iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66594

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

Intrusion Detection using Machine Learning

Mohammed Ameer Hamza[1], Mohammed Ghouse A[2], Mutturaj U[3], Sai Kiran[4]
First-Computer science Department, First- Ballari Institute of Technology and Management

Abstract: ***It is well known that distributed attacks simultaneously launched from many hosts have caused the most serious problems in recent years including problems of privacy leakage and denial of services. Thus, how to detect those attacks at early stage has become an important and urgent topic in the cyber security community.***
For this purpose, recognizing C&C (Command & Control) communication between compromised bots and the C&C server becomes a crucially important issue, because C&C communication is in the preparation phase of distributed attacks. Although attack detection based on signature has been practically applied since long ago, it is well-known that it cannot efficiently deal with new kinds of attacks. In recent years, ML (Machine learning)-based detection methods have been studied widely.SVM (Support Vector Machine) and PCA (Principal Component Analysis) are utilized for feature selection and SVM and RF (Random Forest) are for building the classifier. We find that the detection performance is generally getting better if more features are utilized.

I. INTRODUCTION

The problems and losses caused by cyber-attacks have been increasing greatly in recent years, despite many works for avoiding and detecting cyber-attacks have been done and a huge amount of money has been invested on cyber security. The main reason for this is that attackers have also been more and more sophisticated. Distributed attacks are those launched cooperatively by many compromised hosts. Such attacks are referred to as next generation cyber-attacks in Xu's work and it is well known that such attacks are one kind of the most sophisticated attacks.
According to many reports, distributed attacks have caused the most serious problems/losses in these years. Thus, many researchers and developers in the cyber security community have been working on how to detect and avoid such attacks. In general, the attacker prepares or hijack a C&C server, which is used to send attack instructions to the compromised hosts (bots).
Then, the bots launch an actual distributed attack to the victim(s). Thus, the C&C communication is preparation phase for distributed attack. If such communication is recognized, the upcoming actual distributed attack might be blocked. Thus, the detection of the C&C communication is regarded as early detection of distributed attacks.

II. PROPOSED METHODOLOGY

Their performance is greatly dependent on the information theoretic measure; they behavior well only when a significantly large number of anomalies are present in the data and, moreover, it is difficult to associate an anomaly score with a test instance.
To detect scan attacks, the session information based on the actions of communication protocols (such as TCP and UDP) was also used in the work. Change-points based methods also have been proposed. For example, the work proposed a change-point based method to detect TCP SYN flood attacks and scan attacks by computing the characteristics of the packets.

III. SOFTWARE ARCHITECTURE

A. Distributed Intrusion detection

We focused on the issue of feature selection for early detection of distributed attacks. We implemented the early detection by detecting C&C communication of distributed attacks because those communication is at the preparation phase of distributed attacks. Based on our previous research using 55 features to detect C&C communication, in this paper we investigated what if we remove the features from those of the least importance.

B. DDoS attacks

In those methods, feature selection is obviously very important to the detection performance. We once utilized up to 55 features to pick out C&C traffic in order to accomplish early detection of DDoS attacks. In this work, we try to answer the question that "Are all of those features really necessary?" We mainly investigate how the detection performance moves as the features are removed from those having lowest importance and we try to make it clear that what features should be played attention for early detection of distributed attacks.

©IJRASET: All Rights are Reserved

C. Machine Learning

Distributed attacks are those launched cooperatively by many compromised hosts. Such attacks are referred to as next generation cyber-attacks work and it is well known that such attacks are one kind of the most sophisticated attacks. According to many reports, distributed attacks have caused the most serious problems/losses in these years. Thus, many researchers and developers in the cyber security community have been working on how to detect and avoid such attacks.

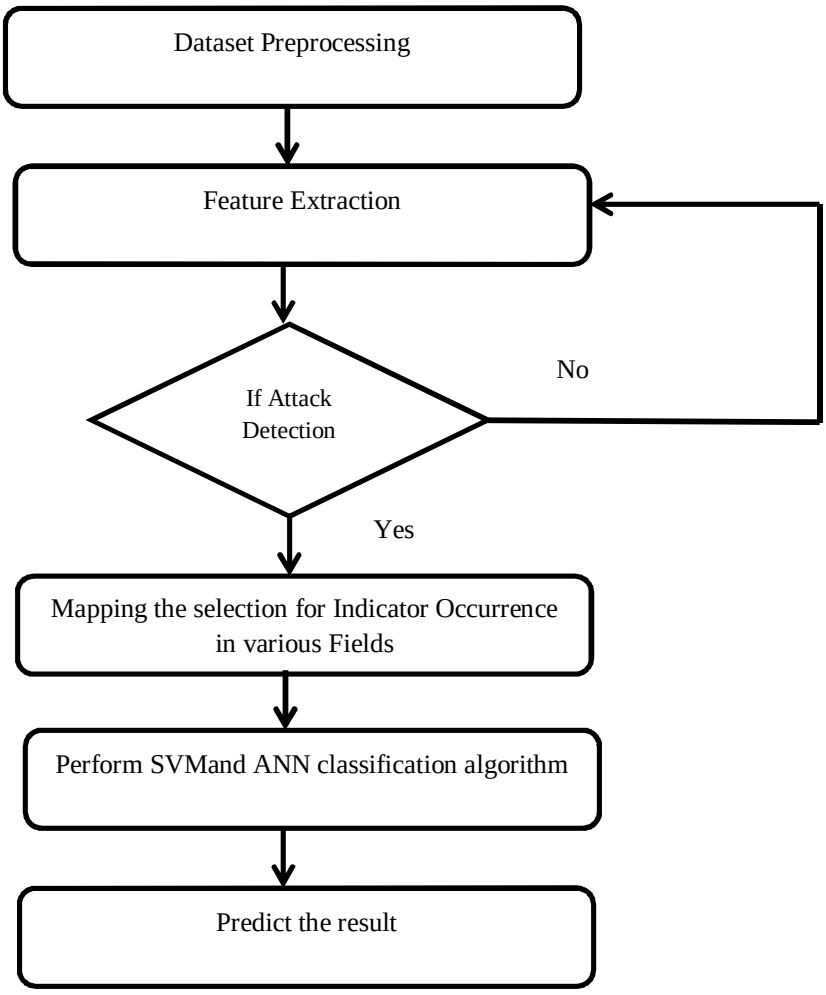
D. Feature Selection

We use honeypot data collected during the period from 2008 to 2013. SVM and PCA are utilized for feature selection and SVM and RF are for building the classifier. We find that the detection performance is generally getting better if more features are utilized. However, after the number of features has reached around 40, the detection performance will not change much even more features are used. It is also verified that, in some specific cases, more features do not always means a better detection performance.

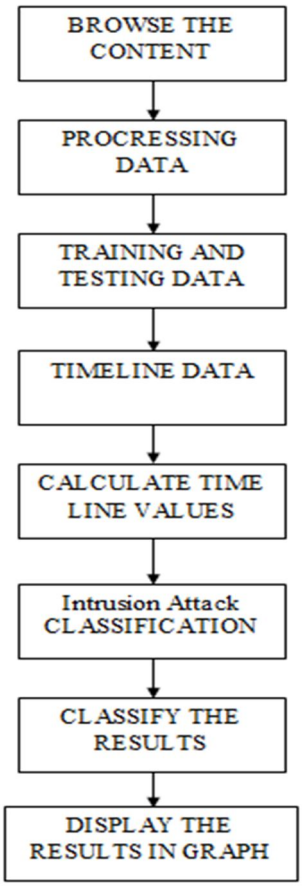
E. Early Detection

There have been many ML (Machine Learning)-based studies on detecting distributed attacks including how to accomplish early detection. Those methods tried to find some particular features of the abnormal traffic to distinguish them from the others. Also, we once used up to 55 traffic features to implement early detection of distributed attacks. In this study, we discuss what will happen if the number of features is decreased gradually.

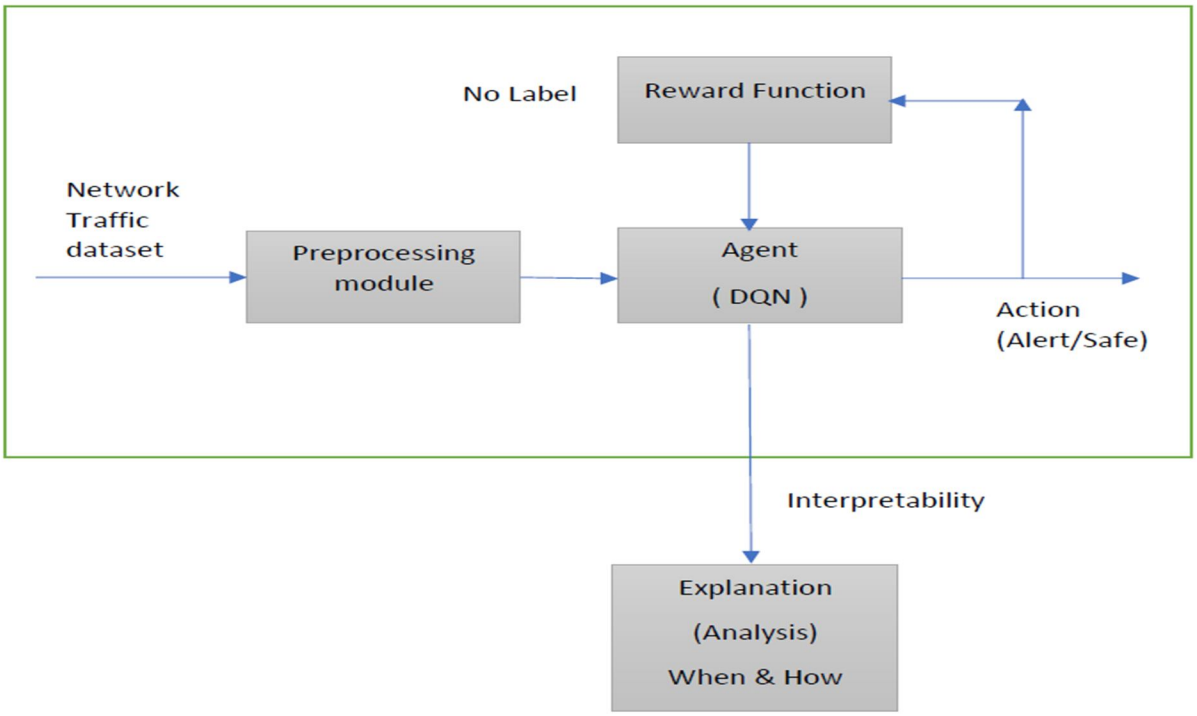
F. Architecture Diagram



G. Activity Diagram

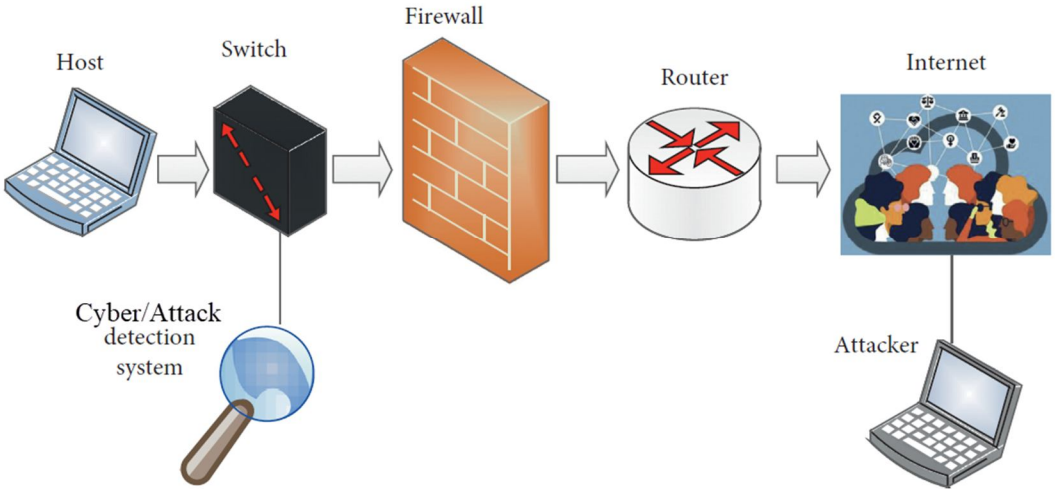


H. Data Flow Diagram



©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

I. Design Diagram



IV. EXISTING SYSTEM

Many approaches have been proposed to detect cyber-attacks including signature-based methods histogram-based methods volume-based methods and information theory-based methods.

It is well known that signature-based methods cannot efficiently deal with new kinds of attacks and new variants. This is because they can only detect the anomalies stored in a predefined database of signatures.

Many statistic histograms are built in histogram-based methods using clean traffic data and all the histograms are mapped into a high-dimensional space. Such methods are easy to understand. But, their false negative rates are often very high.

Volume-based methods need to thresholds that must be determined in advance, which is not easy for most situations. Information theory-based methods also suffer from the following problems.

V. SRS

A. Data Mining

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations-interestingness-metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps. The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations. Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

VI. BIG DATA

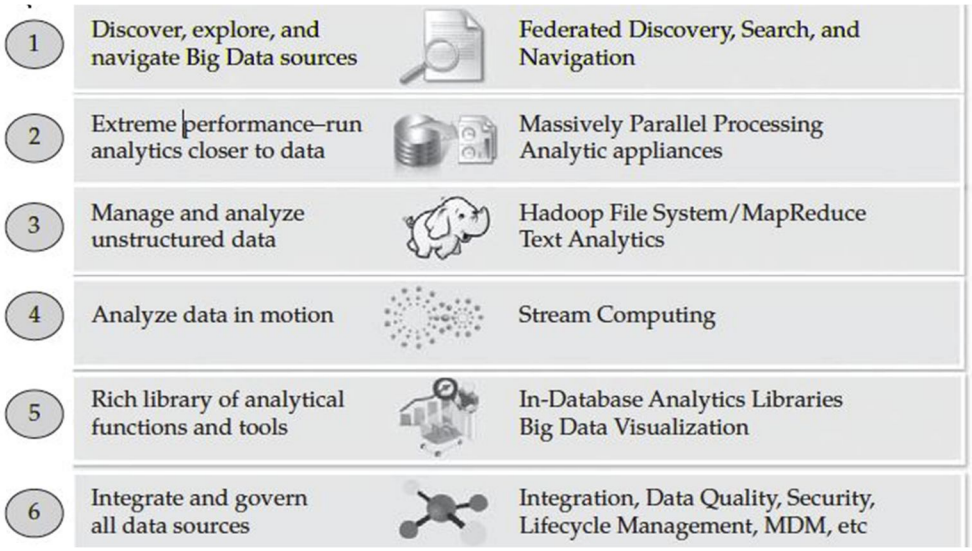
Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools. The challenges include capture, curation, storage, search, sharing, analysis, and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.

Put another way, big data is the realization of greater business intelligence by storing, processing, and analyzing data that was previously ignored due to the limitations of traditional data management technologies

A. The four Dimensions of Big Data

- Volume: Large volumes of data
- Velocity: Quickly moving data
- Variety: structured, unstructured, images, etc.
- Veracity: Trust and integrity is a challenge and a must and is important for big data just as for traditional relational DBs
- Big Data is about better analytics!

B. The Big Data Platform Manifesto



C. Some Concepts

No SQL (Not Only SQL): Databases that “move beyond” relational data models (i.e., no tables, limited or no use of SQL)

- Focus on retrieval of data and appending new data (not necessarily tables)
- Focus on key-value data stores that can be used to locate data objects
- Focus on supporting storage of large quantities of unstructured data
- SQL is not used for storage or retrieval of data
- No ACID (atomicity, consistency, isolation, durability)

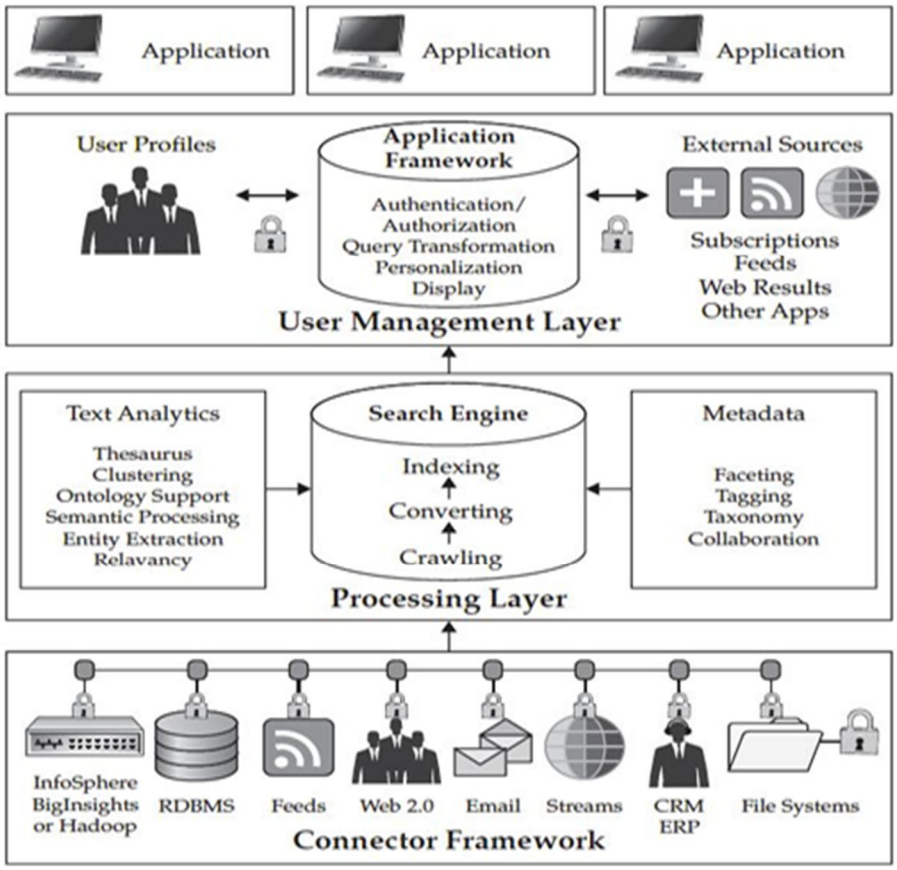
D. Hadoop

- Hadoop is a distributed file system and data processing engine that is designed to handle extremely high volumes of data in any structure.
- Hadoop has two components:
  - The Hadoop distributed file system (HDFS), which supports data in structured relational form, in unstructured form, and in any form in between
  - The MapReduce programming paradigm for managing applications on multiple distributed servers
- The focus is on supporting redundancy, distributed architectures, and parallel processing

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

E. Some Hadoop Related Names to Know

- Apache Avro: designed for communication between Hadoop nodes through data serialization
- Cassandra and Hbase: a non-relational database designed for use with Hadoop
- Hive: a query language similar to SQL (HiveQL) but compatible with Hadoop
- Mahout: an AI tool designed for machine learning; that is, to assist with filtering data for analysis and exploration
- Pig Latin: A data-flow language and execution framework for parallel computation
- ZooKeeper: Keeps all the parts coordinated and working together



VII. WHAT TO DO WITH THE DATA

The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages:

1) Selection
2) Pre-processing
3) Transformation
4) Data Mining
5) Interpretation/Evaluation.

It exists, however, in many variations on this theme, such as the Cross Industry Standard Process for Data Mining (CRISP-DM) which defines six phases:

a) Business Understanding
b) Data Understanding
c) Data Preparation
d) Modeling
e) Evaluation
f) Deployment or a simplified process such as (1) pre-processing, (2) data mining, and (3) results validation.

A. Algorithm

In this study, we discuss what will happen if the number of features is decreased gradually. That is, how the detection performance changes as the features are decreased from those having low importance. Of course, the appropriate number of features is different according to different machine learning algorithms. Honeypot data collected during the period from 2008 to 2013 and the OSS Weak are used in our study. SVM and PCA are used for feature selection and SVM and RF are for building classifier. We find that, the detection performance is generally getting better if more features are used. However, after the number have reached around 40, the detection performance will not change much even more features are used.

B. Technique

Signature-based detection is a direct method to detect botnet in a network and payload analysis is a common technique of signature-based detection. Proposed a botnet detection framework based on signature-based detection, called which is an alert system when some bot behaviors are detected by Snort intrusion detection system. Bot activity can also be detected by picking out C&C session that occurs before actual distributed attacks. Several methods have been proposed to pick out C&C sessions using some features of those sessions. For example, the C&C server only send commands whose packets are of small size. The work analyzes network traffic and uses machine learning to detect IRC-based C&C communication.

VIII. METHODOLOGY

Some methods to detect C&C session based on multiple protocols also have been proposed. The work proposed a C&C traffic detection approach based on analysis of network traffic using seven features the standard deviation of access time and access time. That study claims that their approach is able to effectively detect C&C traffic even in multiple protocols. Our previous research adopts 55 features for the same purpose, which is explained in the next section. In that study, adding some new features to the work we defined a 55-dimensional feature vector to improve the detection performance.

IX. CONCLUSION

In this paper, we focused on the issue of feature selection for early detection of distributed cyber-attacks. We implemented the early detection by detecting C&C communication of distributed attacks because those communication is at the preparation phase of distributed attacks. Based on our previous research using 55 features to detect C&C communication, in this paper we investigated what if we remove the features from those of the least importance. We did this for the purpose of finding that what features are actually critical for early detection of distributed attacks. From our experiment using traffic data collected by honeypots, we observed that the detection performance is generally getting better if more features are utilized. However, after the number of features has reached around 40, the detection performance will not change much even more features are used. We also found the top-10 important features for detecting C&C traffic. We again verified that some “bad” features would deteriorate the detection performance.

X. FUTURE WORK

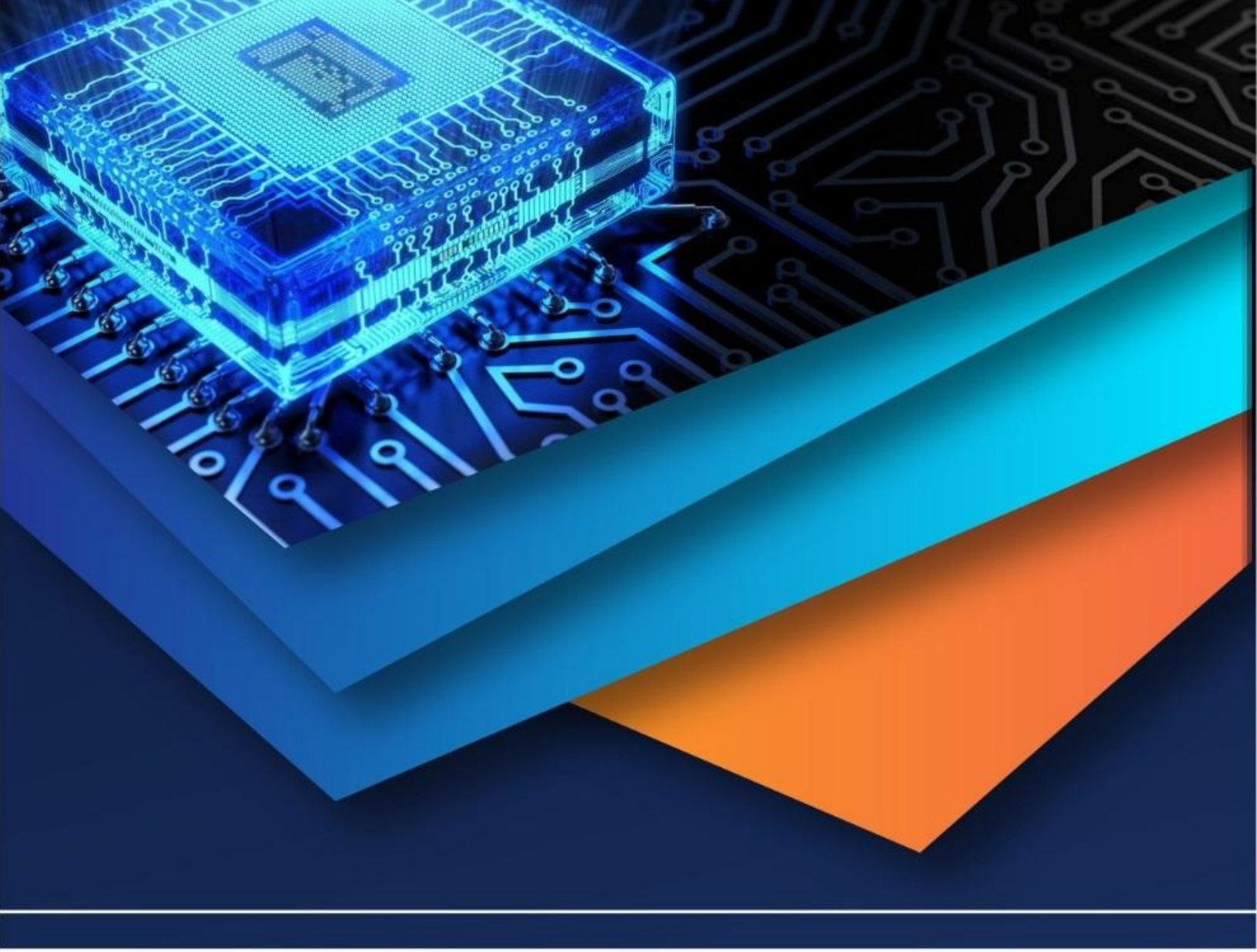
As future work, we will analyze the experiment result in more detail and find the detailed reason why the detection performance changes in such ways. Also, we will verify our observation using other traffic datasets. However, after the number of features has reached around 40, the detection performance will not change much even more features are used. We also found the important features for detecting C&C traffic. We again verified that features would deteriorate the detection performance. That is, some features may have bad influence to the detection performance. Thus, our experiment verified again that the importance of feature selection. It is necessary to investigate in detail why the results are so different for different algorithms of machine learning and feature selection, which will be our future work.

REFERENCES

[1] S. Xu S., ”Collaborative Attack vs. Collaborative Defense,” in Proc. the 4th International Conference on Collaborative Computing (CollaborateCom), pp. 217–228, 2009.
[2] The non-profit anti-spam organization Spamhaus, https://www.spamhaus.org/ (accessed on April 19, 2018).
[3] ”The DDoS That Knocked Spamhaus Offline,” reported by Cloudflare on March 30, 2013. https://blog.cloudflare.com/the-ddos-that-knockedspamhaus-offline-and-ho/ (accessed on April 19, 2018).
[4] ”400Gbps: Winter of Whopping Weekend DDoS Attacks,” reported by Cloudflare on March 3, 2016. https://blog.cloudflare.com/a-winter-of400gbps-weekend-ddos-attacks/ (accessed on April 19, 2018).
[5] ”DDoS Attack Has Varying Impacts on DNS Root Servers,” reported by ThousandEyes on July 19th, 2016. https://blog.thousandeyes.com/ddosattack-varying-impacts-dns-root-servers/ (accessed on April 19, 2018).

[6] Y. Feng, Y. Hori, and K. Sakurai, ”A Proposal for Detecting Distributed Cyber-Attacks Using Automatic Thresholding,” in Proc. the 10th Asia Joint Conference on Information Security (AsiaJCIS), pp. 152–159, 2015.
[7] Y. Feng, Y. Hori, K. Sakurai, J. Takeuchi, ”A Behavior-Based Method for Detecting Distributed Scan Attacks in Darknets,” Journal of Information Processing, Vol.21, No.3, page 527-538, 2013.
[8] Y. Tang, ”Defending against Internet Worms: a Signature-based Approach,” in Proc. the 24th IEEE Annual Joint Conference of the Computer and Communications Societies (INFOCOM), pp. 1384–1394, 2005.
[9] I. Yazid, A. Hanan and M. Aizaini, ”Volume-based Network Intrusion Attacks Detection,” Advanced Computer Network and Security, UTM Press, pp. 147–162, 2008.
[10] A. Kind, M. P. Stoecklin and X. Dimitropoulos, ”Histogram-Based Traffic Anomaly Detection,” IEEE Transactions on Network Service Management, Vol. 6, No. 2, pp. 1–12 (2009).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)