning methods and presented results with an experimental example. In their study, they aimed to help those who research active learning and to encourage research.

In 2020, researchers made a study to emphasize how important to design a system while developing an intrusion detection system for IoT (Derhab et al., 2020). Since the main purpose of the study was to design an intrusion detection system, they created a hybrid model by combining convolutional neural network (CNN) which a deep learning approach with causal convolution, and they called it Temporal Convolution Neural Network (TCNN). They used Bot-Iot dataset for this work. Since the Bot-IoT dataset used in the study is unbalanced, SMOTE technique was used to obtain an efficient model by balancing the dataset, and feature selection methods were also carried out. The results of the study were compared with Logistic Regression, Random Forest, Long short Term Memory and CNN algorithms in terms of accuracy and efficiency and it was seen that the proposed system achieved 99.99% performance.

Parra and his friends created a cloud-based distributed deep learning model that will detect phishing and botnet attacks to maintain security on IoT devices (Parra et al., 2020). The model consists of two basic mechanisms; (1) a built-in Distributed Convolutional Neural Network (DCNN) model to detect phishing and DoS attacks; and (2) a cloud-based temporary Long-Short-Term Memory (LSTM) network model for detecting botnet attacks. They created a dataset of both phishing and non-phishing URLs to train the proposed DCNN plug-in model, and used the existing N_BaIoT dataset to train the backend LSTM model. As a result of experimental studies, it has been seen that the proposed DCNN model can detect phishing attacks with 94.3% accuracy and 93.58% F-1 score, while the back-end LSTM model detects Botnet attacks with 94.80% accuracy.

Some studies show that objective function selection is effective in providing security in IoT devices. Furthermore, the available literature lacks examining vulnerability analysis of objective functions, especially in combined attacks. Therefore, the researchers examined the vulnerability analysis of two different objective functions with machine learning approaches to detect combined attacks against IoT devices through different scenarios (Foley et al., 2020). They created a new IoT dataset for the study as part of the RPL IDS/IPS solution. As a result of the studies, the objective function machine learning approach was successful in detecting combined attacks based on power and network measurements.

Dataset plays an important role when developing models to ensure security in IoT devices. Alsaedi and others proposed a new labeled dataset for a type feature, showing a subclass of attacks on IoT devices (Alsaedi et al., 2020). The proposed dataset is named TON-IoT. The study also explains the proposed dataset and features of Telemetry data of IoT services. TON_IoT currently has the advantages of having attack and non-attack and evenly distributed data for different IoT/IIoT services. Using the proposed dataset, intrusion detection was implemented with various machine learning and deep learning approaches, and the success of the dataset was tested.

2.2. Proposed Approach

In this section, information is given about the methods and dataset used while developing the study.

2.2.1. K-Means Algorithm

We may not always have meaningful data, so we don't know how to relate the data. Clustering, which is an unsupervised learning method, on such datasets provides better results (Likas et al., 2003; Hartigan and Wong, 1979). Clustering algorithms aim to establish relationships between such data. Figure 2 shows the state of the data before and after clustering.

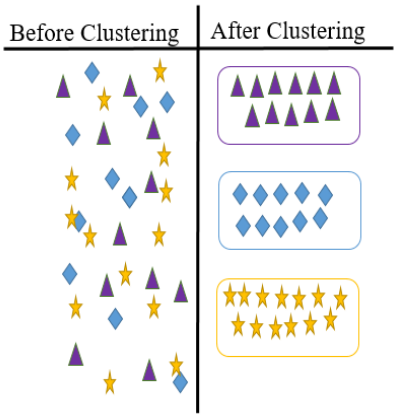

Figure 2. Clustering

K-means algorithm is a clustering algorithm. Its main purpose is to determine the center points by minimizing the distance within the cluster, for this it assumes that the data at hand consists of k clusters and tries to minimize the distances of the points included in the clusters to be formed to the cluster

mean. The algorithm makes clustering based on the Euclidean distance formula given by the Formula (1);

$$\sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} = \sqrt{(x_1 - y_1)^2 - \cdots - (x_N - y_N)^2} \quad (1)$$

Where N is the size of the dataset, and $X = (x_1, x_2, \ldots, x_N)$ and $Y = (y_1, y_2, \ldots, y_N)$ are the points. The algorithm performs the following steps;

1. Start with randomly selected K (number of clusters) central points.
2. Assign each point in the dataset to the set of the center point closest to it (based on the distance calculated by the Euclidean distance).
3. The value of the cluster center is calculated by taking the average of all its own points.
4. This process continues until the location and value of the centers do not change.

Application areas of the algorithm; customer segmentation, game/player analysis, document classification, intrusion/fraud detection. The main problem here is to choose the K value correctly. Because it is necessary to manually (heuristically) assign the value of k. There are several methods for determining the appropriate number K, these are; Elbow Method, Average Silhouette Indeks, GAP statistic. Since the K value was determined according to the number of classes in this study, any of the suggested methods was not used.

2.2.2. Feature Selection (FS)

Features are information about each data in a dataset. For example, the model, color, brand and mileage of a car are features. Feature selection is the process of selecting and finding the most useful features for operations within the dataset (Korkmaz et al., 2020; Korkmaz et al. 2020). The purpose of feature selection is to speed up the training time of the model and make it more meaningful. Feature selection methods are named in three main categories; Embedded Methods, Wrapper Based Methods, and Filter Methods. Each method can be diversified within itself. Since the chi-square test, which is a filter method, is used in this study, detailed information is given only about it.

Chi-Square; It is a method that performs operations on the basis of whether the difference between the observed and expected values is significant and works with categorical data (Tallarida et al., 1987). It is tested whether there is a relationship between the features and the target, and the features that cannot be found are removed from the dataset. It is calculated with Formula (2).

$$X^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i} = N\sum_{i=1}^{n}\frac{(O_i/N - p_i)^2}{p_i} \quad (2)$$

Where O_i is number of data of category i, N is total number of data in the dataset, E_i and Np_i is the expected count of category i, and n the number of feature in the dataset. It can determine whether there is a relationship between the target and the feature with the p value obtained as a result of the test with this method.

2.2.3. Decision Tree (DT)

Tree-based methods have high accuracy, adaptability and ease of interpretation. Decision trees are tree-based supervised learning methods. Decision trees are an easy to interpret machine learning approach commonly used for classification and problems. It is one of the most widely used classification techniques due to its low cost, ease of understanding, reliability, and high performance working with different data (Karatas et al., 2020; Bayazit et al., 2020). A decision tree is a method in which the learned function is represented as a decision tree. Its working principle is based on decision making with Entropy calculation. It creates an associated decision tree and keeps the result in leaf nodes when dividing a dataset into smaller subsets with entropy. Figure 3 shows a decision tree structure.

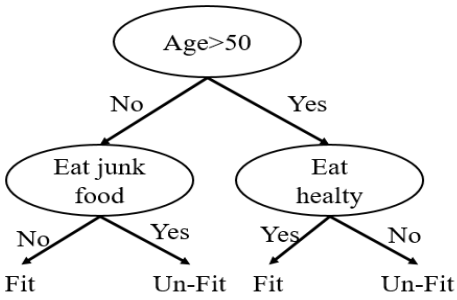

Figure 3. Decision tree structure

2.2.4. Dataset

The most important thing when designing an intrusion detection system is selection of the dataset. Today, although the number of data increases rapidly, it is very difficult to obtain meaningful data. The most important thing in intrusion detection system development for IoT devices is to develop a model by selecting the right dataset. In this direction, the n_Balot dataset was selected and the training and testing processes were carried out with this dataset. This dataset was created by researchers using 9 different IoT devices in 2018. The dataset contains 7,062,606 data and 115 features. 555,932 of them are benign and the rest are attacks. In the data set, operations can be performed for estimations, such as what type of attack, from which devices, apart from normal/attack estimation (Meidan et al., 2018). Labeling possibilities that can be used for classification/clustering in the dataset are given in Table 1.

Table 1. n_Balot dataset labelling possibilities

Benign/ Attack status	**Benign/ Attack Type**	**Benign/ Attak Name**	**Device Model**
Benign	Benign	benign	Donmini
Attack	Gafgty	combo	Ennio
	Mirai	junk	Ecobee
		scan	Philips
		TCP	PT-737E
		UDP	PT-838
		ack	XCS7-1002
		syn	XCS7-1003
		udpplain	Samsung

Data Preprocessing; It is very important to make sense of the dataset to be used in designing an intrusion detection model

and to make it usable. Before training with the dataset used here, the following steps were performed;

1. The dataset is allocated as 9 different folders and 2 different folders within these folders. There are csv files with different normal and attack information in the folders inside. These files have been merged.
2. Each feature has been taken into account when merging files. According to this;
    a. A new column has been added to the incoming csv file and named "*state*". Accordingly, if the name of the incoming file is "*benign.csv*", 0 otherwise 1 label is written.
    b. A new column named "*dirs*." has been added to the incoming csv file. According to the column; If the folder where the file comes from does not have a name, 0 is written, if its name is "*gafgyt attack*", 1 is written, and if its name is "*mirai attack*", 2 is written.
    c. A new column named "*label*" has been added to the incoming csv file. According to the column; Table 2 shows label name and number assigned to it. This column will be used for attack detection. So this is the label of the dataset because main purpose of this work is design a model that can detect different types of attacks and this requires multiple classes.

Table 2. Label names and Numbers

Label	number
Benign	0
Combo	1
Junk	2
Scan	3
Tcp	4
Udp	5
Ack	6
Syn	7
udpplain	8

    d. A new column named "*root*" has been added to the incoming csv file. According to the column; Table 3 shows root name and number assigned to it.

Table 3. Root Names and Numbers

Root file	number
Donmini	0
Ennio	1
Ecobee	2
Philips B120N/10	3
Provision PT-737E	4
Provision PT-838	5
Samsung SNH 1011 N	6
SimpleHome XCS7-1002-WHT	7
SimpleHome XCS7-1003-WHT	8

3. Finally, a new column named "*_neg*" has been added for the negative columns in the dataset, and if the value in the original column is positive, 0 if the value is negative, 1 values are written in the "*_neg*" column. After the operation, the relevant column is deleted from the dataset.

After the data preprocessing was completed, the number of features in the dataset increased from 115 to 119.

2.2.5. Under Sampling (US)

The performance of the developed model decreases considerably when the distribution of the data in the dataset is not in a balanced amount (Yen and Lee, 2009). Operations to reduce or increase (under sampling or over sampling) the number of data are applied to cope with this situation. In this study, under sampling was performed both to speed up the model and to obtain a more successful model. For this process, the RandomOverSampler method in Python has been used, the RandomOverSampler method randomly selects the specified number of data from the desired classes and creates a subset.

In the study, under sampling process was applied to all classes. It is suggested that the sample size should be adjusted with the average or the least data according to the data number of the class when the literature is examined. Therefore, the number of data has been adjusted 255,111 according to the label number 3 that is the "scan" attack, which has the least data in the dataset. Table 4 shows the number of data for classes before and after under sampling.

Table 4. Number of data in the n_Balot dataset

Label	Before	After
Benign - 0	555,932	255,111
Combo - 1	515,516	255,111
Junk - 2	261,789	255,111
Scan - 3	255,111	255,111
Tcp - 4	859,850	255,111
Udp - 5	946,366	255,111
Ack - 6	643,821	255,111
Syn - 7	773,299	255,111
Udpplain - 8	523,932	255,111
Total:	7,062,606	2,295,999

3. Results and Discussion

Experimental results were applied using the Scikit-Learn library with Python programming language on PyCharm Compiler. The study was carried out on a computer indicated by the Table 5. In the study, the dataset is separated for training and testing. Accordingly, 75% of the dataset was used for training and 25% for testing.

The main purpose of the study is to create a fast-running hybrid model that will perform multiple classification operations. For this, operations were performed according to the "label" column in the dataset. Information about this column is provided under the Proposed Approach. In addition, the "state" column was not used in clustering and classification processes, so that the algorithm does not overfit. The other aim of the study is to investigate the importance of data preprocessing and combining different types of algorithms in intrusion detection on

IoT devices. Therefore, under sampling and feature selection process was performed on the dataset, and then clustering was done with the K-means approach.

Table 5. Working Environment

Hardware	Features
CPU	Intel(R) Core(Tm) I7-8700 Cpu @3192Mhz, 6 Cores
Op. Sys.	64 bit, Windows 10
Graphic card	NVIDIA GeForce® GTX 1080 Ti Founders Edition 11G
L1/L2/L3 Cache	384 KB/1.5 MB/12.0 MB
RAM	16.00 GB

According to the points reached as a result of clustering, Decision Tree algorithm was used and accuracy rate was tried to be increased. Decision Tree algorithm was applied without making any changes on the numerically converted dataset. Table 6 shows the results.

Table 6. Only DT

Accuracy (%)	Time (sec)
99.61	217.08

Then, the under sampling was performed on the dataset and the results after working with decision tree are shown in Table 7.

Table 7. US+DT

Accuracy (%)	Time (sec)
99.73	96.13

It has been observed that there is a slight increase in the performance rate, but the running time of the algorithm has decreased considerably. Then, the K-means algorithm was applied to the under sampling applied dataset and clustering was done. The results obtained by clustering were classified with Decision Tree algorithm and the accuracy rate was shown in Table 8.

Table 8. US + K-Means + DT

Accuracy (%)	Time (sec)
99.99	56.90

Finally, the sampling reduction followed by the feature selection process was performed on the dataset, and then the K-means algorithm was applied. The number of features was selected as 100, 50, 30, 25, 15, 5 and 4 in the preliminary studies and it was seen that the accuracy rate did not change up to 5 features. Therefore, 5 important features were selected with the Chi-Square method in the dataset and the operations were carried out accordingly. Then again, Decision Tree algorithm was applied and the results were shown in Table 9.

Table 9. US + FS + K-Means + DT

Accuracy (%)	Time (sec)
99.99	2.90

It was observed that the performance did not change after the feature selection, but the detection time was greatly reduced. Table 10 shows all states, accuracy rates, times and error rates.

Table 10. All results

Algorithm	Accuracy (%)	Error Rate (%)	Time (sec)
Only DT	99.61	0.39	217.08
US + DT	99.73	0.27	96.16
US + K-means + DT	99.99	0.01	56.90
US + FS + K-means + DT	99.99	0.01	2.90

The decrease in the error rate of the result achieved by the use of hybrid algorithm in attack detection is shown in Figure 4.

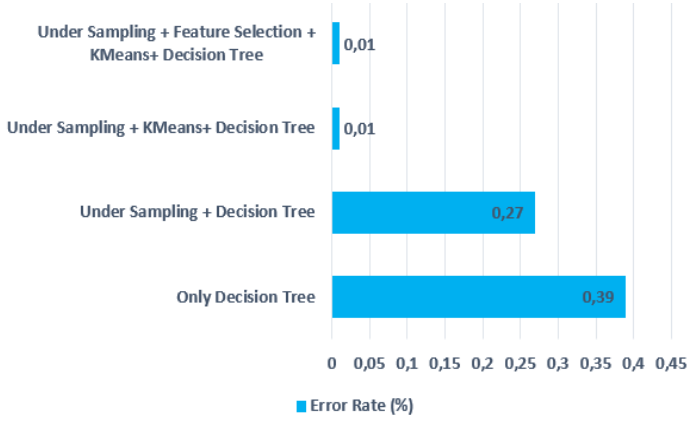

Figure 4. All results in one graphic.

4. Conclusions and Recommendations

Technology has the ability to change the world and the future. IoT devices, which have gained popularity in recent years, allow to facilitate and improve the life of humanity; Using IoT services, people can access and use the information they want from anywhere. Although IoT devices make life easier, security becomes a big problem in the IoT system regarding the protection of user information. It has become a current research to develop new solutions to uncertain attacks and deal with new models to improve security.

The aim of the study is to detect attacks on IoT devices using Decision Tree algorithm, which is used in many subjects and popular machine learning algorithms. For this, it is aimed to increase the performance of the algorithm by using oversampling, feature selection and K-means method. In this way, a hybrid algorithm has emerged. In the study, under sampling and feature selection processes were also carried out, since it was desired to observe how data preprocessing affects the running time of the model. After applying numerical transformations to the n_Balot dataset used in the study, under sampling was applied and all classes were equalized to the number corresponding to the lowest class data number. Then, heuristic evaluations were made with Chi-Square feature selection and it was seen that the system gave the same results with 5 features. After the preprocessing was completed, the model was created, first clustering with K-means was performed, and then a classification was made with the clustering results. And as a result of the study, the error rate of the Decision Tree algorithm was reduced from 0.39% to 0.01% with all these operations. The effects of normalization and

standardization on deep learning algorithms will be examined in future studies according to the values reached as a result of this study.

References

Alsaedi, A., Moustafa, N., Tari, Z., Mahmood, A., & Anwar, A. (2020). TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems. *IEEE Access*, *8*, 165130-165150.

Anthi, E., Williams, L., Słowińska, M., Theodorakopoulos, G., & Burnap, P. (2019). A supervised intrusion detection system for smart home IoT devices. *IEEE Internet of Things Journal*, *6*(5), 9042-9053.

Ashton, K. (2009). That 'internet of things' thing. *RFID journal*, *22*(7), 97-114.

Bayazit, E. C., Sahingoz, O. K., & Dogan, B. (2020, June). Malware detection in Android systems with traditional machine learning models: a survey. In *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-8). IEEE.

Derhab, A., Aldweesh, A., Emam, A. Z., & Khan, F. A. (2020). Intrusion detection system for Internet of Things based on temporal convolution neural network and efficient feature engineering. *Wireless Communications and Mobile Computing*, *2020*.

Foley, J., Moradpoor, N., & Ochenyi, H. (2020). Employing a Machine Learning Approach to Detect Combined Internet of Things Attacks against Two Objective Functions Using a Novel Dataset. *Security and Communication Networks*, *2020*.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100-108.

Hodo, E., Bellekens, X., Hamilton, A., Dubouilh, P. L., Iorkyase, E., Tachtatzis, C., & Atkinson, R. (2016, May). Threat analysis of IoT networks using artificial neural network intrusion detection system. In *2016 International Symposium on Networks, Computers and Communications (ISNCC)* (pp. 1-6). IEEE.

Karatas, G., Demir, O., & Sahingoz, O. K. (2020). Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset. *IEEE Access*, *8*, 32150-32162.

Korkmaz, M., Sahingoz, O. K., & Diri, B. (2020, June). Feature Selections for the Classification of Webpages to Detect Phishing Attacks: A Survey. In *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-9). IEEE.

Korkmaz, M., Sahingoz, O. K., & Diri, B. (2020, July). Detection of Phishing Websites by Using Machine Learning-Based URL Analysis. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461.

Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Shabtai, A., Breitenbacher, D., & Elovici, Y. (2018). N-baiot—network-based detection of iot botnet attacks using deep autoencoders. *IEEE Pervasive Computing*, *17*(3), 12-22.

Raza, S., Wallgren, L., & Voigt, T. (2013). SVELTE: Real-time intrusion detection in the Internet of Things. *Ad hoc networks*, *11*(8), 2661-2674.

Parra, G. D. L. T., Rad, P., Choo, K. K. R., & Beebe, N. (2020). Detecting Internet of Things attacks using distributed deep learning. *Journal of Network and Computer Applications*, *163*, 102662.

Sforzin, A., Mármol, F. G., Conti, M., & Bohli, J. M. (2016, July). Rpids: Raspberry pi ids—a fruitful intrusion detection system for iot. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)* (pp. 440-448). IEEE.

Smys, S., Basar, A., & Wang, H. (2020). Hybrid intrusion detection system for internet of Things (IoT). *Journal of ISMAC*, *2*(04), 190-199.

Tallarida, R. J., & Murray, R. B. (1987). Chi-Square Test. Manual of Pharmacologic Calculations.

Yang, K., Ren, J., Zhu, Y., & Zhang, W. (2018). Active learning for wireless IoT intrusion detection. *IEEE Wireless Communications*, *25*(6), 19-25.

Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, *36*(3), 5718-5727.