International Journal of Communication Networks and Information Security
ISSN: 2073-607X, 2076-0930
Volume 14 Issue 03 Year 2022 Page 16:25

A Supervised Machine Learning Based Intrusion Detection Model for Detecting Cyber-Attacks Against Computer System

Santhosh Kumar Chenniappanadar
Department of Information Technology, Sona College of Technology, Salem, Tamil Nadu, India. - 636005
sanscsk@gmail.com
Sundharamurthy Gnanamurthy
Department of Computer Science and Engineering, Kuppam engineering college, Chittoor (dist), Andhra Pradesh, India. 517425
gnanamurthyspec@gmail.com
Vinoth Kumar Sakthivelu
School of computing, Department of Computer science and engineering, Vel Tech Rangarajan Dr.Sagunthala R&D institute of science and technology, Chennai, Tamilnadu, India. 600062
profsvinoth@gmail.com
***Vishnu Kumar Kaliappan**
Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India. 641407
vishnudms@gmail.com

Article History

Received: 18 June 2022
Revised: 26 September 2022
Accepted: 12 October 2022

Abstract

Internet usage has become essential for correspondence in almost every calling in our digital age. To protect a network, an effective intrusion detection system (IDS) is vital. Intrusion Detection System is a software application to detect network intrusion using various machine learning algorithms. The function of the expert has been lessened by machine learning approaches since knowledge is taken directly from the data. The fact that it makes use of all the features of an information packet spinning in the network for intrusion detection is weakened by the employment of various methods for detecting intrusions, such as statistical models, safe system approaches, etc. Machine learning has become a fundamental innovation for cyber security. Two of the key types of attacks that plague businesses, as proposed in this paper, are Denial of Service (DOS) and Distributed Denial of Service (DDOS) attacks. One of the most disastrous attacks on the Internet of Things (IOT) is a denial of service. Two diverse Machine Learning techniques are proposed in this research work, mainly Supervised learning. To achieve this goal, the paper represents a regression algorithm, which is usually used in data science and machine learning to forecast the future. An innovative approach to detecting is by using the Machine Learning algorithm by mining application-specific logs. Cyber security is a way of providing their customers the peace of mind they need knowing that they have secured their information and money.

KEYWORDS: Intrusion detection, Cyber security; machine learning, Internet of Things (IOT), Denial of services (DOS), Distributed denial of services (DDOS)

1. Introduction

These days, there is a lot of discussion about cyber protection and security from several cyber-attacks. The primary causes of that would be the enormous development of computer networks and the availability of following the development of the Internet of Things, related apps are now being used by individuals and organizations for either individual or commercial goals. In large-scale networks, cyber-attacks cause major physical harm and significant financial losses [1]. There are numerous industrial uses for machine learning, and these applications are indeed going to expand in the future. By 2026, the business for machine learning is projected to grow at an outstanding CAGR of 38%, reaching a value of USD 118 billion. Machine learning regression methods are a key idea with several applications. Regression-based computer vision techniques are used to predict present value. Regression is used to forecast a wide range of possible predicted values using the input data and historical data. Regression is a supervised learning algorithm used in machine learning to assist in mapping a relationship between labelling and statistics that predicts future outcomes.

The specific objective of attack detection is determined by intrusion detection. A computer system or network's processes are observed by intrusion detection, which examines them to detect for deviations or other abnormalities that would be against security rules. The misuse and anomaly techniques are the two types of intrusion detection. Misuse seeks to discover attack signatures in the resource under watch. Understanding normal behaviour and any deviation from it is essential to recognize an anomaly. Due to its efficiency against new attacks, anomaly detection has risen in favour. Anomaly detection can also be done using machine learning methods. For the actual detection process, machine learning algorithms are built and then used on observed data. Machine learning provides a wide range of classification methods that may be developed and applied to find network attacks. Feature reduction techniques can be used to enhance the effectiveness of these classifiers further and to reduce the detection time [2].

Institutions and societies are quite worried about IOT device protection from anomalies. All necessary actions are taken to confirm the physical security and cyber security of the IOT architecture against serious attacks. The network protection also must be extensively evaluated. Traditional intrusion detection methods are too sophisticated and resource-constrained for the Network of Things to be secure (IOT) [3].

The major contributions of the proposed approach are

1. Profiting the cyber security department by enabling the use of real data;
2. permitting for the prediction of future attacks that victims may endure;
3. enabling analysis of machine-learning algorithms to determine the optimal performance

The literature is reviewed, the most recent findings are discussed, and the gaps are explicitly stated under "Related Works." The machine-learning techniques which will be utilised in research are stated in "Materials and Methods." “Results and Discussion” offers the dataset's projections, average accuracy, and a comparison to previous research. Findings and upcoming work are provided under "Conclusions and Future Work."

2. Related Work

In the recent years, many DDoS flooding attack detection and mitigation techniques have been reported. Several methods have been given forth currently for detecting DDoS attacks. Some recent DDoS attack detection research is examined in this section.

Gao et al [5] developed a flow mining-based technique for abnormal attack identification and proved the feasibility of this method through using Disruption information from MIT Lincoln Laboratory as well as virus traces from Slammer and Code Red. It proposes deploying data mining techniques to examine the alarms received by the Distributed Intrusion Prevention and Detection System (IDS/IPS) and Deep Defence network security architecture. The author comprises three main effectiveness of the

suggested defence design; the prototype was made using a range of data mining techniques and then used to identify Cyber-attacks. It quickly detects attacks and boasts a high attack detection rate and FPR.

Gurulakshmi, K., & Nesarani et al [6] DOS attack aims to stop authorized persons from using information or services. Attackers "flood" a system with information to launch a DOS, which is the most common and obvious kind of assault. Inputting a site's URL into your computer initiates are instructing the website's secure server to display the page. In the event that an attacker overloads the server with queries, the server will not be able to process your request because it can only handle a certain variety of requests rapidly. This is frequently referred to as "denial of service" since unable to access that website. An identical attack on your email account will be initiated by an attacker using spam emails. An amateur can simply use this application to initiate attacks against the other sites or systems due to its straightforward user interface.

Sarker et al [7] among machine learning techniques, a tree-based strategy called as tree structure is among the most prominent machine learning classification methods for creating data sets. You are granted a set limit, which restrict the amount of information that can keep on your profile at any given time, whether you use a personal email supplied by your employer or one available through a free app such as Google or Microsoft.

Fischer, E. A et al [8] unlike the previous techniques, In order to address the concerns raised above, they provide a computer learning-based security method called "IntruDTree" in this study that first assesses the significance of security aspects before strengthening a tree for detecting attacks based on the chosen important components. Due to the scope of the domains under research, the estimating method for cyber-attacks and perpetrators is discussed.

Martínez Torres, J., Iglesias el.al [9] the various ways that cyber-security can show itself from a legal perspective have been described and analysed. The investigation of numerous machine learning applications in various fields will then be done in further detail in the parts that follow. Because of this, this work is regarded as a good place to start learning about this use of machine learning techniques. However, it is advised to delve much deeper into each of the topics of interest after consulting the bibliography because there is a vast amount of literature devoted to this topic, making it impossible to include all of it in this work and instead focusing on what has been deemed to be more pertinent and current.

3. Methods and Materials

When a person is a criminal victim, they turn to the detectives who expertise in a certain kind of crime. The database of this unit contains a detailed record of these data. These crimes are reported by police units based on the type, manner, day, etc. They create data based on these traits, evaluate them, and present them graphically. When several assaults are launched simultaneously against with a target, they are listed in police reports as a single strike. Although the database contains several of crimes, the focus in recent years has been on cybercrime. Cybercrime has caused significant material and moral harm, and it has yet to be ended. This section contains a mathematical description of various machine learning techniques for detecting and managing attacks [9].

1. *Establish the initial centroids.*
2. *Generate a logic matrix that represents the position of each point in a group.*
3. *Calculate and eliminate the inequality function.*
4. *Determine the new centroids.*

3.1 Selecting a Dataset

Denial-of-Service-Attacks and Distributed Denial-of –Service-Attacks are of great concern to organizations. They rank among the most dreaded threats and are quite challenging to fend against. DOS attacks can disrupt network and site services. Crime gangs can also utilize them to demand money from companies. Any deliberate attempt to isolate a system or website from its target purposes is referred to as a DOS attack. A successful Attack can disrupt a number of services, costing firm’s

money and bringing unwelcome attention to them. The primary goal of a DOS or DDOS assault is typically not to capture or expose sensitive information. The goal is to simply flood the network with data. In a distributed denial of service attack, a network connection of machines is often controlled by a single computer. A desktop machine launches a DOS assault on a victim. [10]. the bulk of DOS assaults target residential systems and are less effective than Attacks. A DOS assault, which is conducted through one machine to the victim's network, will probably overwhelm and bring down the router for the victim's local network.

A networking DoS can take many different forms, including overloaded a service with demands that appear to be genuine and delivering malformed packets, both of which are intended to make the system fail due to a defect in the system. An Attack is just a type of assault that IDS may identify. [12]. since faulty bandwidth DOS can be screened using a regulation method and is relevant to generic host-based protection, this study concentrates on the first instead of the latter type of DOS. The time of detection is a crucial component of DOS detection techniques. First before service is compromised, a good detection method should catch the DOS assault. This makes detection more difficult and raises the risk of a positive result, which is a crucial problem in DOS diagnosis. A good detection technique needs to be quick and have a low percentage of wrongful convictions.

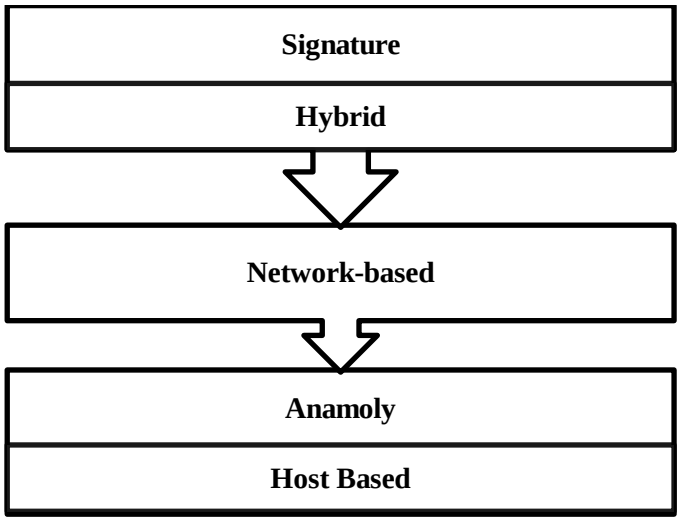

Figure 1. IDS General Classification

Similar to generic IDS, DoS detecting methods can be categorized as signature-based and unusual case. [12]. each classification work will be given a general description, and some of the related techniques will be discussed in the following sections.

3.2 Feature Extraction

The features are found and extracted after choosing a dataset. For the method, this stage is crucial. In order to be represented numerically or Boolean, the category properties of Net Flow data must be converted, which results in an excessively high matrix and memory problems. To minimize the amount of data to be analyses, we employ a strategy to select Net Flow traffic to use a window of time. A pace of two minutes and a time restriction of one minute have been established. The pre-processing of the unidirectional Net Flow dataset then extracts numerical and categorical features that characterize the dataset within the specified window of time.

3.3 Sparse Logistic Regression

To ensure that the system operates with extreme precision, sparse modelling aims to choose discriminant information for the IDS classification task while minimizing duplicate and unnecessary characteristics. Consider a prediction problem with N samples and outcomes $y1, y2, y3, .. J$. Where $j = 12,3,4, \ldots M$ and $M = 2,3,4,$ ..and M is the output numeral of the variable, let X reflect the MR input vector and Y denote the $R1$ production matrix. Class labels are used when X = +2, 2, with (+1)

denoting traditional and (1) denoting assault. A probabilistic dependent model known as regression models is described as follows:

1. $Q\left(yi = +\frac{1}{w}, xi\right) = \frac{1}{1+exp}$

The chance in the assault categorization issue is given by the values of $Q\left(yi = +\frac{1}{w}, xi\right)$. This means that the choice of subcategory would be dependent on a probabilistic assessment with a threshold to maximise expected efficacy.

2. $Q=\int \frac{+2\ r<1.5}{-2\ r>1.5}$

The parameter w's maximum likelihood estimation corresponds to reducing log-likelihood.

3. $(l)w = -\sum_{i=1}^{n} in(1 + \exp(-wxiyi)$

Normalized logistic regression offers remarkable analytic performance in a range of disciplines, including text categorization and picture classification, according to several of the earlier research in the field. We study several limitations on M in great detail. To resolve this issue, the sparseness requirement is used.

4. $w = \arg mix\ y - /Xw + ʎw$
5. $w1=\sum_{j=1}^{r} w1$

As a result, a sparse normalization for the minimizing of has been designed with the intention of extracted features.

6. $m(w) = l(w) + ʎg(w)$

A normalization parameter in this case,$m(w) = 1$ is the l2-norm regularization. Due to the ill-pawedness of the direct approach and the potential overfitting of the classification findings, the logit $m(w)$ can't be solved in this manner, the sparse regularization, which assumes ʎ=0, is a standard technique for avoiding over fitting. The Bayesian applications highest a probability estimate of l is what can be deduced from the l2 norm normalized linear regression's solution (m).

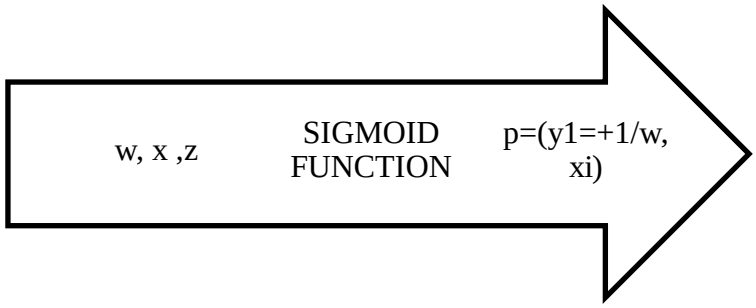

Figure 2 (A) Selection of Features for Sparse Regression Models

Feature sets are represented by the Xi. The white entries in the sparse coefficient vector w denote zero components (sparse data), and the remaining elements are chosen features. When the accelerating gradients technique and distal operator are regularly applied, the algorithm's convergence delivers the best results. The method is described in full in Method 1.

ALGORITHM 1: SPLR pseudo code

INPUT: *Sparse functions f (.) and g (.) with regularisation limitation*

Initialize*: The affine combination parameter w (0) and the step size t (0)*

***OUTPUT**: optimum results w*

- *Determine the search point s*
- *Determine the gradient descent point u* $j+1$
- *Calculate using the proximal operator* $w(j+1)$
- *Update* te^{j+1} *and* $j+1$
- *Continue the previous stages until there is no longer a significant difference between v (i) and v (i+2).*
- *Profit* $v = v(i+2)$

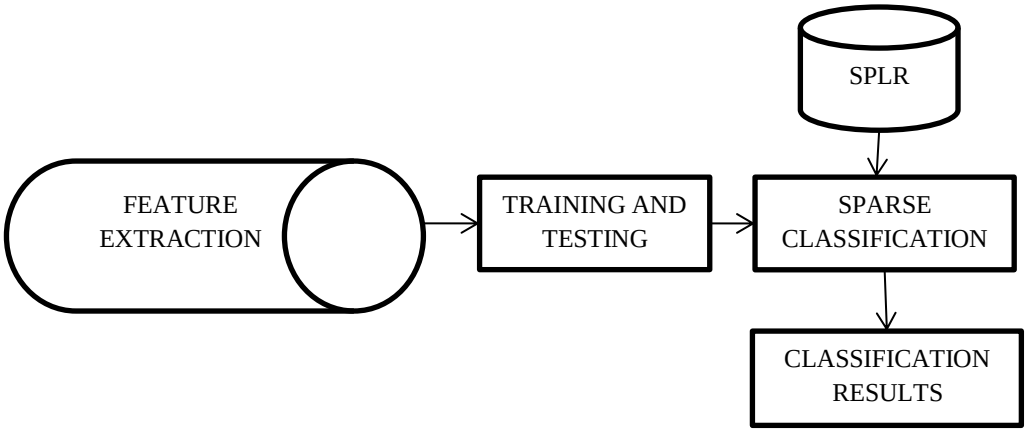

Figure 2(B). SPLR Taxonomy for the IDS

4. Experimental Results and Discussion

4.1 Set Of Data For KDD CUP 98:

KDD CUP 98 has been the most frequently used collection since 1999 for analysing abnormality detection techniques [15]. The combined training and evaluation data contain 39 assaults [14]. Additionally, it has labelled interconnections, which makes it easier to train and run the model and persuaded us to utilize it. Each contact report has 150 bits and there are about 6 per cent of them. As a result, many researchers in the field have used the dataset to create realistic network systems.

Table 1. Count of Tests in the KDD CUP 98 Databases

KDD Databases	DOS	Inquiry	Standard	Total
Entire KDD	3,883,370	41,102	972,780	4,897252
Corrected KDD	229,853	4166	60,593	294,612
10% KDD	391,458	4107	97,277	489,145

The IDS area attracted the interest of numerous researchers shortly after. Fifth classes make up the KDD '99 data, DOS and Probing assaults making up the other four. Attack probability differs between training and testing. It consequently offers the most authentic setting possible for IDS tests. Number of KDD '20 samples, the 20 attributes, and the attack categories are listed in Tables 2.

Table 2. List of Specifications for KDD

S.NO	ATTRIBUTES
1	Length

2	procedure style
3	Facility
4	Standard
5	src_bit
6	dst_bit
7	parcel
8	incorrect_fragmnet
9	crucial
10	Warm
11	n_unsuccessful_logins
12	noted in
13	n_cooperated
14	origin shell
15	su_tried
16	n_origin
17	n_file_formations
18	n_missiles
19	n_access_archives
20	n_outbound_cmds

Table 3. KDD '98 Outbreak Explanations

Occurrence	Cyber-Attack Using Dataset
DOS	Spinal, Terrestrial, shell, smurf, drop
R2L	ftp inscribe, deduction passwd, imap, multihop, phf, detective, warezclient, and warezmaster are all available
U2R	Bumper overflow, Perl, load component, rootkit
Analytical	Bumper overflow, treasure, load segment, rootkit

The KDD '98 data sets are shown in Table 4. From the initial KDD '20 dataset used in [14], they selected several random samples for our investigation.

Table 4. Dataset Descriptions

Class	Preparations	Occurrences
Regular	813,816	78.6
DOS	946,268	23.9
Probe	14,854	2.29
Total	**1,773,934**	**100%**

The detection rate is the most important parameter for evaluating IDS performance (DR). As defined in Equation, this parameter counts the proportion of assaults that were successfully identified out of all attacks. The ratio between the total amount of normal connectivity as stated in Equations and the number of regular connectivity that were incorrectly identified as attacks is used to compute the false alert rate.

To illustrate the suggested SPLR attribute selection technology's capability for categorization with increased effectiveness than KDD '20, [14] and other classifier models. However, as shown in Figures 3, in the experimental results, offer the meaningful features selection and accuracy rate via SPLR. In addition, The OCA and degree of scarcity were exhibited in Figure 4.3 at various settings; if 2 = 0.2, the unit of scarcity is great as shown in Figure 4.2, but the detection accuracy of SPLR is high at the same parameter values 0.9796.

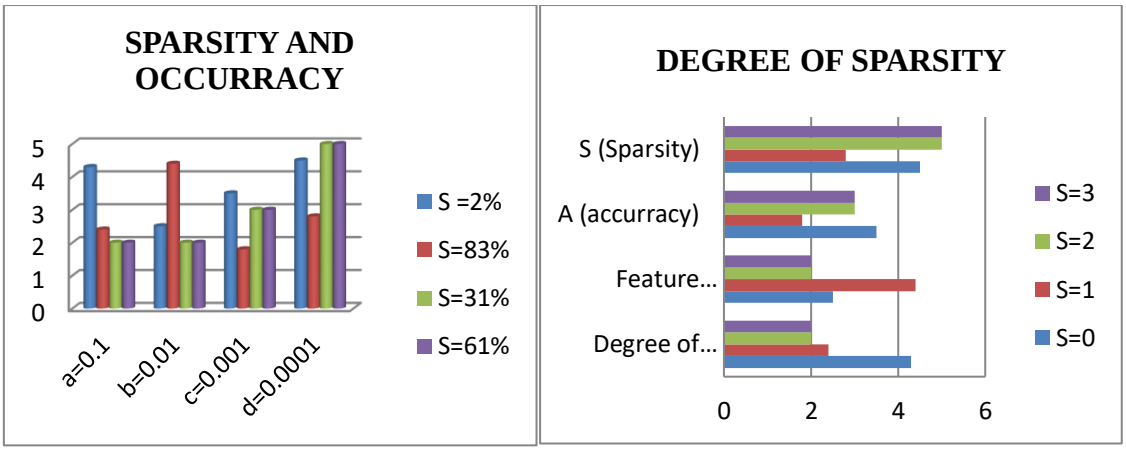

Figure 3 SPLR on KDD'20 Feature Selection and Degree of Sparsity

On the KDD '98 dataset, the SPLR-based FS outperforms VFDT [14] because the VFDT simultaneously chooses descriptive features and classification while the suggested approach simply takes into account the 20 chosen selected features for categorization. This exemplifies how well SPLR extracts valuable and rich data for the IDS. In addition, they analyze the performance of the SPLR-based feature selection and compare the plots of the chosen features with the sparsity variation. The degree of sparsity decreases as the parameter value decreases Figure 4 illustrates how the SPLR first rises before stabilising. The remaining features are sparse or eliminated, as seen in Figure 3. The attributes services, territory missiles, nm arriving comedies hot logins, dst holding srv total, and dst swarm srv amount have been selected. These traits might not be enough for the classification to identify different incursions. As a result, characteristics 18, 20, and 21 have less of an impact. In addition, we selected the final 21 traits based on the values of ʎ. Feature 4.1 is critical for creating the pattern for identifying system infiltration. In other terms, the interruptions depend on the provider.

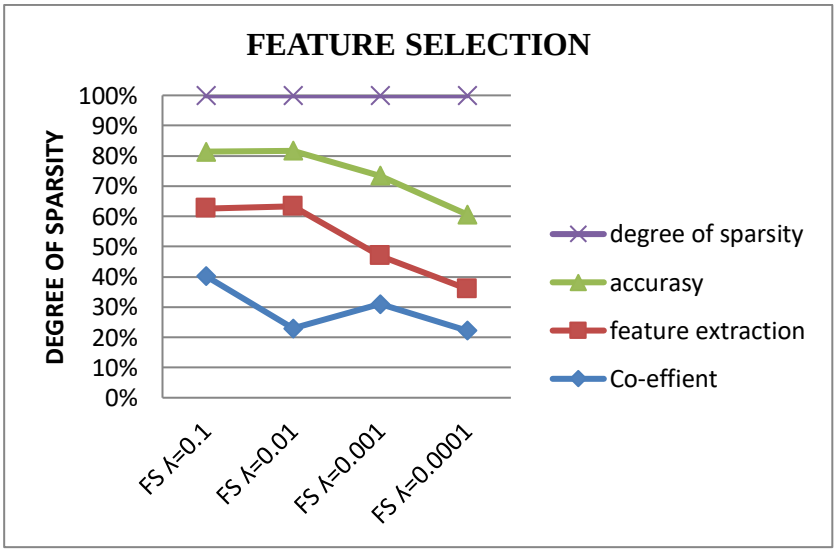

Figure 4 KDD '98 Dataset Feature Selection and Degree of Sparsity Using SPLR

In the situation put forward, classification accuracy for the KDD '99 datasets reaches or approaches the greatest values when the sparsity is around 86% and 96%, or 6% to 21% of the characteristics are chosen.

4.2 Rate of Classification Detection

The outcomes of the suggested technique are contrasted with the experimental outcomes of [14, 15]. Three factors—data quantity, detection accuracy, and median training per sample—are used in the analysis. As a result of similar classification accuracy, the findings show that the SPLR model is

significantly more efficient. Firstly, while using the KDD'98 samples, the SPLR performed better in terms of detection accuracy (98.6%) and FAR (0.39%). Although the average time complexity per sample is lower than the VFDT technique (0.00000 s), it still requires less time training (12.7 s) to develop a model. Second, although other classifiers like genetic code, multivariate adaptive regression splines, naive Bayes, and VFDT are effective approaches, the proposed (SPLR) classifier performs better in terms of extracted features, accuracy rate, predictive accuracy, practice time, and ordinary training time per sample than these other classification methods. Finally, an intriguing and comprehensive answer to IDS classifiers is offered by the SPLR's model-based classifier, which can carry out meaningful features selection throughout the classifier training stage. The testing methods are calculated for the test and training datasets during the testing stage, and the outcomes for all assaults and regularize are displayed in figure 4.3, which represents the overall categorization performance of the recommended system on the KDD cup 98 database. The overall performance of the proposed system is significantly improved by analysing the results, and for all kinds of strikes, it reaches more than 92% effectiveness.

Table 5. defining the Suggested Intrusion System's Class

Samples	**Metrics**	**Suggested Scheme**	
		Preparation	**Challenging**
ENQUIRY	Accuracy	1.912622	0.912822
	Remember	0.37183	0.37383
	F-measured	1.527359	1.52736457
	Correctness	2.906918	2.989823
	Exactness	1.993863	1.993898
	Recall	2.98144	2.984154
DOS	F-measured	1.945236	1.946936
	Correctness	2.94789	2.949869
	Exactness	1.051998	1.051998
	Remember	2.994168	2.994395
NORMAL	F-measured	1.9037653	1.904381
	Correctness	2.910852	2.903019

5. Conclusions and Future Work

With the SPLR model, discriminative feature selection was made possible in this study, which improved attack categorization for an intrusion detection system (IDS). This work's primary and most important contribution is its handling of greater datasets. Dealing with greater information is the first and most important contribution of this work. The suggested approach manages over-fitting and feature repetition by performing feature selection and classification at the same time. The suggested attribute selection approach outperforms alternative identification prototypes, according to our experimental findings. By reducing the total experimental damage and penalizing for feature variable sparseness, the sparse approach unifies the processes of selection and classification of features. The SPLR technique's running times therefore follow a linear relationship with the training samples and feature attributes. Ahead of other approaches suggested in the literature, the feature selection and prediction cost are also more effective. Future study, in our opinion, should focus on exploring current advancements in the IDS field as well as conducting additional research into the effects of various classifiers. This is because the trials we conducted so far have shown promising results.

References

[1] Alqahtani, H., Sarker, I. H., Kalim, A., Hossain, M., Md, S., Ikhlaq, S., & Hossain, S. (2020, March). Cyber intrusion detection using machine learning classification techniques. In *International conference on computing science, communication and security* (pp. 121-131). Springer, Singapore.

[2] Biswas, S. K. (2018). Intrusion detection using machine learning: A comparison study. *International Journal of pure and applied mathematics*, *118*(19), 101-114.

[3] Saba, T., Rehman, A., Sadad, T., Kolivand, H., & Bahaj, S. A. (2022). Anomaly-based intrusion detection system for IoT networks through deep learning model. *Computers and Electrical Engineering*, *99*, 107810.

[4] Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, *2*(1), 1-22.

[5] Gao, Y., Wu, H., Song, B., Jin, Y., Luo, X., & Zeng, X. (2019). A distributed network intrusion detection system for distributed denial of service attacks in vehicular ad hoc network. *IEEE Access*, *7*, 154560-154571.

[6] Gurulakshmi, K., & Nesarani, A. (2018, May). Analysis of IoT bots against DDOS attack using machine learning algorithm. In *2018 2nd International conference on trends in electronics and informatics (ICOEI)* (pp. 1052-1057). IEEE.

[7] Sarker, I. H., Abushark, Y. B., Alsolami, F., & Khan, A. I. (2020). Intrudtree: a machine learning based cyber security intrusion detection model. *Symmetry*, *12*(5), 754.

[8] Fischer, E. A. (2005, February). Creating a national framework for cybersecurity: an analysis of issues and options. Library Of Congress Washington Dc Congressional Research Service.

[9] Martínez Torres, J., Iglesias Comesaña, C., & García-Nieto, P. J. (2019). Machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*, *10*(10), 2823-2836.

[10] Nanda, I. ., Singh, M. ., & Khatua, L. . (2022). Data Security and Anonymization in Neighbourhood Attacks in Clustered Network in Internet of Things (NIoT). International Journal on Recent and Innovation Trends in Computing and Communication, 10(11), 28–32

[11] Chen, Y., Das, S., Dhar, P., El-Saddik, A., & Nayak, A. (2008). Detecting and Preventing IP-spoofed Distributed DoS Attacks. *Int. J. Netw. Secur.*, *7*(1), 69-80.

[12] Delplace *preprint arXiv:2001.06309*, A., Hermoso, S., & Anandita, K. (2020). Cyber Attack Detection thanks to Machine Learning Algorithms. *arXiv*.

[13] Alenezi, M., & Reed, M. J. (2012). Methodologies for detecting DoS/DDoS attacks against network servers. In *The Seventh International Conference on Systems and Networks Communications ICSNC* (pp. 92-98).

[14] Anjaiah, P. ., & Yadav, B. V. R. N. . (2022). IoT Enabled Smart Activity Recognition using Machine Learning Methods. International Journal on Recent and Innovation Trends in Computing and Communication, 10(10), 09–16.

[15] Shah, R. A., Qian, Y., Kumar, D., Ali, M., & Alvi, M. B. (2017). Network intrusion detection through discriminative feature selection by using sparse logistic regression. *Future Internet*, *9*(4), 81.

[16] Kshirsagar, P. R., Yadav, R. K., Patil, N. N., & Makarand L, M. (2022). Intrusion Detection System Attack Detection and Classification Model with Feed-Forward LSTM Gate in Conventional Dataset. Machine Learning Applications in Engineering Education and Management, 2(1), 20–29

[17] Shanmugavadivu, R., & Nagarajan, N. (2011). Network intrusion detection system using fuzzy logic. *Indian Journal of Computer Science and Engineering (IJCSE)*, *2*(1), 101-111.

[18] Al-mamory, S. O., & Jassim, F. S. (2015). On the designing of two grains levels network intrusion detection system. *Karbala International Journal of Modern Science*, *1*(1), 15-25.