Distributed denial of service attack defense system-based auto machine learning algorithm

Mohammad Aljanabi1,2, Russul Hayder 3, Shatha Talib 4, Ahmed Hussein Ali1,2, Mostafa Abdulghafoor Mohammed5, Tole Sutikno6

1Department of Computer, College of Education, AL-Iraqia University, Baghdad, Iraq
2Department of Computer Science, Al Salam University College, Baghdad, Iraq
3Engineer in the Ministry of Education Iraqi Directorate of Education Baghdad Karth III, Baghdad, Iraq
4Computer Science and Information System, Al-Bayan University College, Baghdad, Iraq
5Imam Aadham University College, Baghdad, Iraq
6Department of Electrical Engineering, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

ABSTRACT

The use of network-connected gadgets is rising quickly in the internet age, which is escalating the number of cyberattacks. The detection of distributed denial of service (DDoS) attacks is a tedious task that has necessitated the development of a number of models for its identification recently. Nonetheless, because of major fluctuations in subscriptions and traffic rates, it continues to be a difficult challenge. A novel automatic detection technique was created to address this issue in this work, which reduces the feature space and consequently minimizes the computational time and model overfitting. Data preprocessing is done first to increase the model’s generalizability; then, a feature selection method is used to choose the most pertinent features to increase the accuracy of the classification process. Additionally, hyperparameter tuning—choosing the proper parameters for the learning approach—improved model performance. Finally, the support vector machine (SVM) is compatible with the optimization and the hyperparameters offered by supervised learning methods. The CICDDoS2019 dataset was used to evaluate each of these assays, and the experimental findings demonstrated that, with an accuracy of 99.95%, the suggested model performs well when compared to more modern techniques.

Keywords: Auto ML, Cyber security, DDoS attack, Intrusion detection, Machine Learning, Support vector machine

This is an open access article under the CC BY-SA license.

Corresponding Author:
Mohammad Aljanabi
Department Computer Science, Al Salam University College
Baghdad, Iraq
Email: mohammad.cs88@gmail.com

1. INTRODUCTION

Since the report of the first attack incident by Computer Incident Advisory Capacity in 1999, distributed denial of service (DDoS) attacks have grown to be one of the most difficult network security problems [1]-[5]. The threat of DDoS attacks is still extremely real and growing every year [6]-[8], even though many different defense strategies have been put forth in academics and business. DDoS attacks continue to be the main threat that service providers are contending with. DDoS attacks simultaneously and continuously send a lot of traffic to the target system with the goal of preventing genuine users from accessing a certain network service [9]-[12]. Hackers frequently utilize botnets to launch a DDoS attack in such attempts. Botnets are networks made up of host computers that have been "enslaved" by one or more
attackers, known as "botmasters," in order to carry out destructive operations [13]-[16]. Due to the deployment and connection of billions of susceptible internet of things (IoT) devices as well as the ease with which the majority of IoT devices may be hacked and compromised, the most potent botnets have recently tended to rely on IoT devices [17]. The purpose of launching an attack might differ between different hackers, but there are often five basic motives for doing so, including financial gain, retaliation, ideological challenge, ideological belief, and electronic warfare. What consequences do these attacks have?

Attacks must increasingly be identified and stopped before they reach their target. The most widespread and effective attacks among the numerous types are DoS and DDoS attacks, which have a variety of origins and formats. These attacks are aimed at using up available network resources and bandwidth just to prevent genuine user access to the target network is limited. DDoS attacks often start with two steps; the first is stealth, where attackers set up their attack's launch configuration by building a network of malicious devices or a "botnet" (using DDoS tools on multiple network hosts). The second stage is to attack the target network by triggering the set or bots [18]. DDoS attacks can cost businesses up to $50,000. DDoS assaults are often divided into two categories: Volumetric attack, commonly referred to as a flood attack. This kind of attack has two goals. They first overwhelm the bandwidth of the targeted server by flooding it with traffic to exhaust its bandwidth [19]. The second step is to clear all currently cached data. Attackers frequently start by using less bandwidth by focusing on particular services or apps that have an impact on the performance of other applications. Techniques that detecting the attacks can be broadly divided into three categories [20], [21]: signature-based (abuse-based), hybrid-based, and anomaly-based. With the signature-based technique, previously known attacks are identified by matching the attack signatures [22]. For the skew-based approaches, attacks are identified by detecting patterns that differ from regular traffic or network activity [23]. These are efficient as they can identify unidentified attacks. For the hybrid techniques, they integrate strategies based on anomaly-and signature-based approaches. Several strategies have been put out in recent years to forecast different attacks using machine learning techniques [24]. The following are the primary contributions of our proposed approach.

- The suggested method integrates the oversampling (SMOTE) and under-sampling techniques (Tomek links) to balance the minority class data.
- Suggestion of a hybrid feature selection method for the extraction of the best features with the least amount of training time and with the highest detection rate.
- The support vector machine (SVM) hyperparameters are modified using grid Search to obtain the optimal hyperparameters for enhanced model performance.
- The performance of the proposed method in terms of performance metrics and computing time was evaluated by making a comparison between the existing techniques and the proposed model in the last section.

We briefly go through the newest and most popular techniques for identifying DDoS attacks in this section. Maslan et al. [25] suggested a broad machine learning (ML) approach that reduced functionality while improving DDoS attack detection performance. To determine the function and choose the subset of first 20 features, this method employs built-in function selection and filtering approaches, especially the F test, the light gradient amplification algorithm, and the random forest (RF) algorithm. The proposed model was then tested on more attacks after being trained using the records for a specific type of attack [26]-[28]. The AE-SVM model is intended to quickly identify attacks. To efficiently distinguish attacks from non-attacks, dimensions are downscaled using an automated encoder and trained with the SVM method [29], [30]. The developed model produced good accuracy despite the unbalanced data; it also recorded a excellent accuracy level using 25 functions and decreased the high rates of false positives [31]-[33].

Four sections have been created for the paper. The relevant works are described in section 1, and the proposed method and the performance indicators are discussed in section 2. The results of the experiments and the discussion are found in section 3 while the conclusion of the work is in section 4.

2. PROPOSED METHOD

Preprocessing, model modification and classification are the three phases of the suggested model. data analysis for exploratory is used during preprocessing to examine the data and understand it. After that, a mix of over- and under-sampling techniques. Data quality can also be improved through data cleaning. The next step is to use the function scale to normalize the range of functions before applying a transformation to digitize the categorical data. Similar to this, it is advised to adjust the model using the hybrid function to condense the function space and then tune the hyperparameters to enhance model performance. For various observed learning techniques, the best features and hyperparameters are provided in order to distinguish attacks during classification. In Figure 1, the suggested work's diagram is depicted, and the following parts give a thorough analysis.

Distributed denial of service attack defense system-based auto machine learning ... (Mohammad Aljanabi)
2.1. Dataset
The suggested model was evaluated CIC/DDoS_2019 dataset for performance [1]. The CIC/DDoS_2019 dataset covers more forms of DDoS attacks with high volume compared to other datasets [2], [3]. The dataset includes two different types of attacks which are thinking and exploration. Both forms of attack disguise the identity of the attacker and flood the resources of the victim with response packets by sending packets to reflexive servers using the address IP of the victim as the source IP. The dataset, which includes 88 functions, was created in two days for training and testing. There are 12 different DDoS attacks in the training set.

2.2. Pre-processing
It is a crucial phase in the development of any ML framework and is mostly used for the organization and cleaning data to make sure it is suitable for creation and training of any ML framework. The pre-processing step is very important in machine learning, applying good pre-processing process reduce the execution time and increase the accuracy. The following steps are parts of the pre-processing phase: feature scaling, data cleansing, exploratory data analysis, and transformation.
2.2.1. Data analysis

The data that is visible to the human eye is not necessarily accurate. Exploratory data analysis (EDA) is used to condense, display, and understand accurate data from data sources. Our knowledge of the data set depends on our ability to extract specific statistical measures and information, such as the number, mean, number, odds, peak, and frequency of categorical data. Its features can be applied to modeling once the data analysis has been completed. Outliers, the connection between traits and class imbalances, and other statistical measurements, such as outliers, can be displayed using graphs, box plots, and scatter plots.

2.2.2. Cleaning the data

Data need to be processed for proper model training after the data set has been balanced. Before training the model, the data must be prepared as follows:

- Removing features that not effecting the model (unnecessary)

  The functions such as anonymous 0, source port, destination IP, source IP address, destination port, stream ID, timestamp, and similar HTTP are all eliminated because they are superfluous and socket related. Because different networks have different values for this attribute, therefore, package properties are used to train the model. Additionally, the IP addresses of the attacker and common user can be similar. Furthermore, an ML model can be biased due to the handling problem caused by the use of socket functions to train the model. It is possible to get 80 new features by removing redundant features.

- Data cleaning and imputation

  The majority of ML algorithms demand tests without values being lost. Noise or missing values impair the model's accuracy. In the suggested work, redundant features are removed to reduce the computational cost while groups that contain deficient or NaN, inf values, are not deleted. Since the attack rating on each die offers some basic information, the calculation of the negative values with 0 values, inf values, and missing values is the final step in processing the noisy data.

2.2.3. Feature selection

The family services stage (FSS) of this study employed the Rao algorithm. A randomly generated initial set, which includes a teacher and a group of students that make up the solution set, serves as the initialization step of the Rao algorithm. Rao uses mutation and crossover factors from GA that represent the function of chromosomes to represent its features. This chromosome is updated using the crossover. Every solution in society is viewed as an individual or chromosome (Figure 2). When a chromosome's characteristic gene has a value of 1, it is regarded as a determinant, however when it has a value of 0, it is the opposite.

![Figure 2. Chromosom](image-url)

The proposed method is comprised of the following detailed steps:

Step 1. Randomly initialize the population; the features of each population must differ from that of the others.
Step 2. Determine the best and worst populations based on the classification accuracy for each feature set.
Step 3. Update the solutions based on the specified best and worst solutions and random interactions based on New_set= random_set crossover with (best_set crossover with worst_set).
Step 4. Keep the new set of features if they are better than the old best set (in terms of classification accuracy).
Step 5. Report the best set of features if the termination criteria have been met, else, go to step 3.

Different measures may be assigned to the values of each function in the data set. Training the model at various levels requires complexity, and time, and occasionally results in model errors. We employ a scaling method known as Standard Scaler to prevent this. This technique's goal is to convert the values of the data set's numerical columns to a standardized scale with keeping the distinctions between the ranges of values. The training instance is done using the following default settings:

\[ S = (s - \mu)SD \]  

where \( S \) = Standard Scaler, \( \mu \) = mean, \( SD \) = standard deviation of the training set.

2.2.4. Transformation

Different types of data functions are contained in the current dataset. Since scalar values may be understood by ML algorithms, it is necessary to transform non-scalar values into numeric values using “Tag Encoder” technology. Assign each data category a special number, starting at 0.
2.2.5. Model tuning

The suggested strategy for choosing hybrid features to extract the best features and altering the hyperparameter to select the optimal parameters for improved model performance is discussed in this section. FS the data that pre-processed can then be used with any ML model after processing is done. Role selections are crucial for developing models with optimum performance [1], [2]; this can be divided into three categories: filtering, packaging, and inline [3], [4]. The filter technique uses a single variable to define the set of independent features.

In a multivariate, criteria-based method, important features are selected by sifting out redundant, overlapping, and highly correlated features. The selected roles for the ML are specified. Computing costs are lower for filtering methods than for other methods. The embedding approach operates by assessing chosen sets of functions using an ML algorithm and employing a search strategy to locate a potential subset of functions. This procedure is repeated until the best feature set is obtained with satisfactory outcomes. This is computationally intensive since it searches for multiple feature sets. Function selection in Create file is performed by using built-in techniques.

2.2.6. SVM

As a supervised learning strategy, SVM can be applied to classification and regression problems using support vector classification (SVC) and support vector regression (SVR) respectively. Each data point in SVM is represented by a point in n-dimensional space, with each function value corresponding to a certain coordinate's interpretation. Finding the ideal hyperplane and efficiently classifying the data set are the major goals of SVM. SV, which are the locations nearest to the target groups for hyperplane, are determined as decision boundaries that assist in classifying target groups.

3. RESULTS AND DISCUSSION

This section thoroughly describes the evaluation of the proposed model. The experimental setup is described first, followed by the results. The results in Table 1 demonstrate the strength of the proposed method as well as future directions for future work.

<table>
<thead>
<tr>
<th>Table 1. Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>DDosNet</td>
</tr>
<tr>
<td>ID3</td>
</tr>
<tr>
<td>LSTM</td>
</tr>
<tr>
<td>Proposed method</td>
</tr>
</tbody>
</table>

3.1. Experimental setup

The implementation of the proposed model was done in MATLAB; the experiments were conducted on a PC that has these specifications: Intel Core (TM) i5-10500H CPU @ 2.50 GHz, 2.50 GHz 16 GB RAM, and Windows 11 OS.

3.2. Performance metrics

The evaluation measures are used to gauge how well the suggested solution performs. The “CICDDoS2019” dataset [1] is used to train and test the suggested model that combines a hybrid feature selection method with SVM classifier. The metrics used to determine the model performance are as (2):

\[
\text{Acc} = \frac{TP}{TP + FN + TN}
\]  \hspace{1cm} (2)

Precision (Prc): A measure that determines the ratio of successfully detected DDoS attacks among the overall predicted attacks; it is calculated thus:

\[
\text{Prc} = \frac{TP}{TP + FP}
\]  \hspace{1cm} (3)

Recall (Rc): A measure of the ratio of correctly detected DDoS attacks among the number of actual DDoS attacks; it is calculated as:

\[
\text{Rc} = \frac{TP}{TP + FN}
\]  \hspace{1cm} (4)
F-score (F1): a measure of the harmonic mean of recall and precision for the attack detection; it is calculated as:

\[
F_1 = \frac{2 \times \text{PrC} \times \text{ReC}}{\text{PrC} + \text{ReC}}
\]

4. CONCLUSION

DDoS attacks alter the size and shape of network resources to drain the resources of the targeted network. Hence, this study proposes an automatic detection method that precisely categorizes the attacks to reduce the harmful impact. To prevent sampling bias, the dataset used in this work is first balanced, then, the hybrid function is selected. One technique involves choosing the right features, which is followed by the implementation of hyperparameter adjustment to enhance model performance. Finally, the supervised learning approach is introduced to usher in unique features and optimum hyperparameters for distinguishing between regular traffic and DDoS attacks. The proposed model is observed to be superior to the current ones when these results are contrasted with the existing techniques. The proposed approach can therefore be applied on any network as a predictive model for effective DDoS attack detection.

REFERENCES


Mohammad Aljanabi received the B.Sc. degree in Computer science from the University of Almustansyah, Iraq, in 2010, and the master’s degree in Computer science from BAMU universirt, India and the Ph.D. degree in Computer science from the Malaysia Pahang (UMP) in 2022, my research area is cybersecurity, network security, machine learning, and AI in general. He can be contacted at email: mohammad.cs88@gmail.com.

Russul Hayder received the Bachelor in software engineering degree from the University of Madenat Alelem Iraq, in 2014, and the master’s degree in Software engineering degree science from Iraq commission for Computers and Informatics/Informatics Institute for postgraduate studies in 2018, Iraq. She can be contacted at email: engrusel92@yahoo.com.

Shatha Talib received the B.Sc. degree in Computer Science from the University of Technology, Baghdad, Iraq, in 2014, and the master’s degree in Computer Science from the University of Technology, in 2017. She is currently an assistant teacher in Al-Bayan University College. Her current research interests include cooperative communication, relay selection and bit error rate. She can be contacted at email: shatha.talib@albayan.edu.iq.
Ahmed Hussien Ali was born in Baghdad, Iraq, in 1988. He received the B.Sc. degree in Computer Science from the University of Al-Mustansiriyah, Iraq, in 2010, and the M.Sc. degree from BAMU University, India. He got Ph.D. from ICCL Informatics Institute for Postgraduate Studies, Baghdad, Iraq and Faculty member, Computer Science Department, College of Education, Al-Iraqia University, Adhamiya, Baghdad, Iraq. He can be contacted at email: msc.ahmed.h.ali@gmail.com.

Mostafa Abdulghafoor Mohammed currently works at the Al-Imam Al Aadhham University College. Mostafa does research in Information technology, Computer Communications (Networks), Cloud Computing and Communication Engineering. Their current project is ‘offloading in mobile cloud computing’. He has finish his master from Department of Computer Science at BAMU University, India, and he finish Ph.D in Computer science and IT at University Polytechnic of Bucharest, Romania. He can be contacted at email: alqaisy86@gmail.com.

Tole Sutikno is currently employed as a lecturer in the Electrical Engineering Department at Universitas Ahmad Dahlan (UAD), which is located in Yogyakarta, Indonesia. In 1999, 2004, and 2016, he graduated with a Bachelor of Engineering from Universitas Diponegoro, a Master of Engineering from Universitas Gadjah Mada, and a Doctor of Philosophy in Electrical Engineering from Universiti Teknologi Malaysia. All three degrees are in the field of electrical engineering. Since the year 2008, he has held the position of Associate Professor at the University of Ahmad Dahlan in Yogyakarta, Indonesia. He is currently the Head of the Embedded Systems and Power Electronics Research Group in addition to holding the position of Editor-in-Chief of TELKOMNIKA. His research interests include the areas of digital design, industrial applications, industrial electronics, industrial informatics, power electronics, motor drives, renewable energy, FPGA applications, embedded systems, artificial intelligence, intelligent control, digital libraries, and intelligent control. He can be contacted at email: tole@te.uad.ac.id.