Machine Learning-Based Distributed Denial of Services (DDoS) Attack Detection in Intelligent Information Systems

Wadee Alhalabi, Immersive Virtual Reality Research Group, Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia
Akshat Gaurav, Ronin Institute, USA, & UCRD, Chandigarh University, Chandigarh, India
Varsha Arya, Department of Business Administration, Asia University, Taiwan, & Immersive Virtual Reality Research Group, King Abdulaziz University, Jeddah, Saudi Arabia, & Lebanese American University, Beirut, Lebanon, & Center for Interdisciplinary Research at University of Petroleum and Energy Studies (UPES), Dehradun, India*
Ikhlas Fuad Zamzami, Faculty of business, King Abdulaziz University, Rabigh, Saudi Arabia
Rania Anwar Aboalela, Department Information System, King Abdulaziz University, Rabigh, Saudi Arabia

https://orcid.org/0000-0003-2284-4610

ABSTRACT

The danger of distributed denial of service (DDoS) attacks has grown in tandem with the proliferation of intelligent information systems. Because of the sheer volume of connected devices, constantly shifting network circumstances, and the need for instantaneous reaction, conventional DDoS detection methods are inadequate for the IoT. In this context, this study aims to survey the current state of the art in the topic by reading relevant articles found in the Scopus database, with a brief overview of the IoT and DDoS as this study examines neural networks and their applicability to DDoS detection. Finally, a decision tree-based model is developed for the detection of DDoS attacks. The analysis sheds light on the present trends and issues in this field and suggests avenues for further study.

KEYWORDS:

DDoS, Artificial intelligence, Machine learning, IoT, Neural Networks

INTRODUCTION

Distributed Denial of Service (DDoS) attacks are designed to overwhelm a network with malicious traffic (Kamaljeet Kaur & Parveen Kakkar, n.d.; Q. Zhang et al., 2023; Cvitic´ et al., 2021). DDoS attacks are a significant threat to web-based applications and networks. Lau et al. (Lau et al., n.d.) describes the methods and techniques used in DDoS attacks and lists possible defenses. Salim et al.(Salim et al., 2019) comprehensively surveys DDoS attacks from IoT devices to the cloud...
environment, including attack methods, tools, and state-of-the-art defense measures. Bhuyan et al. (Bhuyan et al., 2013) discusses the two types of DDoS attack architectures: the Agent-Handler architecture and the Internet Relay Chat (IRC)-based architecture. Patil et al. (Patil et al., 2021) presents a comprehensive review of existing distributed frameworks for detecting DDoS attacks and characterizes several existing distributed processing frameworks to select an appropriate one for deploying DDoS attack detection mechanisms.

DDoS is a severe attack caused so far in the world, that crashes many servers, blocks network traffic, and drastically reduces its speed. Various technologies such as Machine Learning, Deep Learning, Blockchain, and Cyber Security have been applied by researchers for handling DDoS attacks. Securing data in an IoT network is challenging due to its decentralized nature and data being shared among millions of devices (Varalakshmi et al., 2021; Stergiou et al., 2021; A. Singh & Gupta, 2022). Machine Learning is a powerful tool for detecting DoS/DDoS attacks. Different machine learning techniques have been used to detect DoS/DDoS attacks, such as supervised, unsupervised, and deep learning. The accuracy of the detection of DoS/DDoS attacks can be improved by combining different machine learning techniques (Verma & Kumar, 2021; B. B. Gupta et al., 2009; Z. Zhang et al., 2017). DDoS attack detection is a challenging task in cloud computing. Artificial Intelligence (AI) based approaches can be used to detect DDoS attacks in cloud computing. Comprehensive reviews of existing DDoS attack detection methods are needed to improve the security of cloud computing (R. Devi & N. Umamaheswari, n.d.; Dahiya & Gupta, 2021; Kumar et al., 2021; Wahab et al., 2017). Feature selection is an important factor in improving the accuracy of ML-based solutions for detecting DDoS attacks. Different datasets such as KDD, UNSW-NB15, and others can affect the accuracy of ML. Several feature engineering strategies can be chosen to improve ML solutions on DDoS attacks (Faiz et al., 2022).

The consequences of a DDoS attack are varied and can affect multiple stakeholders (Abbas et al., 2021; Wassan et al., 2022; Mishra et al., 2021). Somani et al. (Somani et al., 2016) argues that in cloud computing, collateral damage to non-targets can include performance interference, web service performance, resource race, indirect EDoS, service downtime, and business losses. Maciel et al. (Maciel et al., 2018) proposes hierarchical models to assess the impact of a DDoS attack on computer systems, including the likelihood of an attack, attacker benefits, feasibility, the pain factor, and the propensity of the offense. Abhishta et al. (Abhishta et al., 2017) analyzes the impact of DDoS attack announcements on victim stock prices and finds a significant negative impact in cases where the attack creates an interruption in services. Hurst et al. (Hurst et al., 2015) focuses on predicting the effects of DDoS attacks on a network of critical infrastructures and demonstrates a technique for assessing the future impact of disruptions on an integrated critical infrastructure network. Overall, the papers suggest that the consequences of a DDoS attack can be significant and wide-ranging, affecting not only the target but also other stakeholders and potentially causing financial losses.

Intelligent information systems are also affected by DDoS attacks. According to research suggest that DDoS attacks on Intelligent information systems can cause significant damage (Shahzad et al., 2022; AbdulRahman et al., 2020). Kolias et al. (Kolias et al., 2017) warns that the Mirai botnet and its variants can expose the Internet infrastructure to increasingly disruptive DDoS attacks. Lyu 2017 quantifies the reflective DDoS attack capability of household IoT devices, showing that they can be exposed to Internet reflection even if they are secured behind home gateways. Al-Hadhrami et al. (Al- Hadhrami & Hussain, 2021) comprehensively reviews the attacks that can lead to DDoS in Intelligent information system networks and investigates the available solutions used to counter these attacks. Mustapha et al. (Mustapha & Alghamdi, 2018) provides an analysis of the attempts to prevent DDoS attacks, mainly at a network level, and concludes that there is yet to be a perfect solution for IoT security. In this context, we analyze the current development in the field of DDoS attack detection based on neural network techniques.
RESEARCH METHODOLOGY

Data are gathered and analysed from Scopus, the largest database of academic literature. The major purpose of this research is to analysis of development of detection of DDoS Attack detection in Intelligent information systems based on machine learning. To do this, we gathered information on factors including authors, keywords, and sources. We employed metrics like frequency of occurrence, degree of centrality, and significance to evaluate the data and spot trends. We also used network analysis to graphically display the connections between the keywords and isolate groups of similar terms. In sum, this analysis serves as a helpful framework for recognizing major themes and future research topics in the study of data science’s effect on sustainable entrepreneurship.

RESULTS AND DISCUSSION

The information in Figure 1 comes from the Scopus database and spans the years 2016 through 2023, drawing from a total of 165 papers published in 113 periodicals. Each document is an average of 1.92 years old and has been cited 14.13 times. There are a total of 4385 citations, or an annual average of 3.249 citations per document.

Articles, book chapters, conference papers, and conference reviews are the four sorts of documents included here. The majority of these 77 resources are articles, followed by 60 conference papers. There are additionally keyword analyses performed on the papers; they include 775 additional keywords (ID) and 350 author-supplied keywords (DE).

There are 495 unique writers mentioned in the texts (out of a total of 559 authors). Only five of these writers have produced works solely under their own names; the remaining 490 have all contributed to group efforts. There are a total of 28 documents with a single author, for a per-author average of 0.333.

Three people are listed as authors on average, with an additional 3.39 people listed as co-authors. There is a 3.58 cooperation index.

Several inferences may be made from this data. First, articles and conference papers make up the bulk of the documents, suggesting that they are the most common types of publications in the
chosen subjects. Second, the high average number of citations per document is an indication of the papers’ prominence in their domains. Thirdly, there seems to be a lot of cooperation in the chosen disciplines as seen by the large number of writers and co-authors per document. Last but not least, the high cooperation index and the small percentage of single-authored papers lend credence to the notion that teamwork is crucial to scientific progress in these areas.

Figure 2 presents data on the number of articles published in Scopus-indexed journals in different years, from 2016 to 2023. The number of articles has been steadily increasing over the years, with the highest number of publications in 2022. The annual growth rate of the papers are 60.35%.

In 2016, only one article was published in Scopus indexed journals, which is a relatively small number, but it could be an outlier or the data could be incomplete. In 2018 and 2019, the number of publications increased to 11 and 10, respectively. However, the significant increase in publications was observed in 2020, with 23 articles published. In 2021, the number of publications almost doubled to 42, indicating a significant growth in research activity in the selected fields. The trend continued in 2022, with 61 articles published, the highest number of publications in the selected years. However, in 2023, the number of publications decreased to 17, but this could be because the data was collected in the earlier months of 2023.

From this data, we can draw several conclusions. Firstly, the number of articles published in Scopus indexed journals has been increasing steadily over the years, with the highest growth observed in recent years. Secondly, the significant increase in publications in 2020 and 2021 indicates an increased interest in research in the selected fields. Finally, the decrease in the number of publications in 2023 could be due to data incompleteness or seasonal variations, and it is too early to say whether this trend will continue.

ANALYSIS OF AUTHORS

Figure 3 presents the authors’ statistics, the number of articles they have published, and the corresponding fractionalized number of articles. The top author has published 24 articles, representing the entirety of their contribution to the dataset. Among the authors with multiple publications, it is interesting to note that some have a higher fractionalized number than others, indicating that they have
contributed to multiple papers with other co-authors. For instance, author SBAI O has a fractionalized number of 1.5, which implies that they have contributed to one or more papers with other authors. These findings suggest that some authors have a more collaborative approach to research than others, and that collaborations may effectively increase research output. Additionally, the data highlights the importance of individual author contributions in shaping the overall research landscape, as the top author alone has contributed to 24 out of the 47 papers in the dataset.

ANALYSIS OF COUNTRY

Figure 4 illustrates how often articles from different nations appear in print. There were 70 items from India, 43 from China, and 30 from the United States. Although the United Kingdom and Pakistan each had 15 articles published, Iraq and Saudi Arabia each had 21. Australia and Indonesia both published five papers, while Malaysia and Brazil both published seven. Lastly, there were 5 or less papers published in Ireland, Jordan, Algeria, Morocco, Bangladesh, Bulgaria, and France. This information suggests that India and China are the primary sources of scholarly publications, with the United States also making a significant contribution. Iraq, Saudi Arabia, Pakistan, and Malaysia are only a few examples of Middle Eastern and Southeast Asian nations that actively contribute to the published research.

ANALYSIS OF PUBLICATION SOURCE

Bradford’s Law is a mathematical notion that may be used to forecast how widely dispersed certain types of scientific publications will be. In the 1930s, librarian Samuel C. Bradford created a rule that is widely used in bibliometrics and scientometrics to examine the concentration of authoritative works within a certain topic.

Bradford’s law states that there are three distinct areas of scientific literature for each given discipline. On the first level, you’ll find the top-tier journals publishing groundbreaking research. Journals that are nonetheless significant but lack the outsized impact of the first zone may be found in the second. A huge number of journals of lesser importance and usually containing articles of a more specialised type make up the third zone.
The number of journals on the x-axis and the number of articles on the y-axis may be used to create a logarithmic graph showing the distribution of literature throughout these three regions. The resultant curve, frequently called the Bradford curve, is a straight line with two distinct kinks. The transition between Bradford zones is marked by a sharp reversal in the curve’s slope.

In various disciplines, including those of science, technology, and medicine, Bradford’s Law has been used to examine the dispersion of pivotal resources. Finding the leading journals in an area allows scholars to concentrate on the most useful publications while dismissing the rest. This is especially helpful for researchers who are short on time but still want to do a literature review or find the most significant publications in their subject.

Bradford’s Law is a useful statistical theory, although it should be seen more as a suggestion than a hard and fast rule. Factors such as research focus, historical context, and material access all...
play a role in how literature is actually dispersed within a discipline. Researchers should use care when using Bradford’s Law and should constantly double-check their findings with other techniques.

Bradford’s Law is applied to a group of journals in Figure 5 (Zone column). According to Bradford’s Law, there are only a handful of journals that publish the vast majority of research in any given discipline. According to the legislation, the number of journals publishing papers in a particular subject may be roughly distributed into three zones, with each zone containing a different proportion of the field’s total.

ANALYSIS OF TRENDING TOPICS

Figure 6 displays data on the prevalence of certain search terms within a given setting. The terms “internet of things” and “denial-of-service attack” rank first and second, respectively, with 107 and 97 occurrences. This suggests that concerns about the safety of internet-connected gadgets and the hazards they face will be discussed. Similar terms include “intrusion detection,” “distributed denial of service,” and “network security.” The inclusion of “deep learning” and “neural networks” on the list is intriguing since it suggests that internet-connected devices are being protected using machine learning and artificial intelligence strategies. All in all, the numbers show how critical it is to have a secure network and how sophisticated methods are needed to counteract the many security risks that exist today.

Thematic Analysis

A thematic map is a form of map that focuses on a particular theme or topic, as opposed to simply displaying geographic features like physical and political maps. The purpose of a thematic map is to display the geographic spread of a selected property or subject. They excel in revealing and analysing hidden connections and patterns in large datasets.

The density of a region’s population, the incidence of a certain illness throughout the region, or the locations of various land uses inside a metropolis are just some of the topics that may be shown on a thematic map. Thematic maps are useful because they may illustrate patterns and connections that
might not be immediately obvious in a basic table or spreadsheet by using various colours, symbols, or patterns to represent different levels or categories of the subject.

In conclusion, thematic maps are an effective method for gaining insight from and making sense of geographical data. They have many potential applications, including but not limited to environmental research, business analysis, and public health. Figure 7 presents the thematic map related to the current research topic. As presented in Figure 7, the thematic map is divided into the following areas.

- Motor themes: These main concepts are crucial to answering the study question and can be seen in the majority of the data. These are usually determined at the commencement of a study and are regarded as the study’s primary motivation.
- Basic themes: These are the most prominent motifs across the data, and they are important to comprehending the study’s central subject. They are more concrete than motor themes and provide light on the study problem at hand.
- Niche themes: These are the specialised, less widespread trends in the data. They may not be present in all samples, but when they are, they provide light on the research problem.
- Emerging themes: These themes may not have been anticipated at the onset of the study, but they become clear as the data is analysed. They occasionally contradict or expand upon established conclusions, but always provide fresh, useful information to the investigation at hand.

ANALYSIS OF PUBLISHED DOCUMENTS

In this subsection, we analyze the most globally cited documents. Present this analysis helps the new researchers to select the most popular papers in the respective domain.

Figure 7. Thematic map
THEORITICAL AND PRACTICAL IMPL eCATIoNS

In this section, we presented the details about the important observation that we find from this research:

- **Observation 1**: Machine learning techniques can be used to improve the detection of DDoS attacks in IoT. There is a lack of knowledge about how to ensure security in IoT. Therefore, there is a need for further research on this topic (A. Ashraf & Elmedany, 2021; Brdesee et al., 2022; Almomani et al., 2022).

- **Observation 2**: Artificial Intelligence (AI) techniques are performing better accuracy than traditional methods to detect DDoS attacks in WSNs. Support Vector Machine (SVM) and Artificial Neural Networks (ANN) are the most used AI-based techniques to detect DDoS attacks in the wireless sensor network. The performance of AI techniques-based detection systems for DDoS attacks in WSN is remarkable (Al-Naeem et al., 2020; Tembhurne et al., 2022; Li et al., 2022; Stylianou et al., 2022).

- **Observation 3**: AI methods such as deep learning, machine learning, support vector machines, random forest, extreme gradient boosting, neural networks, and recurrent neural networks are effective in detecting cybersecurity attacks in the IoT environment. Smart intrusion detection systems with intelligent architectural frameworks using AI can help to overcome existing security and privacy challenges. AI methods can be used to detect threats based on attack categories (Abdullahi et al., 2022; Gaurav et al., 2022; Pan et al., 2022; Afify et al., 2022; Vijayakumar et al., 2022).

- **Observation 4**: Different detection techniques are available to prevent DDoS attacks on SDN controllers, such as anomaly detection, signature-based detection, and honeypots. These techniques have different characteristics, such as accuracy, scalability, and false positive rate. Resource consumption, privacy, and security issues can arise when using these techniques (Zubaydi et al., 2017; G. Singh et al., 2022; T. Gupta & Panda, 2022; Dwivedi et al., 2021).

- **Observation 5**: DDoS attacks can be identified as a classification problem on network state. Transmission failures or deadline misses can cause disruptions to the process and corruption of the overall control performance. DDoS attack detection and DSR Algorithm with Cryptography can be used to improve security on Wireless Sensor Networks with BS, CH (Kaur et al., 2018; Priyanka & Cherian, 2021).

- **Observation 6**: ICMPv6 protocol is an important part of IPv6 and is responsible for sending and receiving messages. DDoS attacks are a major threat to IPv6 networks and can cause significant economic damage. Anomaly detection techniques can be used to detect ICMPv6-based DDoS attacks, and feature selection techniques based on bio-inspired algorithms can be used to improve detection accuracy (Adnan Hasan Bdair Alghuraibawi et al., n.d.).

Proposed Model

In this subsection, we give the details about the simulation results. We used the DDoS dataset to train the random forest model model. The dataset consists of many unwanted terms; therefore, we preprocess the dataset to extract valuable information from the dataset.

Preprocessing of Dataset

The dataset consists of ‘12794627’ rows and ‘83’ columns. This dataset is a combination of CSE-CIC-IDS2018-AWS, CICIDS2017, CIC DoS dataset(2016) datasets. Each row of the dataset is labeled as ‘DDoS’ and ‘Bening’. Further steps of data preprocessing are as follows:

1. **Data Cleaning**: Many columns of the dataset contain ‘Na’ values or invalid values. These invalid values may affect model training. Therefore, we remove all rows that contain invalid values.
2. **Data Normalization**: In order to get accurate results, the values of the data set should be on the same scale. Therefore, data normalization is necessary. After data normalization, all the data values in the dataset are scaled.

3. **One hot encoding**: As defined previously, the ‘labels’ of the datasets are classified as ‘DDoS’ and ‘Bening.’ But the machine learning model only understands the numerical values. Therefore, before training the model, we have to convert the ‘labels’ into numeric values. This can be done by ‘one hot encoding’. We convert ‘DDoS’ into ‘0’ and ‘Bening’ into ‘1’.

### Table 1. Highly cited papers

<table>
<thead>
<tr>
<th>Paper</th>
<th>DOI</th>
<th>Total Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOSHI R, 2018, PROC - IEEE SYMP SECUR PRIV WORK- SHOPS, SPW (Doshi et al., 2018)</td>
<td>10.1109/SPW.2018.00013</td>
<td>375</td>
</tr>
<tr>
<td>HODOE, 2016, INT SYMP NETW, COMPUT COMMUN, ISNCC (Hodo et al., 2016)</td>
<td>10.1109/ISNCC.2016.7746067</td>
<td>321</td>
</tr>
<tr>
<td>SU J, 2018, PROC INT COMPUT SOFTWARE APPL CONF (Su et al., 2018)</td>
<td>10.1109/COMPSAC.2018.10315</td>
<td>189</td>
</tr>
<tr>
<td>MCDERMOTT CD, 2018, PROC INT JT CONF NEURAL NET- WORKS (McDermott et al., 2018)</td>
<td>10.1109/IJCNN.2018.8489489</td>
<td>173</td>
</tr>
<tr>
<td>JIA Y, 2020, IEEE INTERNET THINGS J (Jia et al., 2020)</td>
<td>10.1109/JIOT.2020.2993782</td>
<td>104</td>
</tr>
<tr>
<td>MANIMURUGAN S, 2020, IEEE ACCESS (Manimurugan et al., 2020)</td>
<td>10.1109/ACCESS.2020.2986013</td>
<td>99</td>
</tr>
<tr>
<td>HWANG RH, 2020, IEEE ACCESS (Hwang et al., 2020)</td>
<td>10.1109/ACCESS.2020.2973023</td>
<td>77</td>
</tr>
<tr>
<td>CHURCHER A, 2021, SENSORS (Churcher et al., 2021)</td>
<td>10.3390/s21020446</td>
<td>59</td>
</tr>
<tr>
<td>ROOPAK M, 2020, ANNU COM- PUT COMMUN WORKSHOP CONF, CCWC (Roopak et al., 2020)</td>
<td>10.1109/CCWC47524.2020.9031206</td>
<td>52</td>
</tr>
<tr>
<td>HUSSAIN F, 2020, PROC - IEEE INT MULTI-TOP CONF, INMIC (Hussain et al., 2020)</td>
<td>10.1109/INMIC50486.2020.9318216</td>
<td>45</td>
</tr>
<tr>
<td>FERRAG MA, 2021, ELECTRON- ICS (SWITZERLAND) (Ferrag et al., 2021)</td>
<td>10.3390/electronics10111257</td>
<td>44</td>
</tr>
<tr>
<td>DE ASSIS MVO, 2020, COMPUT ELECTR ENG (de Assis et al., 2020)</td>
<td>10.1016/j.compeleceng.2020.106738</td>
<td>44</td>
</tr>
<tr>
<td>THANTHARATE A, 2020, ANNU COMPUT COMMUN WORKSHOP CONF, CCWC(Thantharate et al., 2020)</td>
<td>10.1109/CCWC47524.2020.9031158</td>
<td>42</td>
</tr>
<tr>
<td>SOE YN, 2019, PROC INT CONF INF COMPUT, ICIC (Soe et al., 2019)</td>
<td>10.1109/ICIC47613.2019.8985853</td>
<td>36</td>
</tr>
<tr>
<td>PROTOGEROU A, 2021, EVOL SYST (Protogerou et al., 2021)</td>
<td>10.1007/s12530-020-09347-0</td>
<td>23</td>
</tr>
<tr>
<td>JITHU P, 2021, SN COMPUT SCI (Jithu et al., 2021)</td>
<td>10.1007/s42979-021-00516-9</td>
<td>21</td>
</tr>
</tbody>
</table>
Simulation Environment
The experimental test environment in this paper is Windows 11 PC, an Intel (R) Core (TM) i7-7300HQ CPU @ 2.50GHz, 16.00GB RAM. Implement algorithms using Sklearn, Keras, and TensorFlow libraries.

Experiment Results
The simulated results are shown in this section. We employed a Gini index-based approach, which is a type of decision tree classifier, to conduct the tests. The final decision trees for the Gini index based is presented in Figure 8. In a 70:30 split, we used half the data for training and the other half for testing.
From the experiment, it is clear that our proposed model predicts the DDoS attack in an intelligent information system environment with an accuracy of 99%.

CONCLUSION

On the basis of a literature review pertaining to the development of DDoS detection in intelligent information systems using machine learning, we can conclude that machine learning is promising technique for improving the security of intelligent information systems. In addition to that, deep learning, convolutional neural networks, and recurrent neural networks are only a few of the neural network-based methodologies offered by academics for identifying DDoS attack in intelligent information systems. These methods performed very well in real-world DDoS attack detection tests. We also proposed a Gini-index-based DDoS attack detection model for intelligent information systems, which detects DDoS attack with an accuracy of 99%. Nevertheless, further study is required to enhance the scalability and generalizability of these methods and investigate their potential in responding to emerging security risks and assaults. Overall, this study demonstrates the need for more investigation into the application of cutting-edge neural network approaches to the problem of intelligent information systems.

ACKNOWLEDGMENT

This research work was funded by Institutional Fund Projects under grant no. (IFPRC-119-611-2020). Therefore, the authors gratefully acknowledge technical and financial support from Ministry of Education and Deanship of Scientific Research (DSR), King Abdulaziz University (KAU), Jeddah, Saudi Arabia.
REFERENCES


McDermott, C., Majdani, F., & Petrovski, A. (2018). *Botnet detection in the inter- net of things using deep learning approaches.* Institute of Electrical and Electronics Engineers Inc. 10.1109/JICCNN.2018.8489489


Soe, Y., Santosa, P., & Hartanto, R. (2019). *Ddos attack detection based on simple ann with smote for iot environment*. Institute of Electrical and Electronics Engineers Inc. 10.1109/ICIC47613.2019.8985853


Thantharate, A., Paropkari, R., Walunj, V., Beard, C., & Kankariya, P. (2020). Se- cure5g: A deep learning framework towards a secure network slicing in 5g and beyond. In P. R. Chakrabarti S. (Ed.), (p. 852-857). Institute of Electrical and Electronics Engineers Inc. 10.1109/CCWC47524.2020.9031158


