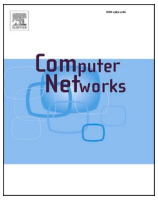
Computer Networks 188 (2021) 107840

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet



Machine learning methods for cyber security intrusion detection: Datasets and comparative study

Ilhan Firat Kilincer [a], Fatih Ertam [b,*], Abdulkadir Sengur [a]

[a] Department of Electrical-Electronics Engineering, Technology Faculty, Firat University, Elazig, Turkey
[b] Department of Digital Forensics Engineering, Technology Faculty, Firat University, Elazig, Turkey

ARTICLE INFO

Keywords:
IDS
KNN
SVM
DT
Machine learning
Cyber security

ABSTRACT

The increase in internet usage brings security problems with it. Malicious software can affect the operation of the systems and disrupt data confidentiality due to the security gaps in the systems. Intrusion Detection Systems (IDS) have been developed to detect and report attacks. In order to develop IDS systems, artificial intelligence-based approaches have been used more frequently. In this study, literature studies using CSE-CIC IDS-2018, UNSW-NB15, ISCX-2012, NSL-KDD and CIDDS-001 data sets, which are widely used to develop IDS systems, are reviewed in detail. In addition, max-min normalization was performed on these data sets and classification was made with support vector machine (SVM), K-Nearest neighbor (KNN), Decision Tree (DT) algorithms, which are among the classical machine learning approaches. As a result, more successful results have been obtained in some of the studies given in the literature. The study is thought to be useful for developing IDS systems on the basis of artificial intelligence with approaches such as machine learning.

1. Introduction

In recent years, especially with the developments in IoT technologies, the number of people and applications using the internet is constantly increasing. Internet usage is increasing according to DataReportal data, which provides information about internet usage in the world. Also, according to DataReportal data, 1 million people of internet users are added every day. The distribution of internet usage by years according to Datareportal data is given in Fig. 1. [1].

Increasing internet usage has also brought many security gaps. Many technologies such as firewall, data encryption, user authentication are used to prevent these security gaps. These security mechanisms prevent many types of attacks. However, these security technologies cannot perform in-depth packet analysis. For this reason, they cannot reach the desired level of attack detection. Intrusion Prevention System (IPS) and IDS systems have been developed to complement the shortcomings of these security mechanisms. These systems can perform deeper data analysis compared to other security systems thanks to their algorithms such as machine learning, deep learning, and artificial intelligence. While IPS systems work as both intrusion detection and prevention mechanisms, IDS systems are used only for intrusion detection and analysis [2–4]. In this study, we focused on IDS systems.

The increase in internet usage and data transfer speeds has also caused many anomalies [5]. As a result, attacks on the internet are constantly increasing. Fig. 2 shows the vulnerabilities and threads report published by Skybox Security in 2020. According to the graph given in Fig. 2, 17 220 new vulnerabilites were determined in 2019 and there is an increase of 3.8% compared to the previous year [6].

Institutions and organizations are constantly increasing their expenditure on cyber security technologies in order to provide a safe and stable service to their users. According to the report published by Crystal Market Research (CMR), the Cyber Security Market, which was valued at approximately USD 58.13 billion in 2012, is expected to reach USD 173.57 billion by 2022. Again, according to this report, the development of cloud storage and technologies such as the Internet of Things have increased the risk of data breaches. The size of the cyber security market published by CMR according to years is given in the graph in Fig. 3. [7].

The increase in internet usage has forced cyber security companies to produce more sensitive systems besides traditional security methods. As a result, proactive cyber security systems such as network behavior analysis, machine learning, threat analysis are developed. Nowadays, IDS systems are one of the technologies that are frequently used to become more sensitive to cyber threats.

IDS systems operate as signature-based or rule-based when detecting

* Corresponding author.
E-mail addresses: ifkilincer@firat.edu.tr (I.F. Kilincer), fatih.ertam@firat.edu.tr (F. Ertam), ksengur@firat.edu.tr (A. Sengur).

https://doi.org/10.1016/j.comnet.2021.107840
Received 15 October 2020; Received in revised form 6 December 2020; Accepted 4 January 2021
Available online 13 January 2021

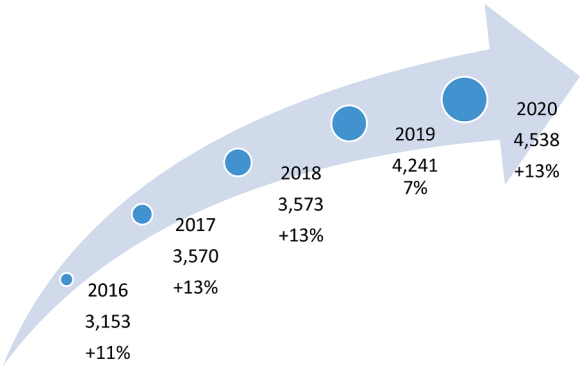


Fig. 1. Number of Internet users (in millions) change by year [1].

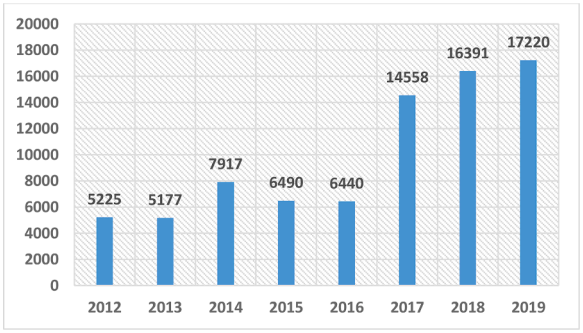


Fig. 2. New CVEs by year and the year those vulnerabilities were identified [6].

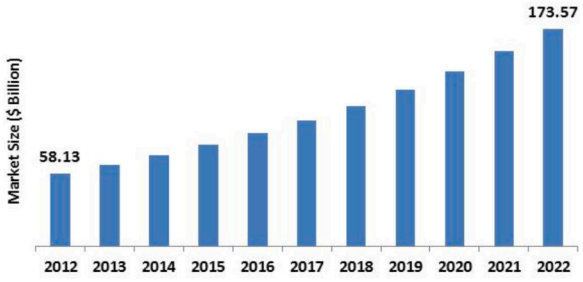


Fig. 3. Cyber Security Market [7].

and analyzing attacks. In signature based intrusion detection systems, IDS matches the data in the network with the attack patterns in its database. If a match is found, an alarm is generated. The major disadvantage of such signature-based IDS systems is that if the attack is not found in the IDS database, these attacks cannot be detected. In rule-based systems, also called anomaly-based IDS, normal behavior states are stored in the IDS database. These systems monitor all events on the network. An alarm is generated whenever there is deviation outside the specified rules. The biggest advantage of anomaly based intrusion detection systems is that they can detect unknown attacks. [2,8,9]. In Fig. 4, taxonomy of IDS systems is given.

New test environments are frequently created for more accurate traffic analysis and attack detection. As a result, new data sets are created. In this sense, some of these datasets have been examined in Section 3.

1.1. Motivation

The increase in internet usage has brought with it many security vulnerabilities. As seen in Fig. 3, the expenditures made by institutions and organizations in this area are constantly increasing in order to provide a stable and secure service to their users. In today's technology, methods such as network behavior analysis and machine learning are used to detect attacks early. Therefore, intrusion detection systems have become the most up-to-date research areas in organizations in the literature and in organizations related to cyber security.

When the literature studies on attack detection systems are examined, the following shortcomings are seen.

- Studies have generally been done on a small number of data sets. This does not provide a clear idea about the performance of the data sets. In this study, the performances of the 5 most frequently used data sets in the literature were examined in order to eliminate this gap in the literature.
- Generally, a limited number of classifiers or only one classifier were used in the studies. It cannot clearly express under which machine learning method the performance of these data sets is at the maximum level. Therefore, the performances of 5 data sets used in this study under the SVM, KNN, and DT classifiers, which are the most used in the literature, have been measured.
- Since the literature studies use a limited number of data sets, the performance of a limited number of attack types can be examined. In the 5 data sets that we considered in the study, there is information about most of the attack types in the literature. Therefore, the performance of attack types can be measured better.

1.2. Contributions

The contributions of this study, conducted using data sets developed on network-based intrusion detection systems, are summarized as follows.

- In the literature, machine learning based IDS models / methods have generally obtained results using a limited number of datasets. This does not give a clear idea of the attack detection capability of IDS datasets. In this research, we presented a machine learning based IDS method and to show general success of our method, the widely preffered five IDS datasets have been used for testing and our results are compared with the state-of-art methods in the literature. By using these datasets, a comparehansive comparisons are obtained. Also, SVM, KNN and DT which are the mostly used conventional classifiers are used to obtain comparable results. Because, the prior methods were generally used these classifiers. We also used variants of these classifiers for comprehensive evaluation. By using these classifiers, a benchmark for our presented model is presented. Behind these, the importance of network intrusion detection systems is emphasized and the studies on network intrusion detection systems in the literature are comprehensively examined and discussed.
- Some of the classes in the UNSW-NB15 dataset have been combined due to the relationship between them. The class merging process for this dataset is explained in detail in the proposed method. In addition, the classifier performance rates in the literature were examined for this dataset and high performance rates were obtained for this dataset (see Tables 4 and 6). Results of the other datasets (CSE-CIC-IDS-2018, UNSW-NB15, NSL-KDD, CIDDS-001 and ISCX 2012) are presented. The calculated results demonstrates general sucess of our method.

Other parts of the study are organized as follows. In Section 2, we provide an overview of related works. Public IDS datasets are presented in Section 3. Section 4 gives information about the classifiers used in the study. The proposed method and experimental results are presented in Sections 5 and 6, respectively. Finally, discussion, results and future studies are presented in Sections 7–9.

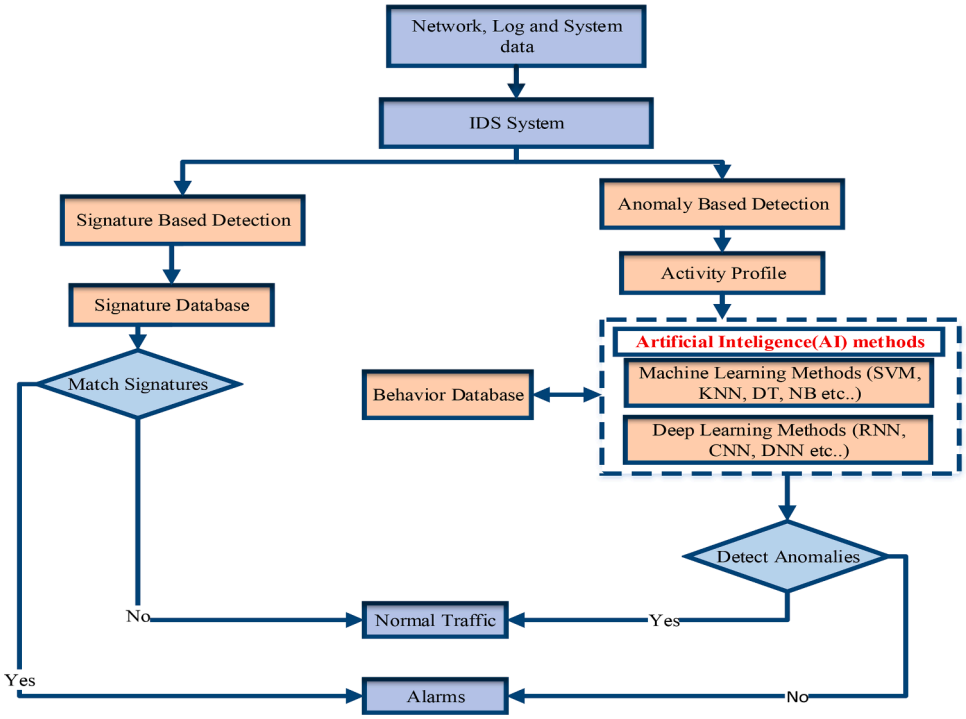


Fig. 4. Taxonomy of IDS systems.

2. Related works

Attackers are updating themselves and the software they use every day and create new attack scenarios. In this context, intrusion detection systems are being developed more and more every day so that the network systems can be effective against the developed malware. For this purpose, there are many literature studies and new studies are performed every day to increase the performance of IDS systems.

In this context, Hajisalem et al. [2] have developed a hybrid classification method using Artificial Bee Colony (ABC) and Artificial Fish Swarm (AFS) in their study. They made feature selection with Fuzzy C-Means Clustering (FCM) and Correlation-based Feature Selection (CFS) techniques. In the last step, they created If-Then rules with CART technique to distinguish normal and anomaly records. This method they developed was applied to NSL-KDD and UNSW-NB15 data sets and an accuracy rate of 99% was obtained. Inayat et al. [3] examines the design parameters of the existing intrusion response system (IRS) in their study. According to this study, there are many comprehensive studies in this field, but in the studies conducted, attack semantics are missing and they use static response metric instead of dynamic approach. This causes the system to generate more false alarms. Teodoro et al. [8] conducted a review of anomaly based intrusion detection systems. In this study, the most well known anomaly based intrusion detection systems, existing platforms, research and development projects are discussed.

Ferrag et al. [10] In their study, they used deep learning methods (recurrent neural networks (RNN), deep neural networks (DNN), restricted Boltzmann machines (RBM), deep belief networks (DBN), convoluted neural networks (CNN), deep Boltzmann machines (DBM), and deep autoencoders (DA) have implemented them on CSE-CIC-IDS2018 and Bot-IoT datasets. Then, classification success of deep learning and classification time of these data sets are compared. In addition, in their study, intrusion detection systems based on deep learning methods were examined, and in this sense, 35 attack detection data sets used in the literature were divided into categories. Sharafaldin et al. [11] created the CSE-CIC-IDS-2017 dataset, since the existing datasets did not meet today's intrusion detection needs. A test environment consisting of network attackers and victims has been set up, to create the data set. In the test environment, attacks such as Brute force, heartbleed attack, botnet, DoS, DDoS, Web attack, Infiltration attack were organized. In addition, system performance was evaluated using machine learning methods. Patil et al. [12] proposed the hypervisor level distributed network security (HLDNS) security framework of cloud computing in their study. In this method intrusions to each server with virtual machines are monitored. Feature similarity-based Fitness Function (FSFF) and Classifier Accuracy based Fitness Function (CAFF) fitness functions are used in conjunction with the BBA algorithm for feature extraction. The performance of the proposed method has been measured using the UNSW-NB15 and CICIDS-2017 datasets. Kim et al. [13] used CNN and RNN deep learning methods on CSE-CIC-IDS 2018 dataset and compared the performance of these two methods.

Kanimozhi et al. [14] classified the CSE-CIC-IDS 2018 data set using ANN, RF, k-NN, SVM, ADA BOOST, NB machine learning methods. Khammassi et al. [15] proposed a feature selection method based on Non-Dominated Sorting Genetic Algorithm II (NSGA-II) and logistic regression. The proposed approach was tested according to the Non-Dominated Sorting Genetic Algorithm Binomial Logistic Regression (NSGA2-BLR) and Non-Dominated Sorting Genetic Algorithm Multinomial Logistic Regression (NSGA2-MLR) methods. The best subsets obtained were classified by C4.5, Random Forest (RF), Naive Bayes (NB) methods. In the study, NSL-KDD, UNSW-NB15 and CIC-IDS2017 data sets were used. Ring et al. [16] conducted a comprehensive review of intrusion detection systems. In the study, data formats of network-based intrusion detection systems were analyzed. In addition, 15 features have been defined to evaluate the suitability of data sets. In addition, these features were collected in 5 groups: General Information, Quality of Data, Data Volume, Recording Environment, and Evaluation. In another study, Kanimozhi et al. [17] classified the CSE-CIC-IDS-2018 dataset using Artificial Inteligence (AI). As a result of the classification, 99.97% success has been achieved.

Moustafa et al. [18] has set up a test environment to better simulate modern network traffic. It produced the UNSW-NB15 data set from the test environment it established. IXIA Perfect-Storm, Tcpdump, Argus and Bro-IDS tools were used while producing the data set. Nine current attack scenarios were created in the test environment, Fuzzers, Analysis,

Table 1
. Literature summary.

Studies	Year	Feature Selection	Method	Datasets Used	Perfromance Metrics
Ferrag et al. [10]	2020	-	Deep disciriminative models. (DNN, RNN, CNN) Unsupervised models (RBM, DBN, DBM, DA)	CSE-CIC IDS2018 BoT-IoT	Accuracy, Accuracy Time(s)
Sharafaldin et al [11]	2018	-	KNN, RF, ID3, Adaboost, MLP, Naive Bayes, QDA	CSE-CIC-IDS2017	Precision, Recall, F1-Score
Patil et al. [12]	2019	BBA+CAFF BBA+FU BBA+FSFF	Random Forest	CSE-CIC-IDS2017 UNSW-NB15	Accuracy, FPR
Kim et al. [13]	2019	-	CNN	CSE-CIC IDS2018	Accuracy
Kanimozhi et al. [14]	2019	-	ANN, RF, k-NN, SVM, Adaboost, NB	CSE-CIC IDS2018	Accuracy, Precision, Recall, F1-Score
Khammassi et al. [15]	2020	NSGA2-BLR NSGA2-MLR	C4.5, RF, NB Tree	CSE-CIC-IDS2017 UNSW-NB15 NSL-KDD	Accuracy
Kanimozhi et al. [17]	2019	MLP	ANN	CSE-CIC-IDS2018	Accuracy, Precision, Recall
Moustafa et al. [19]	2016	-	DT, LR, NB, ANN, EM Clustering	UNSW-NB15 KDD99	Accuracy, FAR
Moustafa et al. [20]	2019	GAA-ADS	k-Means	UNSW-NB15 NSL-KDD	Accuracy, DR, FPR
Zhang et al. [21]	2020	-	Lenet, MSCNN, HAST, MSCNN-LSTM	UNSW-NB15	Accuracy, FAR, FNR
Gottwalt et al. [22]	2019	CorrCorr	CorrCorr	UNSW-NB15	Accuracy, DR, FPR
Khammassi et al. [23]	2017	GA-LR	C4.5, RF, NB Tree	UNSW-NB15 KDD99	Accuracy, DR, FAR
Kasongo et al. [24]	2020	WFEU	FFDNN	UNSW-NB15 AWID	Accuracy
Awad et al. [31]	2019	-	WELM	ISCX2012	Accuracy, Precision, Recall, F1-Score
Tan et al. [36]	2015	EMD	-	ISCX2012 KDDCUP99	Accuracy, DR, FPR
Bouteraa et al. [32]	2020	-	RF, SVM, ELM, RPART, OCSVM	ISCX2012	Accuracy, DR, FAR
Injadat et al. [33]	2018	BO	SVM, KNN, RF	ISCX2012	Accuracy, Precision, Recall, F1-Score
Yassin et al. [34]	2013	-	KMC + NBC	ISCX2012	Accuracy, DR, FAR
Hassan et al. [35]	2020	-	CNN + WDLSTM	ISCX 2012 UNSW-NB15	Accuracy
Hajisalem et al. [2]	2018	FCM, CFS	ABC, AFS	UNSW-NB15 NSL-KDD	Accuracy, DR, FPR
Rashid et al. [25]	2020	Hybrid (CFS, IGR, Gini İndex	k-NN, SVM, Naive Bayes(NB), DNN, DAE	CIDDS-001 NSL-KDD	Accuracy
David et al. [26]	2019	-	Dynamic Threshold Detection Algortihm	DARPA 98 DARPA 2000 Generated Dataset	Accuracy
Ertam et al. [27]	2017	-	NB, bN,RF, SMO, MLP	KDD Cup 99	Accuracy
Kolias et al. [28]	2016	-	Adaboost, Hyperpipes, J48, NB, OneR, RF, ZeroR	AWID	Accuracy
Verma et al. [29]	2018	-	kNN	CIDDS-001	Accuracy
Verma et al. [30]	2018	-	kNN, SVM, DT, RF, NB, DL, ANN, SOMs, EM, k-means	CIDDS-001	Accuracy

Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms, using the IXIA tool. Moustafa et al. [19] examined the complexity of the UNSW-NB15 data set in another study. For this purpose, in the first step, statistical analysis of qualifications is explained. In the second step, feature correlations are examined. In the last step, the performance of the data set with five classifiers was measured and compared with the KDD99 data set. As a result, UNSW-NB15 has been observed to be more complex than KDD99. Moustafa et al. [20] presented another study, Geometric Area Analysis (GAA) technique based on Trapezoidal Area Estimation (TAE) estimation, calculated from Beta Mixture Model (BMM) parameters, for distances between features and observations. This method has been tested on NSL-KDD and UNSW-NB15 datasets. Zhang et al. [21] proposed a unified method combining Multiscale Convolutional Neural Network (MSCNN) with Long Short-Term Memory (LSTM). In the first stage of the method, MSCNN was used to analyze the spatial properties of the data set. In the second stage of the method, LSTM network was used to process temporary features. UNSW-NB15 dataset was used for the training and testing of the model. The method has better accuracy, false alarm rate and false negative speed than models based on traditional neural networks. Gottwalt et al. [22] Corr Corr feature selection method based on multivariate correlation (MC) has been proposed for multivariate correlation-based network anomaly detection. The method was tested on UNSW-NB15 and NSL-KDD datasets.

According to Khammasi et al. [23] in order to classify IDS systems in a shorter time and more successfully, it is necessary to reveal the features that represent the entire data set. In this sense, they applied wrapper logistic regression based on genetic algorithm to select the best features in the intrusion detection system. Tests were performed using KDD99 and UNSW-NB15 datasets and compared with other studies. Kasongo et al. [24] hey conducted tests on AWID and UNSW-NB15 datasets using the Wrapper Based Feature Extraction Unit (WFEU) feature extraction method and Feed-Forward Deep Neural Network (FFDNN). Rashid et al. [25] classified the NSL-KDD and CIDDS-0001 datasets with SVM, Naive Bayes (NB), KNN, Neural networks, DNN and DAE classification algorithms. In order to increase the classifier success, hybrid feature selection and sorting methods were used before the classification algorithms and high accuracy rates were obtained.

David et al. [26] used a statistical approach based on Dynamic threshold algorithm to detect DDoS attacks. DARPA dataset was used to test the method. According to the experimental results, the proposed method required higher detection rates and less processing time than existing methods. Ertam et al. [27] classified on the KDD Cup 99 data set using NB, bayes Net (bN), Random Forest (RF), Multilayer Perception (MLP), Sequential Minimal Optimization (SMO) algorithms. Kolias et al. [28] thoroughly evaluated the attacks on 802.11 networks and

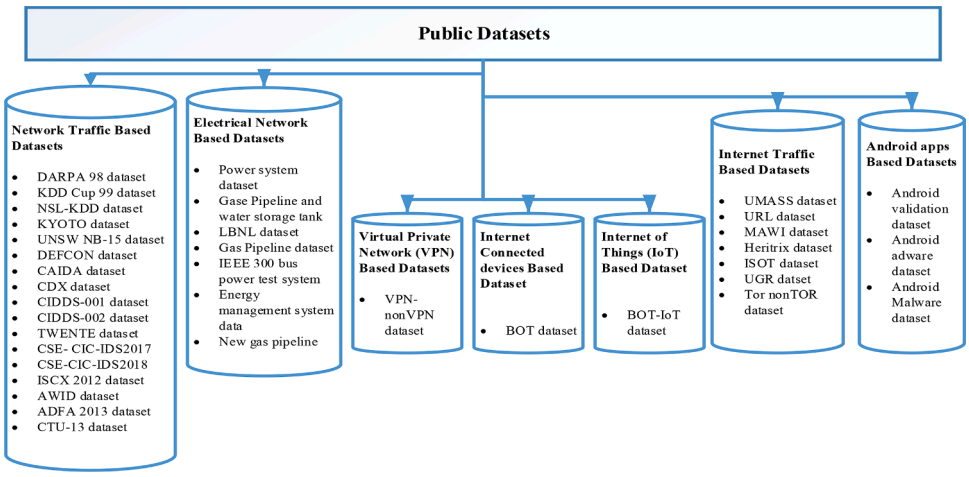


Fig 5. Public datasets and their classes [10].

Table 2
Cyber security intrusion detection datasets.

Data Set Name	Year	Data Set Literature	Data Set Feature Count	Data Set Attack Types (Classes)
CSE-CIC IDS2017	2017	[11,12,14,16,15]	80	DoS Golden Eye, Heartbleed, DoS hulk, DoS Slow http, DoS Slowloris, DDoS, SSH-Patator, FP, Patator, Brute force, XSS, Botnet, infiltration, PortScann, SQL injection.
CSE-CIC IDS2018	2018	[10,13,14,17]	80	DoS Golden Eye, Bening, DoS hulk, DoS Slow http, DoS Slowloris, DDoS-LOIC HTTP, DDoS-LOIC-UDP, DDoS-HOIC, SSH-Patator, FTP Patator, Brute force, XSS, Botnet, infiltration, SQL injection.
UNSW-NB15	2015	[12,15,16,18,19,20,21, 22,23,24,37,2,35]	49	Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms
ISCX-2012	2012	[16,35,31,36,32,34]	80	Normal, Attacker
NSL-KDD	1998	[15,16,20,2,38,39,25,40]	43	Normal, DoS, Remote to Local (R2L), User to Root(U2R), Probing
DARPA	1998	[16,37,38,26,41],	43	DoS, Remote to Local (R2L), User to Root(U2R), Probing
KYOTO	2006-2009	[16,42],	23	Oth, rej, rsto, rstos0, rstr, rstrh, s0, s1, s2, s3, sf, sh, shr
KDD99	1998	[16,19,23,36,38,40,37, 27],	43	DoS, Remote to Local (R2L), User to Root(U2R), Probing
AWID	2015	[16,24,28,43]	156	Normal, Flooding, Injection, Impersonation
CIDDS-001	2017	[16,25,44,45,29,30]	14	Normal, Attacker, Victim, Suspicious, Unknown

presented the AWID dataset in this context. Later, he classified the AWID data set using Adaboost, Hyperpipes, J48, NB, OneR, RF, ZeroR algorithms. Verma et al. [29] classified the data set CIDDS-001, a flow-based and labeled dataset, using k-nearest neighbor classification and k-mean clustering algorithms. For classification, external server and openstack server data were evaluated separately. In another study Verma et al. [30] In another study, classified the CIDDS-001 dataset using k-Nearest Neighbor (kNN), SVM, Decision Tree (DT), RF, NB, DL, Artificial Neural Neworks (ANN), Self Organizing Maps (SOMs), Expectation Maximisation (EM) algorithms.

Awad et al. [31] combined the layered sampling method and different cost function schemes with Weighted Extreme Machine Learning (WELM). They applied their proposed method on the ISCX2012 dataset. Tan et al. proposed the Earth Mover's Distance (EMD) method on ISCX 2012 and KDDCup99 datasets to detect DoS attacks. Bouteraa et al. [32] Conducted a comparative study of data mining techniques for intrusion detection using the ISCX 2012 dataset. Injadat et al. [33] have proposed an effective framework for anomaly detection. They used RF, SVM and k-NN machine learning methods in their studies. They used Bayesian Optimization (BO) technique to adjust SVM, RF and KNN

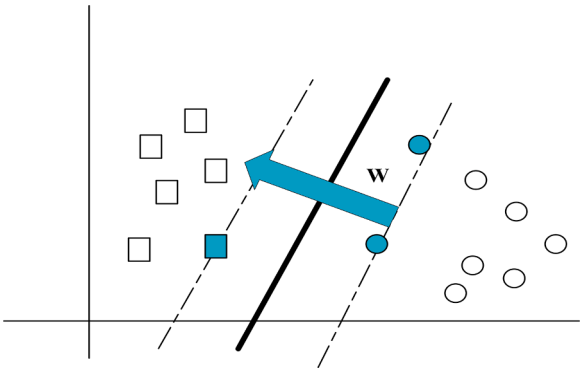


Fig 6. Representation of SVM hyperplanes.

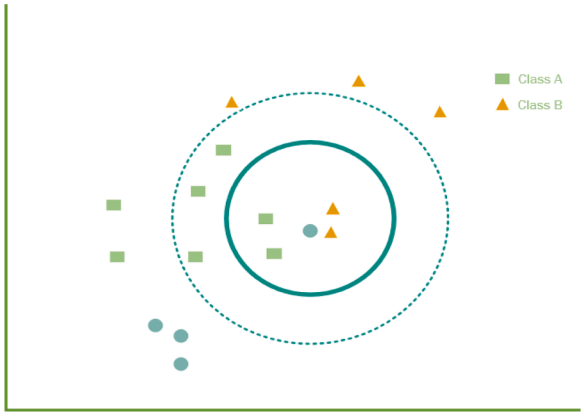


Fig 7. KNN sample drawing.

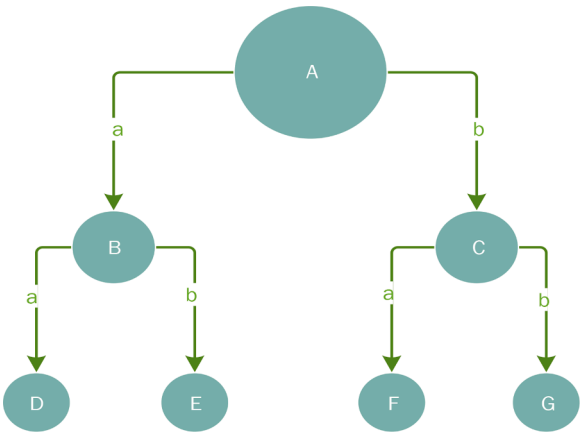


Fig 8. A simple decision tree model.

parameters. Yassin et al. [34] proposed the KMC + NBC algorithm using the K-Means Clusterring (KMC) and Niave Bayes Classifier (NBC) classifiers in an integrated manner. They used the ISCX2012 data set in their studies. Hassan et al. [35] proposed a hybrid deep learning model in order to detect network attacks with high accuracy values. For this purpose, they used CNN and weight-dropped, long short-term memory (WDLSTM) deep learning methods together. They used UNSW-NB15 and ISCX2012 datasets to measure the performance of their proposed method. The summary of the studies on intrusion detection systems is given in Table 1.

3. Public IDS datasets and used cyber security attack types

In this section, commonly used IDS datasets and attack types in these datasets are mentioned.

3.1. Datasets

There are many data sets that can be used in cyber security attack detection. Ferrag et al. [10] divided public data sets into 7 classes according to their content. Data sets and classes are given in Fig. 5.

In this study, network traffic based datasets, which are frequently used in intrusion detection systems, are used. The year, literature studies, feature numbers and attack classes of these data sets are summarized in Table 2. Also some of the most used datasets are explained.

DARPA Dataset: The DARPA dataset is a network-based dataset produced in the MT Lincoln Laboratory in 1998. Training data includes seven weeks of network-based attacks. Test data also includes two-week network-based attacks [10,38,41].

KDD 99 Dataset: KDD Cup 99 data set is based on DARPA'98 data set program. DoS, Remote to Local (R2L), User to Root (U2R), Probing attacks are simulated. The dataset contains 7 weeks of network traffic and consists of about 5 million lines. This dataset is one of the most widely

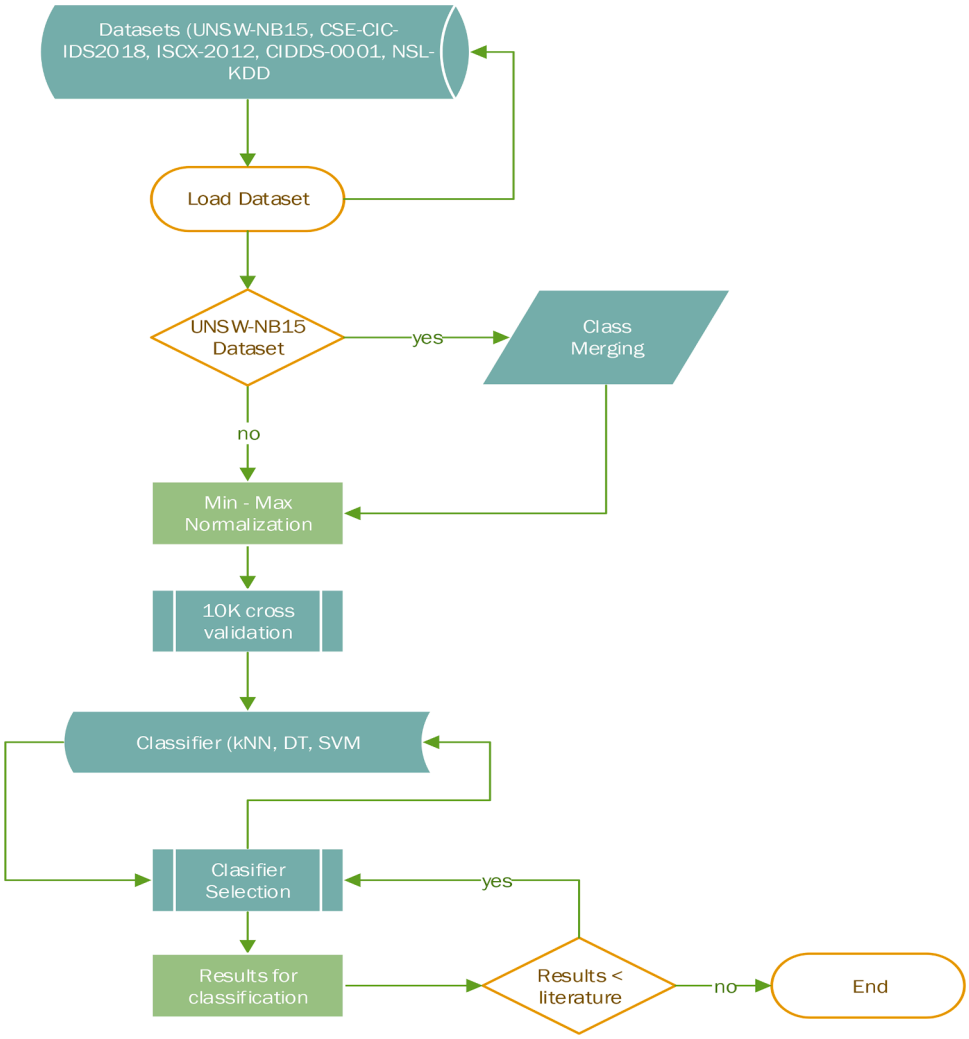


Fig 9. Flowchart of the proposed method.

Table 3
Data count for each class.

Datasets	Classes	Data Count
CSE-CIC-IDS 2018	Benign	1552
	Bot	1541
	Brute Force-WEB	363
	Brute Force-XSS	152
	DDOS attack-HOIC	1552
	DDOS attack-LOIC-UDP	1561
	DoS attacks-GoldenEye	1557
	DoS attacks-Hulk	1563
	DoS attacks-SlowHTTPTest	1523
	DoS attacks-Slowloris	1584
	FTP-BruteForce	1527
	infiltration	1546
	SQL Injection	54
	SSH-Bruteforce	1452
ISCX 2012	Normal	2517
	Attacker	2515
NSL-KDD	Normal	6817
	DoS	11617
	Probe	988
	R2L	53
	U2R	3086
CIDDS-001	Normal	633
	Attacker	985
	Suspicious	655
	Unknown	662
	Victim	1135
UNSW-NB15 with Categorization	Analysis	678
	DoS	2593
	Exploits	2431
	Generic	2089
	Normal	2587
	Reconnaissance	2253
	Worms	40

used datasets for the assessment of intrusion detection models. [10,38,40,46].

NSL-KDD Dataset: NSL-KDD dataset has been developed to solve problems in KDD 99 dataset. It does not contain unnecessary and repetitive records according to the original KDD 99 data set. It contains a reasonable number of records. As a result of removing duplicate and unnecessary records, the data set has been reduced from approximately 5 million records to 150,000 records. It is also divided into predefined training and test subsets for intrusion detection methods. NSL-KDD uses the same properties and classes as KDD CUP 99. DoS, Remote to Local (R2L), User to Root (U2R), Probing attacks in KDD 99 data set are also simulated for this data set. [10,38,40,16].

Kyoto Dataset: This dataset includes 24 features. 14 of these features were created from the KDDCup99 dataset. These 14 features were chosen among 41 features of KDD Cup 99 data set with honeypot systems installed at Kyoto University. In addition to these 14 features, 10 new features have been created for a more effective intrusion detection system [10,42].

CIDDS-001 Dataset: The CIDDS-001 (Coburg Network Intrusion Detection Dataset) data set was created in an anomaly-based network intrusion detection system. Scripts written in nmap and python language were used during the creation of the data set. Portscan, Pingscan, DoS and Brute force attacks have been done. The data set contains 14 features and consists of the Normal, Attacker, Victim, Suspicious, Unknown classes. The data set consists of Openstack and traffic data from an external server. [44,45,29,30]

ISCX-2012 Dataset: This data set was created from seven-day network data. The dataset contains normal and malicious network traffic. Malicious network traffic includes Infiltrating the network from inside, HTTP Denial of Service, Distributed Denial of Service, and Brute Force SSH attacks. There are Normal and Attacker classes [10,47].

AWID Dataset: AWID is a data set created in networks for the IEEE 802.11 standard. The data set contains 156 features. While creating the dataset, 23 different attack methods such as honeypot, rogue Access point, evil twin, deatuhentication attack, dictionary, fregmantation that occur frequently on 802.11 networks are used. The attacks were then classified with the labels Normal, Flooding, Injection and Impersonation. [16,28,43].

CSE-CIC-IDS 2017 Dataset: This dataset was created in 2017 by Communications Security Establishments (CSE) & the Canadian Institute for Cybersecurity (CIC). Offensive and victim networks were created in the test environment prepared for the creation of Dataset. In order to create the dataset, a laboratory environment with attacker and victim networks has been set up. In the network where the attacks were made, there were a switch, a computer with a Kali Linux operating system and three computers with a Windows 8 operating system. In the target network, there is one Windows Server 2016, one server with Ubuntu 16 operating system, one server with Ubuntu 12 operating system, one router and one firewall. Active directory feature has been opened on Windows Server 2016 and all devices in the victim network are in this domain. In addition, the router's uplink port is mirrored in order to listen to all traffic on the victim network [11,10,48].

In the CSE-CIC-IDS dataset, an agent based on the java-B-profile system was written to generate normal traffic. With this agent, some protocol-based attributes such as HTTP, HTTPS, FTP, SSH and e-mail have been reproduced using machine learning and statistical methods. Some tools such as Patator, slowloris, Slowhttps, Metasploit, Ares have been used to generate attack traffic data. Attacks such as brute force, heartbleed attack, botnet, DoS, DDoS, Web attack, Infiltration attack were organized. A total of 14 different attack types are labeled in the dataset. Tagged attack types are: DoS Golden Eye, Heartbleed, DoS hulk, DoS Slow http, DoS Slowloris, DDoS, SSH-Patator, FP, Patator, Brute force, XSS, Botnet, infiltration, PortScann, SQL injection. In addition, CICFlowMeter was used while creating the data set and 80 features were created from the captured traffic [11,48].

CSE-CIC-IDS 2018: It shows the same features as the data set created in 2017. However, more devices were used in the test environment to better model the attacks. To create this dataset, the infrastructure of the attacker network includes 50 machines, the victim network 420 machines and 30 servers. In addition, the victim network is divided into 6 departments: R&D department (Dep1), Management Department (Dep2), Technician department (Dep3), Secretary and operations department (Dep4), IT department (Dep5) and server rooms. In addition, different Windows operating systems (Windows 8.1 and Windows 10) have been installed for all departments except the IT department. An Ubuntu operating system is installed on all computers in the IT department. Different server operating systems such as Windows server 2012 and Windows Server 2016 have been installed for the server room. Thus, a topology with a machine diversity similar to real world networks was created. 7 different attack scenarios, such as Brute force, heartbleed attack, botnet, DoS, DDoS, Web attack, Infiltration attack, were applied to create this data set. [11,49]

UNSW-NB15 Dataset: This dataset was created by the Australian Center for Cyber Security (ACCS). While creating the dataset, IXIA Perfect-Storm, Tcpdump, Argus and Bro-IDS tools were used. The IXIA tool, which is used as a normal and abnormal traffic generator, is built on three virtual servers. Using IXIA tool, 9 attack scenarios were created: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms. The IXIA tool uses the CVE site to create a modern threat environment. Routers used in the test environment are connected to the firewall. The firewall is configured to pass all traffic normally or abnormally. One of the routers was run tcpdump and data packets received during the simulation were recorded. The data obtained using Argus, Bro-IDS tools and twelve algorithms in C # has been extracted into 49 features [10,18,19].

3.2. Attack types

In this section, some of the attack types used in literature studies on

Table 4
Classification results for each dataset.

Datasets	Classification Methods		Accuracy	Precision	Recall	Geometric Mean	F-Measure
CSE-CIC IDS 2018	**SVM Linear**	**Best**	0.9964	0.9864	0.9864	0.9713	0.9803
		Mean	0.9964	0.9862	0.9862	0.9712	0.9802
		Std	0.0000	0.0003	0.0003	0.0001	0.0001
	SVM Quadratic	**Best**	**0.9981**	**0.9941**	**0.9941**	**0.9929**	**0.9935**
		Mean	0.9960	0.9901	0.9901	0.9863	0.9884
		Std	0.0016	0.0023	0.0023	0.0043	0.0031
	SVM Cubic	**Best**	0.9913	0.9876	0.9876	0.9836	0.9858
		Mean	0.9882	0.9665	0.9665	0.9367	0.9573
		Std	0.0027	0.0181	0.0181	0.0402	0.0242
	KNN Fine	**Best**	**0.9918**	**0.9577**	**0.9577**	**0.9463**	**0.9546**
		Mean	0.9914	0.9555	0.9555	0.9444	0.9527
		Std	0.0003	0.0021	0.0021	0.0017	0.0014
	KNN Medium	**Best**	0.9865	0.9582	0.9582	0.9108	0.9414
		Mean	0.9862	0.9543	0.9543	0.9065	0.9398
		Std	0.0002	0.0025	0.0025	0.0025	0.0015
	KNN Cubic	**Best**	0.9865	0.9547	0.9547	0.9108	0.9414
		Mean	0.9859	0.952	0.952	0.9042	0.9381
		Std	0.0005	0.0024	0.0024	0.0049	0.0024
	TREE Fine	**Best**	**0.9992**	**0.9994**	**0.9994**	**0.9994**	**0.9994**
		Mean	0.9989	0.9975	0.9975	0.9976	0.9976
		Std	0.0001	0.001	0.001	0.0009	0.0009
	TREE Medium	**Best**	0.9992	0.9994	0.9994	0.9994	0.9994
		Mean	0.9929	0.9975	0.9975	0.7982	0.9976
		Std	0.0122	0.001	0.001	0.4031	0.0009
NSL-KDD	**SVM Linear**	**Best**	0.9847	0.9517	0.9517	0.8579	0.9156
		Mean	0.9847	0.9491	0.9491	0.8546	0.9133
		Std	0.0001	0.0037	0.0037	0.0047	0.0032
	SVM Quadratic	**Best**	0.9932	0.9635	0.9635	0.9181	0.9447
		Mean	0.9931	0.9627	0.9627	0.9129	0.9423
		Std	0.0001	0.0011	0.0011	0.0074	0.0034
	SVM Cubic	**Best**	**0.9946**	**0.971**	**0.971**	**0.9146**	**0.944**
		Mean	0.9945	0.9676	0.9676	0.909	0.9435
		Std	0.0002	0.0047	0.0047	0.0079	0.0008
	KNN Fine	**Best**	**0.9964**	**0.9808**	**0.9808**	**0.9476**	**0.9657**
		Mean	0.9964	0.9751	0.9751	0.9474	0.963
		Std	0.0001	0.0081	0.0081	0.0003	0.0038
	KNN Medium	**Best**	0.9915	0.9477	0.9477	0.8837	0.9227
		Mean	0.9914	0.9441	0.9441	0.8811	0.9217
		Std	0.0001	0.0051	0.0051	0.0037	0.0013
	KNN Cubic	**Best**	0.9909	0.9388	0.9388	0.8836	0.9199
		Mean	0.9909	0.9319	0.9319	0.8834	0.9165
		Std	0.0011	0.0098	0.0098	0.0002	0.0048
	TREE Fine	**Best**	**0.9992**	**0.9994**	**0.9994**	**0.9994**	**0.9994**
		Mean	0.9939	0.8353	0.8353	0.8353	0.8353
		Std	0.0112	0.3685	0.3685	0.3685	0.3685
	TREE Medium	**Best**	0.9992	0.9994	0.9994	0.9994	0.9994
		Mean	0.9937	0.8451	0.8353	0.8168	0.8351
		Std	0.0113	0.3784	0.3685	0.3867	0.3675
ISCX 2012	**SVM Linear**	**Best**	0.9849	0.9819	0.9819	0.9871	0.9845
		Mean	0.9843	0.9812	0.9812	0.9865	0.9839
		Std	0.0005	0.0005	0.0005	0.0004	0.0005
	SVM Quadratic	**Best**	**0.998**	**0.9978**	**0.9978**	**0.998**	**0.9979**
		Mean	0.9897	0.9877	0.9877	0.991	0.9894
		Std	0.0075	0.009	0.009	0.0062	0.0076
	SVM Cubic	**Best**	0.9849	0.9819	0.9819	0.9871	0.9845
		Mean	0.8746	0.8696	0.8696	0.8734	0.8722
		Std	0.1517	0.155	0.155	0.1581	0.1556
	KNN Fine	**Best**	**0.998**	**0.9978**	**0.9978**	**0.998**	**0.9979**
		Mean	0.9897	0.9878	0.9878	0.9911	0.9895
		Std	0.0076	0.0092	0.0092	0.0063	0.0077
	KNN Medium	**Best**	0.9964	0.9959	0.9959	0.9967	0.9963
		Mean	0.9891	0.987	0.987	0.9906	0.9888
		Std	0.0067	0.0081	0.0081	0.0056	0.0068
	KNN Cubic	**Best**	0.996	0.9954	0.9954	0.9964	0.9959
		Mean	0.989	0.9868	0.9868	0.9905	0.9887
		Std	0.0065	0.0078	0.0078	0.0054	0.0066
	TREE Fine	**Best**	**1**	**0.9994**	**0.9994**	**1**	**0.9994**
		Mean	0.994	0.99	0.991	0.8368	0.8368
		Std	0.0112	0.011	0.01	0.369	0.369
	TREE Medium	**Best**	1	0.9994	0.9994	1	0.9994
		Mean	0.994	0.99	0.991	0.8368	0.8368
		Std	0.0112	0.011	0.01	0.369	0.369
CIDDS-001	**SVM Linear**	**Best**	0.9284	0.9172	0.9172	0.9062	0.9147
		Mean	0.9278	0.9166	0.9166	0.9054	0.914

(*continued on next page*)

Table 4 (*continued*)

Datasets	Classification Methods		Accuracy	Precision	Recall	Geometric Mean	F-Measure
		Std	0.0009	0.0009	0.0009	0.0011	0.001
	SVM Quadratic	**Best**	0.9678	0.9648	0.9648	0.9593	0.9631
		Mean	0.9663	0.9628	0.9628	0.9572	0.9611
		Std	0.001	0.0013	0.0013	0.0011	0.0012
	SVM Cubic	**Best**	**0.9729**	**0.9682**	**0.9682**	**0.9668**	**0.9679**
		Mean	0.9674	0.9636	0.9636	0.9589	0.9622
		Std	0.0025	0.0022	0.0022	0.0036	0.0026
	KNN Fine	**Best**	**0.9781**	**0.9732**	**0.9732**	**0.9726**	**0.9731**
		Mean	0.976	0.9705	0.9705	0.9698	0.9705
		Std	0.001	0.0012	0.0012	0.0012	0.0012
	KNN Medium	**Best**	0.9631	0.9553	0.9553	0.9541	0.9554
		Mean	0.9606	0.9524	0.9524	0.9512	0.9524
		Std	0.0012	0.0015	0.0015	0.0015	0.0014
	KNN Cubic	**Best**	0.9611	0.9534	0.9534	0.9522	0.9534
		Mean	0.958	0.9495	0.9495	0.9482	0.9495
		Std	0.0015	0.0019	0.0019	0.002	0.0019
	TREE Fine	**Best**	**0.9966**	**0.9957**	**0.9957**	**0.9957**	**0.9957**
		Mean	0.9947	0.9935	0.9935	0.9935	0.9935
		Std	0.0008	0.001	0.001	0.001	0.001
	TREE Medium	**Best**	0.9914	0.9894	0.9894	0.9891	0.9893
		Mean	0.9902	0.988	0.988	0.9877	0.9879
		Std	0.0006	0.0007	0.0007	0.0007	0.0007
UNSW-NB15 with Categorization	**SVM Linear**	**Best**	0.9937	0.8521	0.8521	0.8563	0.8534
		Mean	0.9935	0.9133	0.9141	0.511	0.8974
		Std	0.0001	0.0572	0.0574	0.4674	0.0408
	SVM Quadratic	**Best**	**0.9962**	**0.9668**	**0.9668**	**0.8563**	**0.9346**
		Mean	0.9949	0.9143	0.9143	0.512	0.8986
		Std	0.0014	0.0574	0.0574	0.4674	0.0415
	SVM Cubic	**Best**	0.9955	0.9382	0.9382	0.8631	0.922
		Mean	0.9928	0.9004	0.9004	0.5088	0.8896
		Std	0.0037	0.0447	0.0447	0.4646	0.0341
	KNN Fine	**Best**	**0.9546**	**0.8239**	**0.8239**	**0.6858**	**0.8182**
		Mean	0.9534	0.821	0.821	0.6578	0.8157
		Std	0.0006	0.0021	0.0021	0.0225	0.0022
	KNN Medium	**Best**	0.9479	0.8138	0.8138	0.6861	0.7942
		Mean	0.9465	0.8131	0.8128	0.2703	0.7940
		Std	0.0007	0.0181	0.0181	0.3703	0.0256
	KNN Cubic	**Best**	0.9448	0.8103	0.8103	0.6758	0.7905
		Mean	0.9432	0.801	0.821	0.6379	0.7888
		Std	0.001	0.0019	0.0021	0.0211	0.0022
	TREE Fine	**Best**	**0.9984**	**0.979**	**0.979**	**0.9507**	**0.968**
		Mean	0.9982	0.9736	0.9736	0.9505	0.9652
		Std	0.0002	0.0049	0.0049	0.0002	0.0025
	TREE Medium	**Best**	0.9984	0.9942	0.9942	0.9508	0.9754
		Mean	0.9983	0.9926	0.9926	0.9493	0.9741
		Std	0.0001	0.0027	0.0027	0.0026	0.0023
UNSW-NB15 Without Categorization	**SVM Quadratic**	**Best**	0.7795	0.65	0.65	0.4302	0.619
		Mean	0.7738	0.6316	0.6316	0.2086	0.6101
		Std	0.0052	0.0062	0.0062	0.2199	0.0014
	KNN Medium	**Best**	0.7187	0.5606	0.5606	0.2798	0.5355
		Mean	0.7168	0.5371	0.5371	0.2593	0.5234
		Std	0.0013	0.0103	0.0103	0.014	0.0058
	TREE Fine	**Best**	**0.8057**	**0.7229**	**0.7229**	**0.4934**	**0.6687**
		Mean	0.8021	0.6916	0.6916	0.4594	0.6501
		Std	0.0018	0.0198	0.0198	0.0288	0.0114

IDS systems are examined.

Vulnerability Scan: It is the scanning process for the detection of security vulnerabilities in the system. Software such as Nmap, Nessus, OpenVAS are frequently used for this process. This process, in addition to vulnerability scanning, has many functions such as collecting information about the target system, extracting network map of the target system. The first stage of the cyber attack architecture is vulnerability detection. After this stage, the attack methodology is determined [50, 51].

Denial of Service (DoS): It is the type of attack made by exploiting resource capacities such as RAM and CPU of the target system. In this type of attack, the target system is deactivated and the system can not serve [52].

Distributed Denial of Service (DDoS): They are the attacks to deactivate the target system much more quickly. DoS attacks are carried out from multiple sources simultaneously. In DDoS attacks, in order to deactivate the target system, bandwidth is usually exploited or many connection requests are sent to the system at the same time. Thus, the system becomes unable to respond to incoming connections. This attack scenario is done by many methods such as ICMP flooding, HTTP flooding, TCP flooding [53,54].

Brute Force Attack: It is a type of attack, to get full authority on the target system. The aim is to obtain the password of the user, who authorized to system. In order to obtain the login information of the target system, data is uploaded to the target system. In this method, if there is no information available, different username and password combinations are used and the password is tried to be cracked. This type of attack is frequently applied to systems using protocols such as Telnet, SSH, RDP, FTP, HTTP. [55,56].

Exploit: It is a type of attack on the target system to raise authority. The system is seized with pieces of code prepared for vulnerabilities in software or hardware in the systems. There are many types of exploits.

Table 5
Literature studies for CSE-CIC-IDS-2018 dataset.

Reference	Classification Method	Feature Selection	Accuracy	Precision	Recall	F1-Score	ACC Time(s)
Ferrag et al. [10]	**DNN**	-	0.9728	-	-	-	390.2
	RNN		0.9731	-	-	-	334.7
	CNN		0.9376	-	-	-	331.2
	RBM		0.9728	-	-	-	390.1
	DBN		0.9730	-	-	-	344.7
	DBM		0.97371	-	-	-	351.5
	DA		0.97372	-	-	-	341.3
Kim et al. [13]	**CNN**	-	SD-1 (0.9998)	-	-	-	-
			SD-2 (0.9005)	-	-	-	-
			SD-3 (0.9677)	-	-	-	-
			SD-4 (0.9998)	-	-	-	-
			SD-5 (0.9997)	-	-	-	-
			SD-6 (0.9995)	-	-	-	-
			SD-7 (0.9994)	-	-	-	-
			SD-8 (0.8879)	-	-	-	-
			SD-9 (0.7601)	-	-	-	-
			SD-10 (0.9997)	-	-	-	-
Kanimozhi et al. [14]	**ANN**	-	0.9997	0.9996	1.0	0.9998	-
	RF		0.9983	0.9992	0.9988	0.9992	-
	k-NN		0.9973	0.998	0.9988	0.9984	-
	SVM		0.998	0.9	0.9988	0.9994	-
	Adaboost		0.9996	0.9996	0.9988	0.9992	-
	NB		0.992	0.9929	0.9976	0.9953	-
Kanimozhi et al. [17]	**ANN**	**MLP**	0.9998	0.9995	1.0	-	-
Proposed Method	**Fine Tree**	-	**0.9992**	**0.9994**	**0.9994**	**0.9992**	-
	Medium Tree		0.9992	0.9994	0.9994	0.9994	-
	Cubic Tree		0.9865	0.9547	0.9547	0.9414	-
	Fine KNN		**0.9918**	**0.9577**	**0.9577**	**0.9546**	-
	Medium KNN		0.9865	0.9582	0.9582	0.9414	-
	Cubic KNN		0.9865	0.9547	0.9547	0.9414	-
	Linear SVM		0.9964	0.9864	0.9864	0.9803	-
	Quadratic SVM		**0.9981**	**0.9941**	**0.9941**	**0.9935**	-
	Cubic SVM		0.9913	0.9876	0.9876	0.9858	-

But the most dangerous ones are Zero Day Exploits. Because these exploits are written as soon as a security vulnerability is detected in the system and system administrators are not aware of this vulnerability. [57,58].

SQL Injection: It is mostly an attack on Web applications. Vulnerabilities in websites' databases are used and the database is compromised. Thus, user information and confidential information on websites can be accessed. This type of attack ranks first in OWASP (Open Web Application Security Project), which deals with web application security [59, 60].

In addition to these types of attacks, DARPA, KDD99 and NSL-KDD data sets, which are frequently used in studies, have attack types such as Remote to Local (R2L), User to Root (U2R), Probing. In the U2R attack type, the attacker hijacks a normal user account on the system and then exploits vulnerabilities to gain root access to the system. In the R2L attack type, the attacker sends a packet to the target machine over the network and tries to gain local access as the user of this machine. Finally, probing attacks are used to detect vulnerabilities in the target network and are essentially the same as the Vulnerability Scan mentioned above [61].

3.3. Classification methods

In order to compare the performance of the data sets used in this study, the data sets were classified by SVM, KNN and DT methods.

3.4. Support vector machine (SVM)

It is a strong classifier based on statistical methods and structural risk minimization. It is frequently used in classification processes as it is easy to apply and provides good performance. SVM uses a multidimensional hyperplane to classify samples and draws a boundary between points in the plane. The limit should be at the furthest distance to the datasets. Newly added data are classified by looking at this borderline [38, 62, 63]. When learning data is $x_i \in R^d,\ i = 1....n\ ,\ y_i \in \{-1, +1\}$ in SVM, if labeled with $\{x_i, y_i\}$ an example of two classes of data sets is formed in the two-dimensional space given in Figure 6. In Figure 6, the thick line separator shows the hyper plane, while w: shows the normal vector. Assuming that the hyper plane separates positive and negative samples, in the two class linear classifier problem, the normal vector of the hyper plane becomes w ve and the offset value b. Accordingly, the decision limit is $w^T x + b = 0$ line. In this case, the conditions in equation 1 and equation 2 must be ensured [64, 65].

$$w^T x_i + b \geq 1 \tag{1}$$

$$w^T x_i + b \leq -1 \tag{2}$$

3.5. K-Nearest neighbour (KNN)

KNN is a supervised learning method frequently used in classification processes. This algorithm stores existing data and categorizes new incoming data in terms of distance to existing data. The distance of the new incoming data to the existing data is calculated and the close neighborhood of "k" is checked. The data is included in the class of its neighbor, which is closest to it. Euclidean, Manhattan and Minkowski functions are used when calculating the distance of the data from its neighbors. A sample KNN drawing is given in Fig. 7. Accordingly, when an additional red colored ellipse shaped data comes to the clusters consisting of squares and circulars, the distance of this data from its neighbors is measured. The data is included in the round and green data set because the distance "a" from the round and green data is smaller than the distance "b" from the square and blue data. [38,62,66,67,68].

3.6. Decision tree (DT)

Decision trees are one of the frequently used machine learning

Table 6
Literature studies for UNSW-NB15 dataset.

Reference	Classification Method	Feature Selection	Accuracy	DR	FAR	FPR	FNR
Moustafa et al. [19]	**DT**	**-**	0.8556	-	0.1578	-	-
	LR		0.8315	-	0.1848	-	-
	NB		0.8207	-	0.1856	-	-
	ANN		0.8137	-	0.2113	-	-
	EM Clustering		0.7847	-	0.2379	-	-
Patil et al. [12]	**RF**	BBA+FSFF	0.9909	-	-	0.63	-
		BBA+CAFF	0.9927	-	-	0.51	-
		BBA+FU	0.9943	-	-	0.39	-
Khammassi et al. [15]	**C4.5, NB, Tree RF**	NSGA2-BLR	0.9490	-	-	-	-
		NSGA2- MLR	0.6600	-	-	-	-
Moustafa et al. [20]	**k-Means**	GAA-ADS	0.776	0.754	-	0.082	-
			0.86	0.852	-	0.063	-
			0.882	0.871	-	0.061	-
			0.927	0.912	-	0.059	-
			0.928	0.913	-	0.051	-
Zhang et al. [21]	**MSCNN-LSTM**	-	0.898	-	0.474	-	0.086
Gottwalt et al. [22]	**-**	CorrCorr	0.9865	0.9974	-	0.012	-
Khammassi et al. [23]	**C4.5**	GA-LR	0.8149	-	-	-	-
	RF		0.1408	-	-	-	-
	NBTree		0.821252	-	-	-	-
Kasongo et al. [24]	**WFEU**	Binary Classification FFDNN	0.8710	-	-	-	-
		Multiclass Classification FFDNN	0.7716	-	-	-	-
Hajisalem et al. [2]	**FCM, CFS**	ABC, AFS	0.95	0.88	-	0.021	-
			0.968	0.891	-	0.0089	-
			0.974	0.928	-	0.0063	-
			0.983	0.973	-	0.0032	-
			0.989	0.980		0.0013	
Proposed Method	**Fine Tree**	**-**	**0.9984**	-	-	-	-
	Medium Tree		**0.9984**	-	-	-	-
	Cubic Tree		0.9448	-	-	-	-
	Fine KNN		**0.9546**	-	-	-	-
	Medium KNN		0.9479	-	-	-	-
	Cubic KNN		0.9448	-	-	-	-
	Linear SVM		0.9937	-	-	-	-
	Quadratic SVM		**0.9962**	-	-	-	-
	Cubic SVM		0.9955	-	-	-	-

Table 7
Literature studies for NSL-KDD dataset.

Reference	Classification Method	Feature Selection	Accuracy	DR	FAR	FPR
Khammassi et al. [15]	**C4.5, NB, Tree RF**	NSGA2-BLR	0.9965	-	-	-
		NSGA2- MLR	0.9899	-	-	-
Moustafa et al. [20]	**k-Means**	GAA-ADS	0.95	0.942	-	0.011
			0.953	0.951	-	0.0007
			0.977	0.964	-	0.0002
			0.988	0.987	-	0.0002
			0.997	0.996	-	0.0002
Hajisalem et al. [2]	**FCM, CFS**	ABC, AFS	0.967	0.902	-	0.0082
			0.97	0.923	-	0.006
			0.973	0.951	-	0.0034
			0.976	0.986	-	0.0015
			0.99	0.99		0.0011
Rashid et al. [25]	**SVM**	Hybrid (CFS, IGR, Gini İndex)	1.00	-	-	-
	KNN		0.9980	-	-	-
	NB		0.9860	-	-	-
	NN		0.9990	-	-	-
	DNN		0.9990	-	-	-
	Auto Encoder		0.9860	-	-	-
Proposed Method	**Fine Tree**	**-**	**0.9992**	-	-	-
	Medium Tree		0.9991	-	-	-
	Cubic Tree		0.9909	-	-	-
	Fine KNN		**0.9964**	-	-	-
	Medium KNN		0.9915	-	-	-
	Cubic KNN		0.9909	-	-	-
	Linear SVM		0.9847	-	-	-
	Quadratic SVM		0.9932	-	-	-
	Cubic SVM		**0.9946**	-	-	-

Table 8
Literature studies for CIDDS-001 dataset.

Reference	Classification Method	Feature Selection	Accuracy	Precision	Recall	F1-Score
Verma et al. [29]	**KNN**	-	0.995	-	-	-
			0.996	-	-	-
			0.994	-	-	-
			0.995	-	-	-
			0.993	-	-	-
			0.995	-	-	-
Verma et al. [30]	**kNN**	-	0.995	-	-	-
	SVM		0.953	-	-	-
	DT		0.999	-	-	-
	RF		0.999	-	-	-
	NB		0.871	-	-	-
	DL		0.941	-	-	-
	ANN		0.638	-	-	-
	EM		0.459	-	-	-
	SOMs		0.384	-	-	-
	k-means		0.381	-	-	-
Proposed Method	**Fine Tree**	-	**0.9966**	**0.9957**	**0.9957**	**0.9957**
	Medium Tree		0.9914	0.9894	0.9894	0.9893
	Cubic Tree		0.9611	0.9534	0.9534	0.9534
	Fine KNN		**0.9781**	**0.9732**	**0.9732**	**0.9731**
	Medium KNN		0.9631	0.9553	0.9553	0.9554
	Cubic KNN		0.9611	0.9534	0.9534	0.9534
	Linear SVM		0.9284	0.9172	0.9172	0.9147
	Quadratic SVM		0.9678	0.9648	0.9648	0.9631
	Cubic SVM		**0.9729**	**0.9682**	**0.9682**	**0.9679**

Table 9
Literature studies for ISCX2012 dataset.

Reference	Classification Method	Feature Selection	Accuracy	Precision	Recall	F1-Score	DR	FPR
Awad et al. [31]	**WELM**	-	0.9910	-	-	-	-	-
Tan et al. [36]	**EMD**	-	0.9012	-	-	-	0.9004	0.0792
Bouteraa et al. [32]	**OCSVM**	-	0.7630	-	-	-	0.106	0.0991
	ELM		0.9287	-	-	-	0.8284	0.0504
	RPART		0.9807	-	-	-	0.9582	0.0146
	SVM		0.9953	-	-	-	0.9936	0.0043
	RF		0.9985	-	-	-	0.9967	0.0012
Injadat et al. [33]	**SVM**	**BO optimization**	0.9984	0.9998	1.00	-	-	0.003
	k-NN		0.9993	0.9999	1.00	-	-	0.001
	RF		0.9992	0.9999	1.00	-	-	0.001
Yassin et al. [34]	**KMC + NBC**	-	0.99	-	-	-	0.988	0.022
Proposed Method	**Fine Tree**	-	**1.00**	**0.9994**	**0.9994**	**0.9994**	-	-
	Fine KNN		**0.998**	**0.9978**	**0.9978**	**0.9979**	-	-
	Medium KNN		0.9964	0.9959	0.9959	0.9963	-	-
	Cubic KNN		0.996	0.9954	0.9954	0.9959	-	-
	Linear SVM		0.9849	0.9819	0.9819	0.9845	-	-
	Quadratic SVM		**0.998**	**0.9978**	**0.9978**	**0.9979**	-	-
	Cubic SVM		0.9849	0.9819	0.9819	0.9845	-	-

Table 10
Computational complexity calculation of the proposed method.

Phase	Big-O notation
Min-Max normalization	$O(L)$
Classification	$O\ (10k)$
Total	$O(\ L + 10k)$

methods in classification due to their speed and efficiency. Classification with decision trees is carried out in two steps. In the first step, the tree is created. In the second step, classification rules are obtained from this tree. Decision trees mainly consist of roots, branches and leaves. The decision making process is from root to leaf. In the decision process, branches are followed [38,69]. A simple decision tree structure is given in Fig. 8.

4. Proposed method

IDS systems must be constantly updated to prevent attacks that develop day by day. In this context, anomaly based network intrusion detection systems have become a technology frequently used in the attack detection stage. In this study, CSE-CIC-IDS-2018, UNSW-NB15, NSL-KDD, CIDDS-001 and ISCX2012 data sets that can simulate current network attack types were studied. The following steps have been implemented in the proposed method. In addition, a flow diagram summarizing these steps is given in Fig. 9.

- Step 1: A certain amount of sample data was taken from each data set. Empty lines in some data sets have been removed.
- Step 2: The data sets obtained are normalized with the min-max normalization.
- Step 3: The data sets are classified with SVM, KNN and DT machine learning methods.
- Step 4: The classifier results are compared with the literature. At this stage, it was observed that the classifier performance of CSE-CIC-IDS-2018, NSL-KDD, CIDDS-001 and ISCX2012 data sets were high and similar to the literature, and the performance rates of the UNSW-NB15 data set were not at the desired level both in the literature and in this study. As can be seen in Table 4, the classification result of

this data set in its raw form is in the range of 80-85%, and the raw classifier result of this data set has similar results with the studies conducted with the same classifier in the literature.

- Step 5: At the last stage, it was seen that some classes are related to each other in the UNSW-NB15 data set. Accordingly, Backdoors, Exploits and Shellcode attacks are combined into one class as Exploits attack [70]. In the next step, the Fuzzers and DoS classes have just been merged under the DoS class. [18]. In addition, when the UNSW-NB15 data set was classified in matlab environment with all its classes, it was seen that the Fuzzers and DoS classes were similar. In this sense, UNSW-NB15 dataset was created from Analysis, DoS, Exploits, Generic, Reconnaissance, Normal and Worms classes.

The proposed method was tested with Matlab program installed on a Windows Server 2012 R2 virtual machine with 32 Gb RAM, 4 core CPU. Each set of data is classified with SVM, KNN and DT classifiers in Matlab environment. 10 K cross validation was used for each classifier and 100 iterations were run in each data set. In addition, only Accuracy value was not taken in the classifier outputs like most studies. "Accuracy", "Precision", "recall", "F-measure" and "Geometric Mean" values were also taken in the classifier outputs.

Since there is a large amount of data in the datasets, a certain amount of data was received for each class in each dataset. The data amounts for each class from the datasets are given in Table 3.

5. Experimental results

In the study, Accuracy, Precision, Geometric mean and F-measure parameters were calculated by running 100 iterations of SVM, DT and KNN classifiers. These parameters are given in Eqs. 3, 4, 5 and 6 respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$Geometric\_mean = \sqrt{\frac{TP \cdot TN}{(TP + FN) \cdot (TN + FP)}} \tag{5}$$

$$F - Measure = \frac{2TP}{2TP + FP + FN} \tag{6}$$

Different algorithms are used for each SVM, DT and KNN classifier used for classification of data. SVM Linear, SVM Quadratic, SVM Cubic algorithms were used for SVM classifier, KNN Fine, KNN Medium and KNN Cubic algorithms were used for the KNN classifier, and Tree Fine and Tree Medium algorithms were used for the DT classifier. Results are given in Table 4.

When Table 4 is examined, for the CSE-CIC-IDS 2018 data set, the best SVM result is the SVM Quadratic algorithm with 99.81%, the best KNN result is the KNN Fine algorithm with 99.18% and the best DT result is the Fine tree algorithm with 99.92%. For this data set, DT classifier provided the best result with 99.92%. For the NSL-KDD dataset, the best SVM result belongs to the SVM Cubic algorithm with 99.46%, while the best KNN result is the KNN Fine algorithm with 99.64% and the best DT result is the Fine Tree algorithm with 99.92%. DT classifier gives the best result with 99.92% for NSL-KDD data set. For the ISCX 2012 data set, the best SVM result is the SVM Quadratic algorithm with 99.8% accuracy, the best KNN result is the KNN Fine algorithm with 99.8% and the best DT result is the Fine Tree algorithm with 100% performance. For the ISCX 2012 data set, the DT classifier reached 100% accuracy. For the CIDDS-001 data set, DT classifier provided the best result with 99.66% accuracy. For the categorized version of the UNSW-NB15 data set, the best SVM result was achieved by the SVM Quadratic algorithm with 99.62%, the best KNN result with the KNN Fine algorithm with 95.46% and the best DT result with the Fine Tree algorithm with 99.84%. For the uncategorized version of the UNSW-NB15 data set, the best result was obtained by DT method with 80.57%. Since the success rates of SVM and KNN methods are lower than DT, the algorithms that give the highest success in these classifiers are included in the table.

6. Discussion

In this section, the performance of machine learning methods applied in the study is compared with the studies in the literature.

In Table 5, the studies performed on the CSE-CIC-IDS2018 data set and the performance of the proposed method are compared. According to the table, 99.81%, 99.18% and 99.92% success rates were obtained

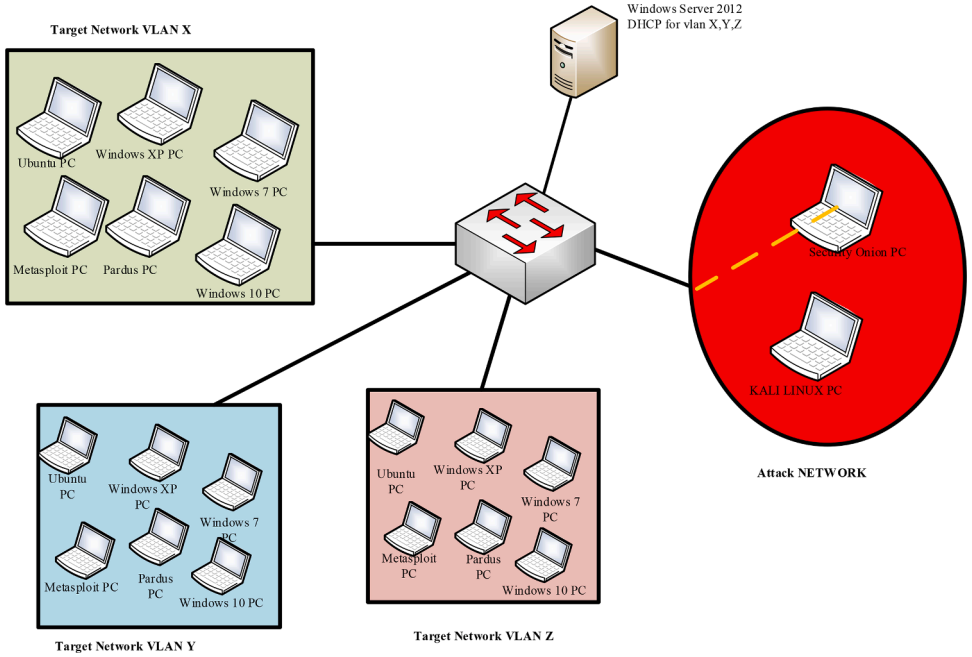


Fig. 10. planned network topology.

for the CSE-CIC-IDS 2018 data set, and accuracy rates were similar to the literature.

In Table 6, literature studies conducted on the UNSW-NB15 data set and the proposed method are compared. As mentioned in the proposed method, the classes that are related to each other in this data set are gathered in the same category. In Table 4, the categorized version and original version of this data set are classified in the same classifiers and the results are given. As can be seen in Table 4, categorizing the classes increased the performance rate in the UNSW-NB15 data set. In addition, as can be seen in Table 6, the categorization of the classes has provided the highest performance rate in the literature for this data set.

In Table 7, literature studies conducted on the NSL-KDD data set and the proposed method are compared. For NSL-KDD data set, 99.46%, 99.64% and 99.92% accuracy rates were obtained and there are similar rates in the literature.

Literature studies and recommended method on CIDDS-001 data set are compared in Table 8. High performance rates were obtained for this data set. The accuracy values of 97.29%, 97.81% and 99.66% obtained from this data set are similar to the literature.

In Table 9, the success rates of the literature studies for the ISCX2012 data set were compared with the proposed method. According to the table, the best accuracy value in the proposed method belongs to DT algorithm with 100%. In addition, the results obtained through this data set are similar to the literature.

Literature studies with intrusion detection data sets are summarized in Table 5 to Table 9. The bottom row of all tables summarizes the results obtained for each data set used in this study. According to the Table 5, 99.81%, 99.18% and 99.92% success rates were obtained for the CSE-CIC-IDS 2018 data set and similar accuracy rates with the literature. For NSL-KDD data set, 99.46%, 99.64% and 99.92% accuracy rates were obtained and there are similar rates in the literature. 99.8%, 99.8% and 100% accuracy rates were obtained for the ISCX 2012 data set. Since there are two classes in this data set, it is thought that 100% success has been achieved. For CIDDS-001 data set, 97.29%, 97.81% and 99.66% accuracy rates were obtained and similar rates with the literature. The CIDDS-001 data set used in the study was taken over the External server. There is no categorization process for classes in CSE-CIC-IDS 2018, NSL-KDD, ISCX 2012 and CIDDS-001 datasets. The related classes from the classes in the UNSW-NB15 data set have been combined. As a result, 99.62%, 95.46% and 99.84% success rates were achieved for the UNSW-NB15 data set. These accuracy rates for the UNSW-NB15 data set are the highest in the literature. This shows how accurate the categorization process for the UNSW-NB15 dataset has been done.

The complexity analysis of the study is shown in Table 10. The Big O notation is used to calculate the computational complexity of the proposed method.

L is the lenght of the dataset given in Table 10 and k shows the complexity of the classifier used. Big-O notation $10k$ was chosen for the classification due to the 10-fold cross valitadion. The kNN, DT and SVM classifiers were used for the study. Average complexity value is $O(nd)$ for kNN, $O(NKd)\,for\,DT$ and $O(n^3)$ for SVM. Among the abbreviations given here, n is the number of samples used for training, d is the dimension, and K is the number of features. The dimensions of the datasets are 12670×42 for the UNSW-NB15 dataset, 17200×80 for the CSE-CIC-IDS2018 dataset, 5032×7 for the ISCX-2012 dataset, 4070×11 for the CIDDS-0001 dataset, 22561×40 for the NSL-KDD dataset. The first value given from the dataset dimensions represents the record count and the second value represents the feature count.

7. Conclusion

With the development of smart technologies, the internet is used in every area of daily life. With the widespread use of the Internet, the types of attacks developing day by day. In order to prevent attacks, these attacks must be detected first. In this sense, IDS systems have been developed to detect attack traffic and IDS data sets have been created to simulate attack types. In this study, CSE-CIC IDS-2018, UNSW-NB15, ISCX-2012, NSL-KDD and CIDDS-001 data sets, which are frequently used for intrusion detection, were used. As a result of the study, it was seen that the classifiers used for all data sets obtained similar or higher performance with the literature. In addition, in terms of classifier performance, it has been observed that the DT classifier is more successful than the other classifiers used. DT's success rates in the range of 99% -100% for CSE-CIC IDS-2018, ISCX-2012, NSL-KDD and CIDDS-001 data sets are similar to the literature. In the study, a categorization process was made for the UNSW-NB15 data set. As a result, the performance rates achieved for the UNSW-NB15 dataset in all classifiers are ahead of studies in the literature. In this sense, the categorization process for the UNSW-NB15 dataset shows that it was done correctly.

8. Future works

In future studies, it is aimed to create a new data set by establishing a network topology as in Fig. 10. As can be seen from the network topology, different vLAN networks will be configured in order to reflect a real network topology. Computers in each vLAN network will automatically obtain IP from the DHCP server. Computers with different operating systems will be used in each vLAN. Attack scenarios will be prepared with Kali Linux computer in the attack network. The traffic generated by the Kali Linux computer will be monitored using the Kibana, Squert and Squil tools on the Security Onion computer. In addition, a more comprehensive data set will be obtained by using network attacks such as Mac flooding, dhcp snooping and arp spoofing, which are not used in most data sets. With the software to be written in Python language, necessary features will be taken for each attack scenario. After the features are taken, the data will be classified with deep learning and machine learning methods. To measure the performance of the data set created at the last stage, it will be compared with the existing data sets. Thus, in this field, it is aimed to provide a current and comprehensive data set to the literature.

When the existing data sets are examined, the types of attacks frequently used in local networks such as Mac flooding, dhcp snooping and arp spoofing are not included. This is one of the shortcomings of the current IDS data sets.

The planned future study is aimed to contain new types of attacks in addition to the attack methods used in many data sets. However, the fact that these attack methods have not been tested yet creates an uncertain situation at the point of extracting the property data of these attacks. The mentioned situation is seen as a disadvantage of the future works goal.

Declaration of Competing Interest

There is no 'Conflict of Interest' in the publication of the manuscript "Machine Learning Methods for Cyber Security Intrusion Detection: Datasets and Comparative Study".

References

[1] Digital 2019: global digital overview, (2019). https://datareportal.com/reports/digital-2019-global-digital-overview.
[2] V. Hajisalem, S. Babaie, A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection, Comput. Netw. 136 (2018) 37–50, https://doi.org/10.1016/j.comnet.2018.02.028.
[3] Z. Inayat, A. Gani, N.B. Anuar, M.K. Khan, S. Anwar, Intrusion response systems: foundations, design, and challenges, J. Netw. Comput. Appl. 62 (2016) 53–74, https://doi.org/10.1016/j.jnca.2015.12.006.
[4] A.S. Ashoor, S. Gore, Difference between intrusion detection system (IDS) and intrusion prevention system (IPS), Commun. Comput. Inf. Sci. (2011) 497–501, https://doi.org/10.1007/978-3-642-22540-6_48.
[5] J. Jabez, B. Muthukumar, Intrusion detection system (ids): anomaly detection using outlier detection approach, Procedia Comput. Sci. 48 (2015) 338–346, https://doi.org/10.1016/j.procs.2015.04.191.
[6] I. Quepons, Vulnerability and trust, PhaenEx 13 (2020) 1–10, https://doi.org/10.22329/p.v13i2.6220.

[7] C.M. Research, Cyber security market by component, security type, deployment, organization and application - global industry analysis and forecast to 2022, (2017) 202. https://www.crystalmarketresearch.com/report/cyber-security-market.

[8] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, E. Vázquez, Anomaly-based network intrusion detection: techniques, systems and challenges, Comput. Secur. 28 (2009) 18–28, https://doi.org/10.1016/j.cose.2008.08.003.

[9] S. Chakrabarti, M. Chakraborty, I. Mukhopadhyay, Study of snort-based IDS, in: ICWET 2010 - Int. Conf. Work. Emerg. Trends Technol. 2010, Conf. Proc., 2010, pp. 43–47, https://doi.org/10.1145/1741906.1741914.

[10] M.A. Ferrag, L. Maglaras, S. Moschoyiannis, H. Janicke, Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study, J. Inf. Secur. Appl. (2020) 50, https://doi.org/10.1016/j.jisa.2019.102419.

[11] I. Sharafaldin, A.H. Lashkari, A.A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization, in: ICISSP 2018 - Proc. 4th Int. Conf. Inf. Syst. Secur. Priv., 2018, pp. 108–116, https://doi.org/10.5220/0006639801080116.

[12] R. Patil, H. Dudeja, C. Modi, Designing an efficient security framework for detecting intrusions in virtual network of cloud computing, Comput. Secur. 85 (2019) 402–422, https://doi.org/10.1016/j.cose.2019.05.016.

[13] J. Kim, Y. Shin, E. Choi, An intrusion detection model based on a convolutional neural network, J. Multimed. Inf. Syst. 6 (2019) 165–172, https://doi.org/10.33851/jmis.2019.6.4.165.

[14] V. Kanimozhi, D.T.P. Jacob, Calibration of various optimized machine learning classifiers in network intrusion detection system on the realistic cyber dataset Cse-Cic-Ids2018 using cloud computing, Int. J. Eng. Appl. Sci. Technol. 04 (2019) 209–213, https://doi.org/10.33564/ijeast.2019.v04i06.036.

[15] C. Khammassi, S. Krichen, A NSGA2-LR wrapper approach for feature selection in network intrusion detection, Comput. Netw. 172 (2020), 107183, https://doi.org/10.1016/j.comnet.2020.107183.

[16] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, A. Hotho, A survey of network-based intrusion detection data sets, Comput. Secur. 86 (2019) 147–167, https://doi.org/10.1016/j.cose.2019.06.005.

[17] V. Kanimozhi, T.P. Jacob, Artificial intelligence based network intrusion detection with hyper-parameter optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing, ICT Express 5 (2019) 211–214, https://doi.org/10.1016/j.icte.2019.03.003.

[18] N. Moustafa, J. Slay, UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set), in: 2015 Mil. Commun. Inf. Syst. Conf. MilCIS 2015 - Proc, 2015, https://doi.org/10.1109/MilCIS.2015.7348942.

[19] N. Moustafa, J. Slay, The evaluation of network anomaly detection systems: statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set, Inf. Secur. J. 25 (2016) 18–31, https://doi.org/10.1080/19393555.2015.1125974.

[20] N. Moustafa, J. Slay, G. Creech, Novel geometric area analysis technique for anomaly detection using trapezoidal area estimation on large-scale networks, IEEE Trans. Big Data 5 (2019) 481–494, https://doi.org/10.1109/TBDATA.2017.2715166.

[21] J. Zhang, Y. Ling, X. Fu, X. Yang, G. Xiong, R. Zhang, Model of the intrusion detection system based on the integration of spatial-temporal features, Comput. Secur. 89 (2020), 101681, https://doi.org/10.1016/j.cose.2019.101681.

[22] F. Gottwalt, E. Chang, T. Dillon, CorrCorr: a feature selection method for multivariate correlation network anomaly detection techniques, Comput. Secur. 83 (2019) 234–245, https://doi.org/10.1016/j.cose.2019.02.008.

[23] C. Khammassi, S. Krichen, A GA-LR wrapper approach for feature selection in network intrusion detection, Comput. Secur. 70 (2017) 255–277, https://doi.org/10.1016/j.cose.2017.06.005.

[24] S.M. Kasongo, Y. Sun, A deep learning method with wrapper based feature extraction for wireless intrusion detection system, Comput. Secur. (2020) 92, https://doi.org/10.1016/j.cose.2020.101752.

[25] A. Rashid, M.J. Siddique, S.M. Ahmed, Machine and deep learning based comparative analysis using hybrid approaches for intrusion detection system, (2020).

[26] J. David, C. Thomas, Efficient DDoS flood attack detection using dynamic thresholding on flow-based network traffic, Comput. Secur. 82 (2019) 284–295, https://doi.org/10.1016/j.cose.2019.01.002.

[27] F. Ertam, I.F. Kilinçer, O. Yaman, Intrusion detection in computer networks via machine learning algorithms, in: IDAP 2017 - Int. Artif. Intell. Data Process. Symp., 2017, https://doi.org/10.1109/IDAP.2017.8090165.

[28] C. Kolias, G. Kambourakis, A. Stavrou, S. Gritzalis, Intrusion detection in 802.11 networks: empirical evaluation of threats and a public dataset, IEEE Commun. Surv. Tutor. 18 (2016) 184–208, https://doi.org/10.1109/COMST.2015.2402161.

[29] A. Verma, V. Ranga, Statistical analysis of CIDDS-001 dataset for network intrusion detection systems using distance-based machine learning, Procedia Comput. Sci. (2018) 709–716, https://doi.org/10.1016/j.procs.2017.12.091.

[30] A. Verma, V. Ranga, On evaluation of network intrusion detection systems: statistical analysis of CIDDS-001 dataset using machine learning techniques, Pertanika J. Sci. Technol. 26 (2018) 1307–1332.

[31] M. Awad, A. Alabdallah, Addressing imbalanced classes problem of intrusion detection system using weighted extreme learning machine, Int. J. Comput. Netw. Commun. (2019), https://doi.org/10.5121/ijcnc.2019.11503.

[32] I. Bouteraa, M. Derdour, A. Ahmim, Intrusion detection using data mining: a contemporary comparative study, in: Proc. - PAIS 2018 Int. Conf. Pattern Anal. Intell. Syst., 2018, https://doi.org/10.1109/PAIS.2018.8598494.

[33] M. Injadat, F. Salo, A.B. Nassif, A. Essex, A. Shami, Bayesian optimization with machine learning algorithms towards anomaly detection, in: 2018 IEEE Glob. Commun. Conf. GLOBECOM 2018 - Proc., 2018, https://doi.org/10.1109/GLOCOM.2018.8647714.

[34] W. Yassin, N.I. Udzir, Z. Muda, Anomaly-based intrusion detection through K-means clustering and Naive Bayes classification, in: Proc. 4th Int. Conf. Comput. Informatics, ICOCI, 2013, p. 2013.

[35] M.M. Hassan, A. Gumaei, A. Alsanad, M. Alrubaian, G. Fortino, A hybrid deep learning model for efficient intrusion detection in big data environment, Inf. Sci. (Ny). (2020), https://doi.org/10.1016/j.ins.2019.10.069.

[36] Z. Tan, A. Jamdagni, X. He, P. Nanda, R.P. Liu, J. Hu, Detection of denial-of-service attacks based on computer vision techniques, IEEE Trans. Comput. (2015), https://doi.org/10.1109/TC.2014.2375218.

[37] G. Maciá-Fernández, J. Camacho, R. Magán-Carrión, P. García-Teodoro, R. Therón, UGR'16: a new dataset for the evaluation of cyclostationarity-based network IDSs, Comput. Secur. 73 (2018) 411–424, https://doi.org/10.1016/j.cose.2017.11.004.

[38] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, C. Wang, Machine learning and deep learning methods for cybersecurity, IEEE Access. 6 (2018) 35365–35381, https://doi.org/10.1109/ACCESS.2018.2836950.

[39] A deeper dive into the NSL-KDD data set, (n.d.). https://towardsdatascience.com/a-deeper-dive-into-the-nsl-kdd-data-set-15c753364657.

[40] M. Tavallaee, E. Bagheri, W. Lu, A.A. Ghorbani, A detailed analysis of the KDD CUP 99 data set, in: IEEE Symp. Comput. Intell. Secur. Def. Appl. CISDA, 2009, p. 2009, https://doi.org/10.1109/CISDA.2009.5356528.

[41] R.P. Lippmann, D.J. Fried, I. Graf, J.W. Haines, K.R. Kendall, D. McClung, D. Weber, S.E. Webster, D. Wyschogrod, R.K. Cunningham, M.A. Zissman, Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation, in: Proc. - DARPA Inf. Surviv. Conf. Expo. DISCEX 2000, 2000, pp. 12–26, https://doi.org/10.1109/DISCEX.2000.821506.

[42] J. Song, H. Takakura, Y. Okabe, Description of Kyoto University Benchmark Data, (2010) 10–12. http://www.takakura.com/Kyoto_data/BenchmarkData-Description-v5.pdf.

[43] U.S.K.P.M. Thanthrige, J. Samarabandu, X. Wang, Machine learning techniques for intrusion detection on public dataset, in: Can. Conf. Electr. Comput. Eng., 2016, https://doi.org/10.1109/CCECE.2016.7726677.

[44] M. Ring, S. Wunderlich, M. Ring, Technical report CIDDS-001 data set, 16 (2017) 361–369. https://www.hs-coburg.de/fileadmin/hscoburg/Forschung/WISENT_cidds_Technical_Report.pdf.

[45] M. Ring, S. Wunderlich, D. Grüdl, D. Landes, A. Hotho, Flow-based benchmark data sets for intrusion detection, in: Eur. Conf. Inf. Warf. Secur. ECCWS, 2017, pp. 361–369.

[46] KDD Cup 1999 Data, (n.d.). http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[47] A. Shiravi, H. Shiravi, M. Tavallaee, A.A. Ghorbani, 31, Toward developing a systematic approach to generate benchmark datasets for intrusion detection, Comput. Secur. (2012) 357–374, https://doi.org/10.1016/j.cose.2011.12.012.

[48] Intrusion Detection Evaluation Dataset (CICIDS2017), (n.d.). https://www.unb.ca/cic/datasets/ids-2017.html.

[49] CSE-CIC-IDS2018 on AWS, (n.d.). https://www.unb.ca/cic/datasets/ids-2018.html.

[50] S. Wang, Y. Gong, G. Chen, Q. Sun, F. Yang, Service vulnerability scanning based on service-oriented architecture in web service environments, J. Syst. Archit. (2013), https://doi.org/10.1016/j.sysarc.2013.01.002.

[51] A. Tantawy, S. Abdelwahed, A. Erradi, K. Shaban, Model-based risk assessment for cyber physical systems security, Comput. Secur. (2020), https://doi.org/10.1016/j.cose.2020.101864.

[52] C.L. Zhang, G.H. Yang, A.Y. Lu, Resilient observer-based control for cyber-physical systems under denial-of-service attacks, Inf. Sci. (Ny). 545 (2021) 102–117, https://doi.org/10.1016/j.ins.2020.07.070.

[53] K.B. Virupakshar, M. Asundi, K. Channal, P. Shettar, S. Patil, D.G. Narayan, Distributed denial of service (DDoS) attacks detection system for openstack-based private cloud, Procedia Comput. Sci. (2020), https://doi.org/10.1016/j.procs.2020.03.282.

[54] L. Barki, A. Shidling, N. Meti, D.G. Narayan, M.M. Mulla, Detection of distributed denial of service attacks in software defined networks, in: 2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI, 2016, p. 2016, https://doi.org/10.1109/ICACCI.2016.7732445.

[55] M.M. Najafabadi, T.M. Khoshgoftaar, C. Kemp, N. Seliya, R. Zuech, Machine learning for detecting brute force attacks at the network level, in: Proc. - IEEE 14th Int. Conf. Bioinforma. Bioeng. BIBE, 2014, p. 2014, https://doi.org/10.1109/BIBE.2014.73.

[56] D. Stiawan, M.Y. Idris, R.F. Malik, S. Nurmaini, N. Alsharif, R. Budiarto, Investigating brute force attack patterns in IoT network, J. Electr. Comput. Eng. (2019), https://doi.org/10.1155/2019/4568368.

[57] Exploit-db, anatomy of exploit - world of shellcode, (n.d.) 35646. https://www.exploit-db.com/papers/35646 (accessed September 22, 2020).

[58] U.K. Singh, C. Joshi, D. Kanellopoulos, A framework for zero-day vulnerabilities detection and prioritization, J. Inf. Secur. Appl. (2019), https://doi.org/10.1016/j.jisa.2019.03.011.

[59] S.W. Boyd, A.D. Keromytis, SQLrand: preventing SQL injection attacks, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). (2004), https://doi.org/10.1007/978-3-540-24852-1_21.

[60] W.G.J. Halfond, J. Viegas, A. Orso, A classification of SQL injection attacks and countermeasures, 2008.

[61] S. Mukkamala, G. Janoski, A. Sung, Intrusion detection using neural networks and support vector machines, Proc. Int. Jt. Conf. Neural Networks (2002), https://doi.org/10.1109/ijcnn.2002.1007774.

[62] M.A.M.S. Omar Zakaria, N.W. Ahmad, M.A. Zaidi, A classification method for data mining using SVM-weight and Euclidean distance, Aust. J. Basic Appl. Sci. 5 (2011) 2053–2059.
[63] N.R. Sabar, X. Yi, A. Song, A Bi-objective hyper-heuristic support vector machines for big data cyber-security, IEEE Access. 6 (2018) 10421–10431, https://doi.org/10.1109/ACCESS.2018.2801792.
[64] M.I. Gürsoy, A. Subaşi, DVM ile EEG işaretlerinin siniflandirilmasinda TBA, BBA ve DAA'nin performansinin karsilastirilmasi, in: 2008 IEEE 16th Signal Process. Commun. Appl. Conf. SIU, 2008, https://doi.org/10.1109/SIU.2008.4632748.
[65] V. Jakkula, Tutorial on support vector machine (SVM), Sch. EECS, Washingt. State Univ (2011) 1–13.
[66] S.H. Bouazza, N. Hamdi, A. Zeroual, K. Auhmani, Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers, in: 2015 Intell. Syst. Comput. Vision, ISCV 2015, 2015, https://doi.org/10.1109/ISACV.2015.7106168.
[67] O. Yaman, F. Ertam, T. Tuncer, Automated Parkinson's disease recognition based on statistical pooling method using acoustic features, Med. Hypotheses. (2020) 135, https://doi.org/10.1016/j.mehy.2019.109483.
[68] A comparative study of classification techniques in data mining algorithms, Int. J. Mod. Trends Eng. Res. 4 (2017) 58–63, https://doi.org/10.21884/ijmter.2017.4211.vxayk.
[69] C.E. Brodley, M.A. Friedl, Decision tree classification of land cover from remotely sensed data, Remote Sens. Environ. 61 (1997) 399–409, https://doi.org/10.1016/S0034-4257(97)00049-7.
[70] Exploit-db, anatomy of exploit - world of shellcode, (n.d.). https://www.exploit-db.com/papers/35646.

Ilhan Firat Kilincer was born in Elazig, Turkey in 1986. He received B. S. degree in Electric Electronic engineering from Kocaeli University, Kocaeli, Turkey, in 2012. He is currently pursuing the Ph.D. degree in electrical engineering. He works as nework security expert in the Computer Center, Firat University, Elazig

Fatih Ertam received the Ph.D. degree in software engineering from the Firat University, Turkey, in 2016. He is currently an Associate Professor of digital forensics engineering, Faculty of Technology, Firat University, Elazig, Turkey. His current research interest includes artificial intelligent, deep learning, digital forensics and network security.

Abdulkadir Sengur graduated from the department of Electronics and Computer Education at Firat University in 1999. He obtained his M.S. degree from the same department and the same university in 2003. His Ph.D. degree was from the department of Electronic Engineering at Firat University in 2006. He is a professor at Firat University. His interest areas include pattern recognition, machine learning and image processing.