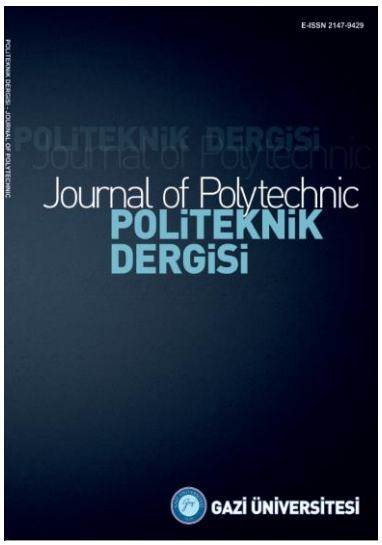

POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Machine learning based network intrusion detection with hybrid frequent item set mining

Hibrit sık kullanılan öğe kümeleme ile makine öğrenmesi tabanlı ağ sızma tespiti

***Yazarlar (Authors)**: Murat FIRAT[1], Gokhan BAKAL[2], Ayhan AKBAS[3]*

ORCID[1]: 0009-0009-0113-9868

ORCID[2]: 0000-0003-2897-3894

ORCID[3]: 0000-0002-6425-104X

To cite to this article: Firat M., Bakal G. ve Akbas A., "Machine Learning based Network Intrusion Detection with Hybrid Frequent Item Set Mining", ***Journal of Polytechnic***, 27(5): 1937-1943, (2024).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Firat M., Bakal G. ve Akbas A., "Machine Learning based Network Intrusion Detection with Hybrid Frequent Item Set Mining", ***Politeknik Dergisi***, 27(5): 1937-1943, (2024).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1386467

Machine Learning based Network Intrusion Detection with Hybrid Frequent Item Set Mining

Highlights

- *A novel hybrid feature selection approach is proposed using frequent item set mining.*
- *Distinct machine learning models are constructed to show the proposed approach performance.*
- *A recent intrusion detection dataset is used to test the proposed hybrid approach.*

Graphical Abstract

A hybrid feature selection method is integrated with several classification algorithms for the detection of malicious network traffic.

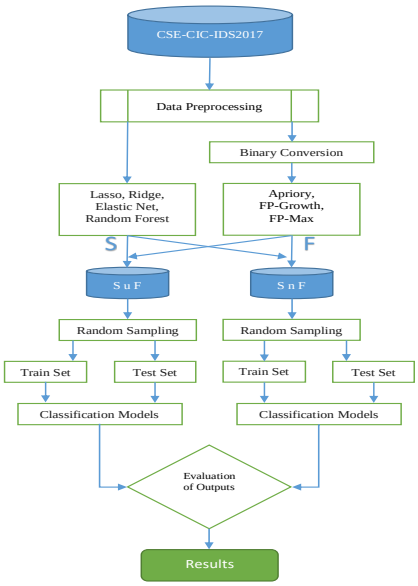

Figure. Proposed method

Aim

This study aims to propose a hybrid feature selection approach that enhances the classification performance of Intrusion Detection Systems (STS/IDS) equipped with artificial intelligence methods.

Design & Methodology

Classical feature selection methods and frequent item set mining approaches are employed to construct a hybrid model, specifically focusing on network traffic data encompassing both ordinary and attack records.

Originality

In this study proposes a novel hybrid feature selection method, integrating classical techniques and frequent item set mining is proposed. The originality lies in the application of this hybrid approach to improve the classification performance of Intrusion Detection Systems.

Findings

Outcomes reveal a notable 3% improvement in classification performance when applied with the Logistic Regression algorithm on a dataset comprising over 225,000 records.

Conclusion

The proposed approach optimizes the performance of IDS systems, leveraging classical methods and frequent item set mining and yields promising results in improving the classification accuracy for network traffic data.

Declaration of Ethical Standards

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Machine Learning Based Network Intrusion Detection with Hybrid Frequent Item Set Mining

Araştırma Makalesi / Research Article

Murat FIRAT[1], Gökhan BAKAL[2], Ayhan AKBAŞ[3*]

[1]Computer Engineering Department, Cankiri Karatekin University, Cankiri, Türkiye
[2]Computer Engineering Department, Abdullah Gul University, Türkiye
[3]Institute for Communication Systems, University of Surrey, Guildford, UK

(Geliş/Received : 06.11.2023; Kabul/Accepted : 25.12.2023; Erken Görünüm/Early View : 18.01.2024)

ABSTRACT

With the development and expansion of computer networks day by day and the diversity of software developed, the damage that possible attacks can cause is increasing beyond the predictions. Intrusion Detection Systems (STS/IDS) are one of the practical defense tools against these potential attacks that are constantly growing and diversifying. Thus, one of the emerging methods among researchers is to train these systems with various artificial intelligence methods to detect subsequent attacks in real time and take the necessary precautions. However, the ultimate goal is to propose a hybrid feature selection approach to improve the classification performance. The raw dataset originally enclosed 85 descriptor features (attributes) for classification. These attributes are extracted using CICFlowMeter from a PCAP file where network traffic is recorded for data curation. In this study, classical feature selection methods and frequent item set mining approaches were employed in feature selection for constructing a hybrid model. We aimed to examine the effect of the proposed hybrid feature selection approach on the classification task for the network traffic data containing ordinary and attack records. The outcomes demonstrate that the proposed method gained nearly 3% improvement when applied with the Logistic Regression algorithm on classifying more than 225,000 records.

Keywords: Intrusion detection systems, frequent ıtem set mining, hybrid feature selection, machine learning.

Hibrit Sık Kullanılan Öğe Kümeleme ile Makine Öğrenmesi Tabanlı Ağ Sızma Tespiti

ÖZ

Bilgisayar ağlarının gün geçtikçe gelişmesi ve genişlemesi ve geliştirilen yazılımların çeşitliliği ile olası saldırıların neden olabileceği zararlar tahminlerin de ötesine geçmektedir. Sızma Tespit Sistemleri (STS/IDS), sürekli büyüyen ve çeşitlenen bu potansiyel saldırılara karşı pratik savunma araçlarından biridir. Bu nedenle, araştırmacılar arasında ortaya çıkan metotlardan biri, bu sistemleri çeşitli yapay zeka yöntemleri ile eğiterek gerçek zamanlı olarak sonraki saldırıları tespit etmelerini ve gerekli önlemleri almalarını sağlamaktır. Ancak, asıl hedef, sınıflandırma performansını iyileştirmek için hibrit bir özellik seçimi yaklaşımı önermektir. Ham veri seti başlangıçta sınıflandırma için 85 tanımlayıcı özellik içermekteydi. Bu nitelikler, veri kürasyonu için ağ trafiğinin kaydedildiği bir PCAP dosyasından CICFlowMeter kullanılarak çıkarılmıştır. Bu çalışmada, hibrit bir model oluşturmak için klasik özellik seçimi yöntemleri ve sık öğe kümesi madenciliği yaklaşımları özellik seçiminde kullanılmıştır. Önerilen hibrit özellik seçimi yaklaşımının, sıradan ve saldırı kayıtlarını içeren ağ trafiği verileri için sınıflandırma görevine etkisini incelemeyi amaçladık. Sonuçlar, önerilen yöntemin, 225.000'den fazla kaydı sınıflandırmada Lojistik Regresyon algoritması ile uygulandığında yaklaşık %3'lük bir iyileşme sağladığını göstermektedir.

Anahtar Kelimeler: Sızma tespit sistemleri, sık kullanılan öğe kümesi madenciliği, hibrit özellik seçimi, makine öğrenmesi.

1. INTRODUCTION

The information technology field is experiencing a rapid proliferation of security threats in terms of both quantity and variety, leading to notable advancements in security technologies. Hardware and software tools, including antivirus programs, Intrusion Detection Systems (IDS), and firewalls, have been specifically developed to address these challenges.

System administrators employ one or more of these tools in conjunction to safeguard their systems against potential attackers. However, relying solely on a single security measure is insufficient to ensure comprehensive protection. Consequently, Intrusion Detection Systems have become indispensable for system administrators. These systems are designed to identify and respond to potential attacks originating from local networks or the internet, preferably before they occur, but also subsequently if necessary.

Presently, researchers are actively engaged in the pursuit of developing more powerful, faster, and higher success-rate intrusion detection systems. Although there are numerous intrusion detection systems documented in the existing literature that utilize artificial intelligence and its subfields, our proposed study distinguishes itself through the novel integration of classical feature selection methods and frequent item set mining, which is

*Corresponding Author
e-mail : a.akbas@surrey.ac.uk

combined with machine learning techniques within a hybrid structured model. Through our novel approach, we aim to introduce an original method for intrusion detection.

Contribution. Research on Intrusion Detection Systems (IDS) has been developed over the years to propose more successful IDS. However, many researchers have faced challenges in finding up-to-date and sufficient datasets to test and evaluate their studies. Consequently, studies have often been evaluated using the same few datasets. Due to the continuous evolution of malicious software and changing attack strategies, studies conducted on these reference datasets may fall short in addressing current problems. In our proposed study, the CIC-IDS2017 dataset, which is one of the most up-to-date publicly available datasets, was used to target the accurate detection of recent attacks. The uniqueness of our study lies in the hybrid approach of combining classical feature selection methods with frequent itemset mining approaches for feature selection, accompanied by the utilization of machine learning techniques to observe any improvements in the classification performances.

The structure of the paper is as follows: first section provides general information on the area. In Section II, related studies are given and relevant research conducted in the field are summarized. In the first part of Section III, the details about the used dataset is presented, including an explanation of the attack types present in the target (label) column. The following part in Section III discusses the preprocessing steps taken to address any missing or erroneous parts of the dataset, making it ready for application. Lastly, in Section III, we focus on feature selection studies, describing the methods used in detail and explaining the machine learning models employed and the continued methodology. In Section IV, the findings of the study are presented using numerical values. Section V concludes the study by summarizing the obtained results.

2. RELATED WORK

There are numerous studies in the literature on intrusion detection [1]–[3]. Altunay and Albayrak [4] utilized the CIC IDS2018 dataset which was created by the Canadian Cybersecurity Institute (CSE) and other synthetically generated datasets. They employed a convolutional neural network-based intrusion detection system for feature selection on these datasets. The results showed an achievement rate of 98.32% on the CIC IDS2018 dataset and a rate of 98.8% on the CIC IDS2018+Synthetic dataset.

In another study by Karaman et al. [5], where they implemented the Artificial Neural Networks approach on the CIC IDS2018 dataset. The study yielded impressive results, with threat detection accuracy reaching 99.11%, botnet attack detection accuracy at 93.23%, DDoS attack detection accuracy at 99.31%, DoS attack detection accuracy at 92.26%, and BruteForce attack detection accuracy at 99.26%.

An important research was done by Bakhshi and Ghita [6] who employed deep learning methods (CNN, LSTM, RNN, CNN+GRU) on the NSL-KDD, UNSW-NB15, and CIC-IDS-2017 datasets. The results of the study demonstrated significant achievements, with a performance of 93.56% (using CNN) on the NSL-KDD dataset, 91.21% (using CNN+GRU) on the NB15 dataset, and 90.17% (using CNN+GRU) on the CICIDS-2017 dataset.

In a study carried out by Wei et al. [7], they tested the methods of Automatic Encoder and Independent RNN with One-Dimensional CNN (1DCAE-IndRNN) on the CIC And Mal 2017 dataset, achieving a detection success rate of 98%. Arslan [8] performed a research study in which preprocessing data and machine learning techniques were applied to the CSE CIC-IDS2018 dataset, reaching a performance of 99.5% using the ExtraTree method and 98.5% using the Random Forest method.

In a study [9], Atay et al. preferred to make use of various methods such as LGBM, CNN, LGBM+Random Forest, CNN+Random Forest, and Random Forest+Random Forest on the CSE CIC-IDS2018 dataset. The highest performance of 98% was achieved using the CNN+Random Forest method. On the CIC IDS2017 dataset, Özekes and Karakoç, [10] utilized the Decision Tree and Random Forest methods, obtaining a performance of 99.95% for the Decision Tree method and 99.966% for the Random Forest method. Tokyürek [11], in a study he conducted, employed the Apriori and FP-Growth algorithms on market data using the Weka program. It was found that with a minimum support value of 0.35, the Apriori algorithm took three minutes to generate rules, while the FP-Growth algorithm took nearly three hours for rule generation.

Hidayanto et al. [12] used the Apriori and FP-Max methods on a specific dataset (Id-SIRTII/CC-Indonesia). As a result of the application, SQL attacks, Malware Virus DNS, and DoS attacks were detected with a minimum support value of 95% using both algorithms. Out of the 620 discovered rules, eight rules dominated 90% of them.

Using a different approach, Awadh and Akbas [1] proposed the TF.IDF and C4.5 methods on the UNSW-NB15 dataset in their research. The highest accuracy of 99.43% was achieved on a binary class dataset with a segment size of 1000.

In a recent research [13], Moustafa and Slay developed a hybrid model incorporating Feature Selection and Association Rule Mining methods on the UNSW-NB15 and NSLKDD datasets. The proposed model increased accuracy, reduced false alarm rates, and shortened processing time. Likewise, Aung and Oo [14] employed a modified FP-Growth algorithm on the KDD dataset, demonstrating the model's efficiency in identifying infrequent items and its ability to process large amounts of data.

Nalavade and Meshram [15] used modified association rules (using the signed Apriori algorithm) on the KDD99

Cup dataset, generating 11 frequent item sets and 35 rules. They successfully detected attacks with a low false positive rate. Sokhangoee and Rezapour [16], studied association rules and applied association rules and a combination of genetic algorithms for feature selection and classifier methods on The SAC'13, The Spambase, and The ICC datasets. They achieved an average accuracy rate of 87.99% and 95.24% for the two approaches, respectively, compared to the baseline methods. In May 2018, Cekmez et al. [17] conducted a study in which they applied automatic feature extraction and automatic encoders on the NSL-KDD dataset. They achieved a performance of 91.3% on the KDDTest+ dataset and 85.5% on the KDDTest21 dataset.

3. MATERIAL AND METHOD

A. Dataset

In this study, we utilized the publicly available CICIDS2017 dataset, created by the Canadian Institute for Cybersecurity [18] to allow researchers interested in the field. This dataset was constructed taking into account the limitations of previously established datasets by examining attack structures and normal traffic flows during a test period of 5 days.

Table 1. Some of the features and their descriptions

ID	Feature	Description
0	Source Port	The sending device's port
1	Destination Port	The receiving device's port
2	Protocol	Protocol used
3	Flow Duration	Duration of flow in microseconds
4	Total Fwd Packets	Total number of packets in forward direction
5	Total Backward Packets	Total number of packets in backward direction
.	.	.
.	.	.
.	.	.
66	Idle Mean	Average active duration of flow before going idle
67	Idle Std	Std. deviation of active duration of flow before going idle
68	Idle Max	Maximum active duration before going idle
69	Idle Min	Minimum active duration before going Idle
70	Attack	Attack label

The dataset was generated in the PCAP (Packet Capture Data) file format and subsequently converted to the CSV (Comma-Separated Values) format using the CICFlowMeter network traffic flow generator application. The dataset consists of 225,745 samples and 84 traffic features, with an additional label feature added, resulting in 85 columns. As part of the preprocessing conducted for the study, the attribute columns, initially 85, were reduced to 71, and the number of samples, initially 225,745, was reduced to 225,711. Table I provides descriptions of some attributes of the obtained dataset. The last column, originally labeled as "Label" in the dataset, contains the values BENIGN (Normal) and DDoS (Distributed Denial of Service). In the study, it was referred to as "Attack," and its cell values were transformed to 0 and 1. Distributed Denial of Service (DDoS) attacks refer to a type of attack where an excessive amount of requests is sent to a targeted server through previously compromised devices such as computers, mobile phones, tablets, smart devices, etc., with the aim of rendering the server unable to provide services. These attacks operate with the objective of pushing the capacity limits of a network server (e.g., a web server). DDoS attacks aim to disrupt the normal functioning of the server by overwhelming it with a high volume of requests from numerous clients, surpassing its capacity to respond to such an excessive number of requests. Servers have a limited capacity to handle a certain number of requests within a given time frame. Additionally, the channel connecting the network resources to the internet has a restricted bandwidth. When the quantity of requests received from clients surpasses the capacity limits of any component within this infrastructure, service disruptions occur. As a result of these disruptions, the responses to client requests are significantly delayed, and in some cases, certain requests may go unanswered or be entirely ignored. To be able to send an excessive number of requests to the target, the attacker establishes a "zombie army" using previously infected personal computers, mobile phones, tablets, smart devices, and other similar devices. The size of this zombie army and its capability to generate requests determine the magnitude of the attack.

B. Data Preprocessing

In the utilized dataset, a total of 68 cells containing missing or infinite values were identified, and the corresponding rows were removed considering the size of the dataset. Subsequently, a total of 10 attributes were excluded from the dataset as all column values were determined to be zero. The name of the last column containing the label values was transformed to "Attack," and it was converted to binary values, with 0 representing "Benign" (Normal) and 1 representing "DDoS" (Attack). Additionally, the dataset was duplicated to apply frequent itemset mining algorithms, and all cells in the duplicated dataset were transformed into binary values of 0 and 1. During this transformation, the average value of the column where the respective attribute was located was taken as the threshold value. For cell values below the threshold, a value of 0 was assigned, while values equal to or above the threshold were assigned a value of 1, thereby creating a new dataset.

C. Feature Selection

Feature selection is defined as the process of selecting the best subset that can represent the original dataset [19]. In this study, feature selection was performed by considering that not all features in the dataset are necessary, and the most influential features were identified for prediction purposes. To determine these features, a hybrid method was employed, combining classical methods such as Lasso Regression, Ridge Regression, ElasticNet Regression, and Random Forest Classifier, with frequent itemset mining algorithms including Apriori, FP-Growth, and FP-Max. Firstly, a classical feature selection was conducted using Lasso Regression, Ridge Regression, Elastic Net Regression, and Random Forest Classifier methods (Figure. 1).

The **Lasso Regression** (Least Absolute Shrinkage and Selection Operator) utilizes the absolute value of coefficients (L1 penalty- L1 regularization), leading to the complete neglect of certain features. It plays a significant role in feature selection and is also employed to reduce overfitting.

On the other hand, **Ridge Regression** employs the squares of weights (L2 penalty - L2 regularization). This regression method aims to find the parameter pair with the smallest variance (average of the squared differences from the mean). It tends to shrink all weights towards zero to an equal extent (Figure. 2).

	Lasso Regression		Ridge Regression	
1	Bwd Packet Length Std	30.923053	Bwd Packet Length Std	37.943879
2	ACK Flag Count	29.761833	Total Length of Bwd Packets	25.910856
3	Packet Length Std	16.328752	Packet Length Std	21.162387
4	Flow IAT Mean	3.375144	Total Fwd Packets	19.355160
5	Packet Length Variance	1.207272	Average Packet Size	16.330339
6	Total Fwd Packets	0.476519	Subflow Bwd Packets	13.383797
7	act_data_pkt_fwd	0.262608	Bwd Packets/s	11.492405
8	Avg Fwd Segment Size	-0.000000	Flow IAT Mean	9.624607
9	Average Packet Size	0.000000	Fwd IAT Std	6.466631
10	Down/Up Ratio	-0.000000	Bwd Header Length	4.705843
11	ECE Flag Count	-0.000000	Active Std	3.926928
12	Source Port	-0.000000	Flow IAT Std	3.634790
13	PSH Flag Count	-0.000000	Bwd IAT Total	3.092734
14	Avg Bwd Segment Size	0.000000	Fwd Packet Length Max	2.942838
15	SYN Flag Count	-0.000000	Subflow Fwd Bytes	2.901544
16	FIN Flag Count	-0.000000	Total Length of Fwd Packets	1.903692
17	Packet Length Mean	0.000000	Packet Length Variance	1.532421
18	Max Packet Length	0.000000	ACK Flag Count	1.457024
19	RST Flag Count	-0.000000	Avg Fwd Segment Size	1.422686
20	Subflow Fwd Packets	0.000000	Fwd Packet Length Mean	1.422686

	Elastic Net Regression		Random Forest Classifier	
1	Bwd Packet Length Std	27.365466	Destination Port	8.661165
2	ACK Flag Count	22.028426	Bwd Packet Length Std	6.697136
3	Packet Length Std	15.220442	Bwd Packet Length Max	6.571627
4	Flow IAT Mean	8.359915	ACK Flag Count	5.448238
5	Total Fwd Packets	6.468059	Bwd Packet Length Mean	5.388956
6	Average Packet Size	3.206071	Total Length of Bwd Packets	4.051165
7	Packet Length Variance	2.727069	URG Flag Count	3.833347
8	Max Packet Length	1.970210	Bwd Packet Length Min	3.772755
9	act_data_pkt_fwd	1.032004	Avg Bwd Segment Size	3.630978
10	Packet Length Mean	0.959765	Init_Win_bytes_forward	3.626812
11	Bwd Packet Length Mean	0.905055	Fwd Packet Length Min	3.614074
12	Avg Bwd Segment Size	0.868207	Min Packet Length	3.092483
13	Down/Up Ratio	-0.000000	min_seg_size_forward	2.681445
14	Source Port	-0.000000	Flow Packets/s	2.444384
15	ECE Flag Count	-0.000000	Subflow Bwd Bytes	2.233641
16	PSH Flag Count	-0.000000	Flow IAT Mean	2.081051
17	SYN Flag Count	-0.000000	Total Backward Packets	1.871282
18	FIN Flag Count	-0.000000	Subflow Bwd Packets	1.796236
19	RST Flag Count	-0.000000	Max Packet Length	1.748291
20	Subflow Fwd Packets	0.000000	Bwd IAT Mean	1.674695

Figure 1. First 20 features with respect to their weights

Elastic Net regression combines the L1 and L2 approaches, meaning that the coefficients of this linear regression model are trained with both L1 and L2. When performing calculations in Elastic Net regression, it utilizes a merged structure of Lasso regression and Ridge regression predictors.

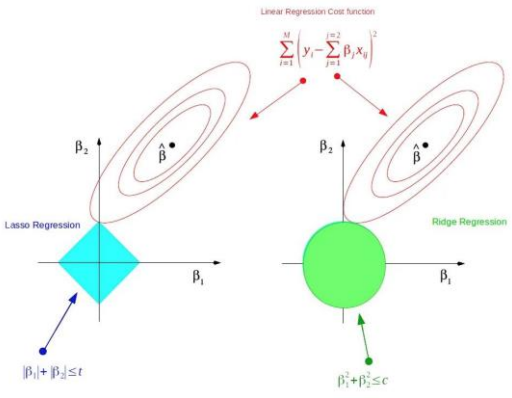

Figure 2. Illustration of Lasso and Ridge Regression

Random Forest classifier is a model that uses multiple decision trees to generate ideal models and aims to achieve more accurate classification [20]. This model attempts to overcome the overfitting problem of Decision Tree algorithms by selecting a large number of subsets from the dataset and training them.

In the study, a hybrid modeling approach was employed by using the intersection and combination of features obtained through four classical feature selection methods and frequent itemset mining techniques.

Frequent Itemset Mining consists of two steps: identifying frequently occurring items in a dataset and extracting strong association rules based on these items. Association rules are used to identify relationships and correlations among datasets. These rules reveal frequently observed feature value patterns within the working dataset.

One of the most classical and fundamental algorithms used for association rule inference is the **Apriori algorithm**. The Apriori algorithm derives its name from the fact that it gains knowledge from the previous step, hence the term "prior." It operates in a forward-repeating manner and is a classic algorithm used to discover frequently occurring itemsets in databases. According to the working logic of the Apriori algorithm, if an itemset with k elements (k-itemset) satisfies the minimum support constraint, so will its subsets, too.

The **FP-Growth algorithm** (Frequent Pattern Growth) is an efficient algorithm for generating fast rules in large data sets. By storing the database in a frequent pattern tree structure, it is fast and low-cost. Unlike the Apriori algorithm, the FP-Growth algorithm does not repeatedly scan the database but only scans it twice, regardless of its size, resulting in much faster outcomes.

On the other hand, the **FP-Max algorithm** is developed to investigate and discover only the maximum frequent itemsets in a database. It is based on the FP-Growth algorithm. In the study, implementations were conducted using the Apriori, FPGrowth, and FP-Max algorithms.

D. Modelling

In the study, various models were tested, and consistent results were not obtained with methods other than Logistic Regression (Eq. 1). Therefore, this method has been primarily considered.

$$f(x) = \frac{1}{1+e^{-x}} \quad (1)$$

Logistic Regression (LR), despite its name containing "regression," is a classification model and is one of the fundamental models among classification models [21], [22]. It utilizes the Maximum Likelihood method to determine the line that distinguishes between two classes. Logistic Regression employs the Sigmoid (Logistic) function for classification (Fig. 3).

This model calculates the probability of an observation belonging to one of the two categories (e.g., present-absent, positive-negative, yes-no, 1-0) based on the

training data. In this method, the target variable is typically categorical, often binary.

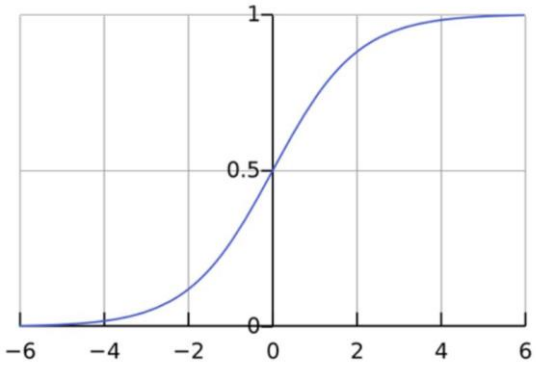

Figure 3. The sigmoid function curve

The LR model has several advantages over other classical Machine Learning techniques. Some of these advantages are as follows:

Simplicity: From a mathematical perspective, these models are less complex compared to other Machine Learning methods. Therefore, they are highly applicable and easy to use.

Speed: Logistic Regression requires less computational power and memory, allowing it to process large volumes of data at high speeds. This makes it ideal for users who need quick results.

Flexibility: Logistic Regression can be used to find answers to questions with two or more limited responses. It can also be used for data preprocessing and its outputs can be processed as a new dataset using other Machine Learning techniques.

Visibility: Logistic Regression provides more visibility in software processes compared to other data analysis methods. As the calculations are less complex, error corrections are also easier.

Since Logistic Regression is a classification algorithm, it cannot predict real values when continuous data is involved. As a result, by using the intersection and combination of features obtained through four classical feature selection methods and frequent itemset mining techniques, final feature spaces were obtained and the classification models were built by using those retained features.

4. RESULTS AND DISCUSSIONS

Lasso Regression, Ridge Regression, Elastic Net Regression, and Random Forest classifier were initially applied to the preprocessed dataset containing 71 columns. The results obtained from these methods were combined using Association Rule Mining techniques like Apriori, Fp-Growth, and FPMax to perform hybrid feature selection, taking the union (FuS-61 columns) and intersection (FnS-14 columns) of the results. Subsequently, various machine learning methods from the Sklearn machine learning library [23], including Logistic Regression, Neural Network (Multi-layer Perceptron - MLP), XGBoost, and Random Forest, were tested for normal and abnormal (attack) network traffic classification.

The original dataset used in the study contains 97,718 labeled samples of normal (BENIGN) traffic and 128,027 labeled samples of attack (DDoS) traffic, totaling 225,745 records. After preprocessing, the data was split into training and testing sets at various ratios, and experiments were conducted. Based on these experiments, it was decided to use 80% of the total data for training and the remaining 20% for testing purposes (Figure. 4).

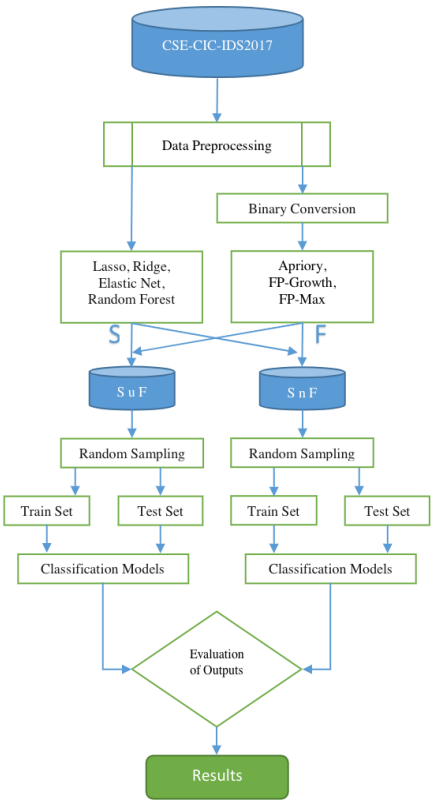

Figure 4. Flowchart of the proposed system

As shown in Table 2, we used the accuracy metric to see how well the classification methods performed on the test dataset. Even though we tried several models with the data, we could not get consistent results, and we did not report them. This study emphasizes the need for clear and accurate reporting by keeping the different classification outcomes from the six methods we studied. We admit that the majority of models achieved superior performances, and the proposed hybrid approach did not yield any better results. However, the main objective was to observe any improvement when we applied the proposed feature selection method. Unlike the other models, the logistic regression model gained a 3%

accuracy improvement when we integrated the hybrid feature selection approach.

Table 2. Performance results of the applied methods on the test set

Methods	All (71 Col)	FuS(61 Col)	FnS (14 Col)
Logistic Regression	0.94772	0.94081	**0.97687**
Naive Bayes	0.80820	0.80876	0.64023
Neural Network (MLP)	0.99692	0.98117	0.98615
Random Forest	0.99988	0.99991	0.99937
XGBoost	0.99991	0.99991	0.99951
CatBoost	0.99988	0.99991	0.99946

Based on the obtained test results, the following evaluations have been made:

- The Neural Network (MLP) method did not show improvement on either of the datasets.
- The XGBoost method did not show any improvements on either of the datasets.
- The Random Forest method showed a quite small amount of improvement as approximately 0.01% on the combined (FuS) dataset.
- The CatBoost method also yielded a negligible improvement of approximately 0.01% on the combined (FuS) dataset.
- The Naive Bayes algorithm achieved a tiny improvement rate of approximately 0.05% on the combined (FuS) dataset. Surprisingly, the model was also negatively affected on the combined (FuS) dataset having a nearly 16% drop in accuracy.
- The Logistic Regression method achieved an accuracy improvement rate of close to 3% on the intersection (FnS) dataset.

5. CONCLUSION

In cybersecurity, Intrusion Detection Systems (IDS) play a critical role in defense systems by enabling early detection of malicious activities such as intrusion, eavesdropping, and attacks, allowing for timely action. Having an effective IDS has become essential for protection against constantly increasing malicious activities. Therefore, ongoing research efforts have been dedicated to developing better IDS.

In this study, a hybrid feature selection method is integrated with several classification algorithms for the detection of malicious network traffic, which is one of the commonly encountered types of cyber attacks. The CIC-IDS2017 dataset, one of the most comprehensive and up-to-date publicly available datasets in the field, was used as the main dataset. The performance of the models trained and tested using over 250,000 samples was evaluated using the accuracy performance metric. The evaluations revealed that the proposed model improved the accuracy value by approximately 3.0%, from the baseline accuracy of 94.77% in the model using the Logistic Regression method. However, the other models utilizing distinct classification algorithms did not show any considerable improvements. Even though not all the models provided promising results, the Logistic Regression, Random Forest, XGBoost, and CatBoost models yielded at least some minor improvements. As a result, an accuracy value of 97.68% was achieved by the LR model employed with the proposed hybrid approach, indicating, to some extent, the potential effectiveness of the proposed method.

With the proposed study, it is possible to detect and take action against Distributed Denial of Service (DDoS) attacks experienced in network environments. Another important aspect of the study is the identification of the types of these attacks. Some examples of different attack types include Malware, Phishing, SQL Injection, DoS/DDoS, Man in The Middle, PortScan, Cryptojacking, SSH Patator, FTP Patator, Zero Day Exploit, Web Attack, Web Attack XSS, Web Attack BruteForce, Bot, Passwords Attack, Heartbleed, Eavesdropping Attack, Birthday Attack, and Infiltration methods. In future studies building upon the proposed research, the aim is to categorize network traffic based on the types of malicious activities it represents.

DECLARATION OF ETHICAL STANDARDS

The authors of this article declare that the materials and methods employed in their research do not require ethical committee approval and/or legal-special permission.

AUTHORS' CONTRIBUTIONS

Murat FIRAT: Performed the experiments, analyzed the results and wrote the initial draft of the manuscript.

Gökhan BAKAL: Co-supervised the study, analyzed the results, and wrote the final manuscript.

Ayhan AKBAŞ: Co-supervised the study, analyzed the results, and wrote the final manuscript.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] Awadh K. and Akbas A., "Intrusion detection model based on TF.IDF and C4.5 algorithms", ***Politeknik Dergisi***, 24:(4), 1691–1698, (2021).

[2] Akbas A. and Buyrukoglu S., "Deep belief network based wireless sensor network connectivity analysis," ***Balkan Journal of Electrical and Computer Engineering***, 11: 262–266, (2023).

[3] Uyan O. G., Akbas A., and Gungor V. C., "Machine learning approaches for underwater sensor network

parameter prediction," ***Ad Hoc Networks***, 144:103-139, (2023).

[4] Altunay H. C. and Albayrak Z., "Network intrusion detection approach based on convolutional neural network," ***Avrupa Bilim ve Teknoloji Dergisi***, 26: 22–29, (2021).

[5] Karaman M. S., Turan M., and Aydin M. A., "Yapay Sinir Ağı Kullanılarak Anomali Tabanlı Saldırı Tespit Modeli Uygulaması," ***Avrupa Bilim ve Teknoloji Dergisi***, Ejosat Ek Ozel Sayi (HORA): 10–17, (2020).

[6] Bakhshi T. and Ghita B., "Anomaly detection in encrypted internet traffic using hybrid deep learning," ***Security and Communication Networks***, 1–16, (2021).

[7] Wei S., Zhang Z., Li S., and Jiang P., "Calibrating network traffic with one-dimensional convolutional neural network with autoencoder and independent recurrent neural network for mobile malware detection," ***Security and Communication Networks***, (2021):1–10, (2021).

[8] Arslan R. S., "Fasttrafficanalyzer: An efficient method for intrusion detection systems to analyze network traffic," ***Dicle Universitesi Muhendislik Fakultesi Muhendislik Dergisi***, 12:(4) 565–572, (2021).

[9] Pehlivanoglu M. K., Remzi A., and Odabas D. E., "Iki seviyeli hibrit makine ogrenmesi yontemi ile saldiri tespiti," ***Gazi Muhendislik Bilimleri Dergisi***, 5:(3), 258–272, (2019).

[10] Ozekes S. and Karakoc E. N., "Makine ogrenmesi yontemleriyle anormal ag trafiginin tespit edilmesi," ***Duzce universitesi Bilim ve Teknoloji Dergisi***, 7:(1), 566–576, (2019).

[11] Tokyurek E., "Birliktelik kural cikarim algoritmalari kullanilarak market sepet analizi," ***Master's thesis, Bilecik Seyh Edebali Universitesi***, Fen Bilimleri Enstitusu, (2019).

[12] Hidayanto B. C., Muhammad R. F., Kusumawardani R. P., and Syafaat A., "Network intrusion detection systems analysis using frequent item set mining algorithm fp-max and apriori," ***Procedia Computer Science***, 124:751–758, (2017).

[13] Moustafa N. and Slay J., "A hybrid feature selection for network intrusion detection systems: Central points," ***arXiv preprint*** arXiv:1707.05505, (2017).

[14] Aung K. M. M. and Oo N. N., "Association rule pattern mining approaches network anomaly detection," ***Ph.D. dissertation***, Meral Portal, (2015).

[15] Nalavade K. and Meshram B., "Mining association rules to evade network intrusion in network audit data," ***International Journal of Advanced Computer Research***, 4:(2), 560, (2014).

[16] Sokhangoee Z. F. and Rezapour A., "A novel approach for spam detection based on association rule mining and genetic algorithm," ***Computers & Electrical Engineering***, 97: 107655, (2022).

[17] Cekmez U., Erdem Z., Yavuz A. G., Sahingoz O. K., and Buldu A., "Network anomaly detection with deep learning," in 2018 ***26th Signal Processing and Communications Applications Conference (SIU)***. IEEE, 1–4, (2018).

[18] IDS 2017 Datasets- canadian institute for cybersecurity, https://www.unb.ca/cic/datasets/ids-2017.html, (Accessed on 06/30/2023).

[19] Budak H., "Ozellik secim yontemleri ve yeni bir yaklasim," ***Suleyman Demirel Universitesi Fen Bilimleri Enstitusu Dergisi***, 22: 21–31, (2018).

[20] Erkantarci B. and Bakal G., "An empirical study of sentiment analysis utilizing machine learning and deep learning algorithms," ***Journal of Computational Social Science***, 1–17, (2023).

[21] Bakal G., Talari P., Kakani E. V., and Kavuluru R., "Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations," ***Journal of biomedical informatics***, 82:189–199, (2018).

[22] Bakal G. and Kavuluru R., "Predicting treatment relations with semantic patterns over biomedical knowledge graphs," in ***International Conference on Mining Intelligence and Knowledge Exploration***. Springer, 586–596, (2015).

[23] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., and Duchesnay E., "Scikit-learn: Machine learning in Python," ***Journal of Machine Learning Research***, 12: 2825–2830, (2011).