DETAILED ANALYSIS OF CYBER ATTACK DETECTION USING MACHINE LEARNING

D. Joseph Jeyakumar[1]
[1]Professor, Department of Electronics & communication Engineering,
J.N.N Institute of Engineering ,Chennai,TamilNadu ,India.
B. Shanmathi[2]**, M.Priya**[3]**, Senthil Kumar.V**[4]**,R. Nithya**[5]**,D.V. Swathy**[6]
[2,5,6]Assistant Professor, Department of Electronics & communication Engineering,
J.N.N Institute of Engineering ,Chennai ,Tamil Nadu ,India.
[3,4]Assistant Professor, Department of Computer Science and Engineering, J.N.N
Institute of Engineering ,Chennai,Tamil Nadu ,India.
Corresponding Email :reserachphdtk@gmail.com

Abstract

Internet of Things (IoT) is one of the computing disciplines that is growing the fastest, the reality is that IoT is vulnerable to a wide range of attacks in the increasingly hostile online environment. To tackle this, practical countermeasures, like as network anomaly detection, need to be created to protect IoT networks. Attacks can never be totally avoided, but early attack identification is essential for effective protection. Numerous techniques for identifying malware were developed, including network and system techniques. However, expanding detection to a wide range of apps remains a difficulty. The reasons why cyber security matters so much in the constantly developing and fast growing field of cyber security are nearly impossible to quantify or explain. Concerns about cyber security affect not only governments, businesses, and educational institutions operating on the global network or internet, but also people and families. Using machine learning, we will create the cyber security scene.

Keywords: Machine learning, Data security, Cyber attacks, Internet of Things

1. Introduction

Within the cyber realm, cyber-attacks are on the rise. To reduce or eliminate the number of cyber-attacks, improved security procedures should be implemented. DDoS assaults, man-in- the-middle attacks, data espionage, PROBE, User-To-Root, and Remote-To-Local attacks areonly a few examples. Intruders or hackers use these attacks to get illegal entry to any non- public system, websites, data, or even our machines. As a result, internal or external hackers utilize advanced methods or develop ways to irritate or break any defense mechanisms to protect sensitive data and financial data. Effective intrusion weaponry should be able to stop or handle a variety of hacker-created or planned attacks.

The study of technologies, methods, and practices meant to protect networks, devices, programs, and information from assaults, damage, or unauthorized access is known as cyber security. [1] In 2016, there were significant improvements in machine learning approaches such as self-driving cars, linguistic interaction processes, [2]health field, and smart virtual assistants. They must be utilized to locate useful data from a variety of audit datasets that are relevant to intrusion detection.

[3]With the use of machine learning techniques, we will implement these principles in cyber security to improve the IDS defense measures. We'll need to feed the data into the machine- learning algorithm first. [4]The dataset sample trains the model, making it a trained prototype. The next step is to apply and use the machine-learning equation after we've fed thedataset specimen.

This intrusion detection system's protective features are boosted thanks to a machine learning methodology. [5] There are two types of machine learning methodologies: unsupervised and supervised machine learning. They're distinguished by the type of information (i.e., input) they choose. The term "supervised learning" refers to systems that are provided a set of labeled training data[6] and are tasked with figuring out what distinguishes the labels. Learning algorithm relates to systems that are provided unlabeled training data and are tasked with inferring classes on their own. Typically, tagged data is extremely rare, or even the jobof labeling data is demanding in and of itself, and we may not have been ready or ready to seeif labels exist. [7]

The fast growth of the Internet of Things (IoT) in recent years has led to significant advancements in Industry 4.0, smart buildings, and the cloud environment—all of which process sensitive personal data and need to be protected against cybersecurity attacks. (8) Cybersecurity attacks have increased dramatically across a number of industries, including home automation, healthcare, energy, agriculture, robotics, and manufacturing. Because of their wide range of functions, IoT device sensors produce enormous volumes of data, which makes authentication, privacy, and security essential[9]. To ensure IoT security, conventional approaches and standards were previously used. But the popularity of using different artificial intelligence (AI) techniques to detect cybersecurity vulnerabilities has increased recently.

The Internet of Things (IoT) is made up of networked devices that are gradually being constructed on a large scale while considering a variety of aspects through cloud and fog processing, which can enhance the execution of real-time applications. 10] Industry 4.0's interconnected CPSs are the backbone of the smart sector, facilitating increased data transit between networks through development. These consist of IoT for smart cities, big data AI, medical IoT, and IIoT. To address the security issues with the Internet of Things, a number of researchers have developed intrusion detection systems (IDSs) using various AI algorithms. 11]

For instance, [12] offered a distributed service architecture for multidirectional data collection for edge computing enhancement, which encourages the development of security and dependability protection. [13] proposed a deep learning algorithm-based IoT cybersecurity threat detection solution. They compared conventional machine learning techniques with the DL model. [14] recognised AI for the identification of harmful attacks in IIoT and CPS and proposed a hybrid smart classic control technique for modelling attacks on the data input of non-linear CPS using shared systems. As the Internet becomes more intricately entwined with social life, it is revolutionising education and employment practises, but it also poses an increasing threat to our security.

Cybersecurity refers to a combination of technologies and practices that secure computers, networks, programs, and data from assaults, as well as unauthorized access, modification, or destruction [15]. A network system includes both a network and an information security system. Firewalls, antivirus programs, and IDSs are all part of these systems (IDS). IDSs aid in the detection, determination, and identification of illegal system activity including copying,alteration, and deletion.

Internal and external intrusions are examples of security vulnerabilities. Misuse-based also called anomaly-based, signature-based, and hybrid network analysis are the 3 basic forms of network theory for IDSs. The goal of misuse-based detection systems is to identify known threats by analyzing their signatures. [16] They're used to detect recognized forms of attacks without causing a lot of false alarms. Operators, on the other hand, frequently have to manually upgrade system rules and signatures. Based on misunderstood technology, new (zero-day) threats cannot be identified.

2. Methods:

2.1 Model for Comprehensive Cybersecurity Audits

Cyberattacks and cyber threats are persistent and complex problems for private companies and government institutions nowadays. Organizations should construct and promotecybersecurity awareness and culture to protect against cyber thieves, as a general alert. Data Technology, such as IT, and Information Security, such as InfoSec audits, which were once cost-effective, are now attempting to merge into cybersecurity evaluations to deal with cyber attacks, cyber hazards, and cyber threats that develop in an intense cyber landscape. However, the increasing variety and quality of assaults, as well as the more complex cyber risk landscape, are putting current cybersecurity audit models to the test and necessitating the development of a completely new cybersecurity auditing model. This book examines the most basic procedures and methodologies used by world leaders in the field of cybersecurity assurance and auditing. The true breadth, strengths, and limitations of these techniques and theoretical framework are highlighted through a study of their methods and theoretical foundation, to achieve a good and coherent synthesis. As a result, this article proposes an innovative and thorough cybersecurity audit model for use in performing cybersecurity auditsin enterprises and nations. For all structural useful domains, the CSAM assesses and validates auditing, preventative, persuasive, and detective measures. On the Cybersecurity side, CSAM has been verified, implemented, and approved. To validate every concept, a research case analysis is undertaken.

2.2 Feature selection to identify botnets using ML methods

It is granted a unique strategy to try to offer choices to view botnets at their C&C part. One big disadvantage is that scholars have suggested possibilities based on their experience, but there is no method to evaluate these options because some of them may have a weaker detection accuracy than alternatives. To achieve the current goal, we identify the set of features that supports botnet linkages at their C&C section and maximizes the detection accuracy of those botnets. A genetic equation (GA) was familiarised with and selected as the set of alternatives with the highest detection accuracy. We frequently employ the deeplearning equation C4.5, which distinguishes between botnet connections and those that are not. The datasets used in this study were obtained from the ISOT and ISCX sources. The simplest variables in a GA and the equation C4.5 were induced after some tests. We usually conduct experiments together to obtain the easiest set of possibilities for each botnet investigated (specific) and for each type of botnet (generic). The results are presented at the end of the publication, and they indicate a significant reduction in characteristics and agreater detection rate than previous studies.

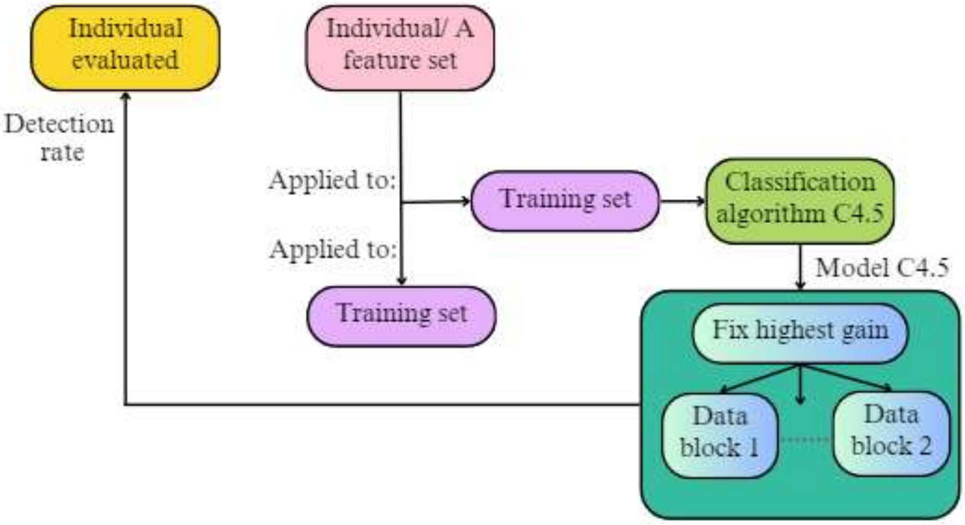


Figure 1 Feature selection to identify botnets using ML methods

2.3 IDS using Deep Belief Network

The challenges that occur in neural intrusion detection systems, such as data redundancy, vastamounts of data, and long-term training, are simple to fall into the locally optimal. The use ofDBN and PNN in an IDS method is suggested. Using DBN's nonlinear wit, the original data will be first recreated to low-dimensional information while retaining the data's key properties. Second, the number of hidden units per layer is optimized using the particle swarm optimization method to get the easiest learning efficiency. After that, PNN is used to categorize the low-dimensional data. Lastly, the KDD CUP 1999 database is used to assess the effectiveness of the aforementioned techniques.

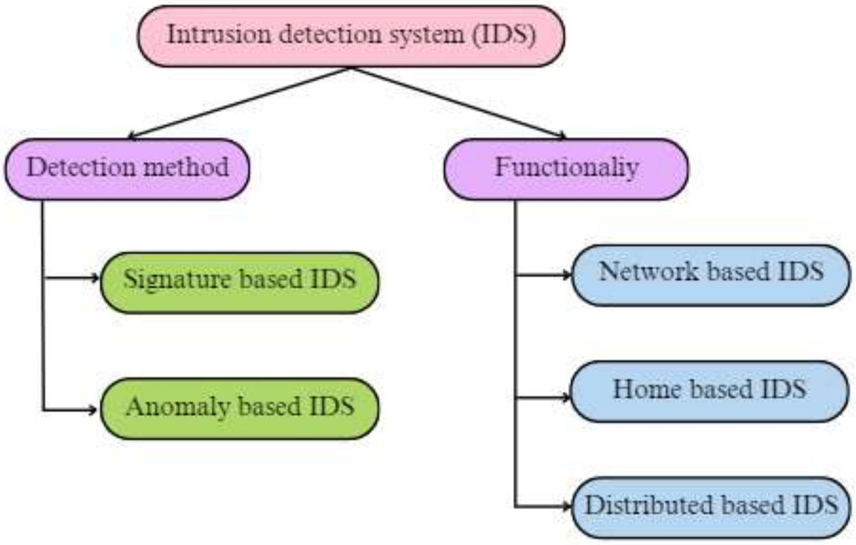


Figure 2 IDS using deep belief network2.4.Impact of ML Techniques on IDS

With the rapid rise of laptop systems and user-generated content consumption, secure and dependable networks are becoming increasingly important. Because it has been found that thenumber of different types of network assaults has increased over time, it is required todevelop a supply of efficient automatic methods to identify attack identification circumstances. One of the assault systems that detect incursions returning from the internet is the IDS. Many methods for intrusion detection systems have been identified in the literature. Mining approaches were popular in the past when it came to visualizing intrusion detection. Using the well-mined information over the info supplied within the system, the features of incoming invasions were determined. An intrusion is proclaimed if a similar object is discovered within the features of the well-mined data. As a result of the current investigation, a variety of IDS algorithms were built to support this criterion, and the accuracy has improved. Over the earlier approaches, a brief inspection is done out. The entire procedure is broken into two parts: information preparation and detection. In addition, the data preprocessing methods are classified into Extraction of features and transformation methods that assist operating methods over the alternatives. Deep-learning and natural process methodologies are also used to classify detection methods.

2.4Increasing the Assurance Model for Cybersecurity

Designing, defining scope and objectives, enlightening terms of interactions, performing the audit, corroboratory evidence, assessing risks, reporting the audit findings, and scheduling follow-up tasks are all phases that a group of auditors will go through every time they participate in an IT, data protection, or audit standards. Creating a cybersecurity audit was similar to designing any other type of audit. This, however, will necessitate a significant amount of effort due to the high quality of several cybersecurity sectors. Most cyber abilities, on the other hand, are not covered by the scope of internal audits. This framework comprises threat monitoring, the advancement life cycle, security program, third-party control, data control, access management, hazard management, the importance of implementing cyber threat restrictions as part of a larger structure and tactic, the need for assurance, that will be achieved through review meetings, cyber risk evaluations, data management and safety, risk analytics, strategic planning, and resilience management.

2.5 Database IDS Using Machine Learning and Octraplet

For host network systems, many intrusion detection solutions have been created. There are, however, only a few important studies in the field of data malware detection. A method was one of the most recent works revealed. This presents a method for detecting information incursion that detects misuse. Here, common information trends have been well-mined and have been preserved as traditional characteristics. The biggest disadvantage was that no job profiles are generated. The users carry out completely distinct actions, which are aided by their responsibilities. User profiles cannot be the sole criterion. Users can carry out actions that are supported by roles, and they will be flagged as malevolent. Time signatures were supported by Lee et alproposed .'s true-time IDS. Temporal data items are used in actual data systems, and their contents change over time. As a result, every time it is upgraded, a device interface is created. Throughout a particular period, temporal data is provided. An alarm is signaled if a deal seeks to change temporal features which have already been modified over that amount. However, this method has the drawback of focusing primarily on updates rather than role profiles. Log files are used by Hu Panda to create user profiles. Hold on to frequently accessed data and tables for comparison. The issue with this technique would be that knowledge management becomes extremely difficult as the size of the data becomes excessively large and the number of users grows rapidly.

3. Result and Discussion3.1.Data Set

It'll show you how to make a permission dataset in this section. Malicious programs are categorized into 177 families, whereas benign applications are grouped. The permission data is translated into a binary format dataset, with 1 indicating that the app requests authorization and 0 indicating that it does not. To create a complete database for data processing, permission files from both hazardous and benign apps are combined.

3.1 Pruning Data at Multiple Levels

1. Using a Negative Rate to Evaluate Permission Rating

First, we use PRNR to generate both the benign and malicious permission ranking lists. After that, we use the PIS to gradually add rights based on the ranking lists. For each phase, two rights from both categories are placed on the major permission list, the outcomes of which areshown in Fig. 3 and 4.

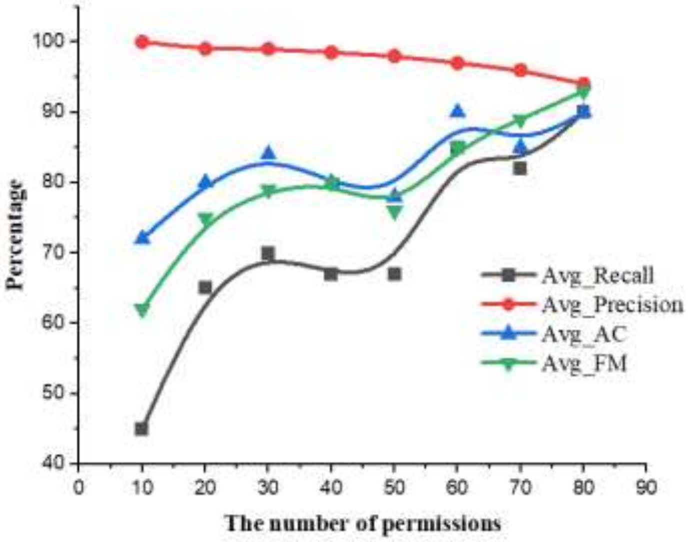


Fig. 3: Permission Incremental System Malware Detection Performance

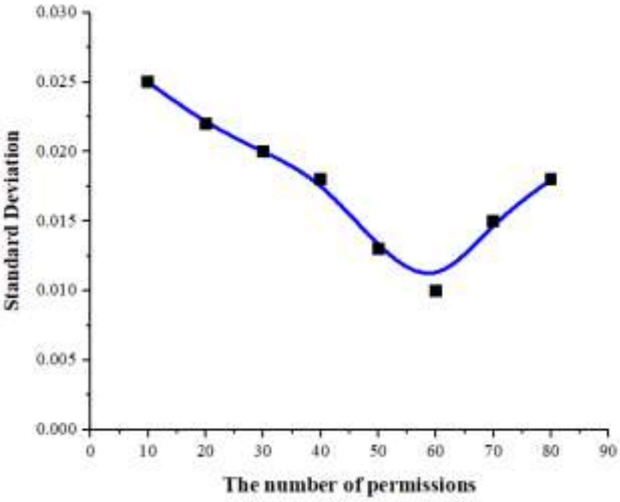


Fig. 4: Standard Error with an Increasing Number of Permissions

Figures 3 and 4 show that improvement occurs in prediction performance, retention, and F-measure with an increase in permissions. As demonstrated in Fig. 3, the precision decreases little with each round but consistently stays over 95%. Recall, accuracy, and F-measure all level after the level of permissions reaches 66, which is an intriguing discovery. Figure 4 also displays the F-measure's standard error. Table 1 displays the results of using 67 ML algorithms.

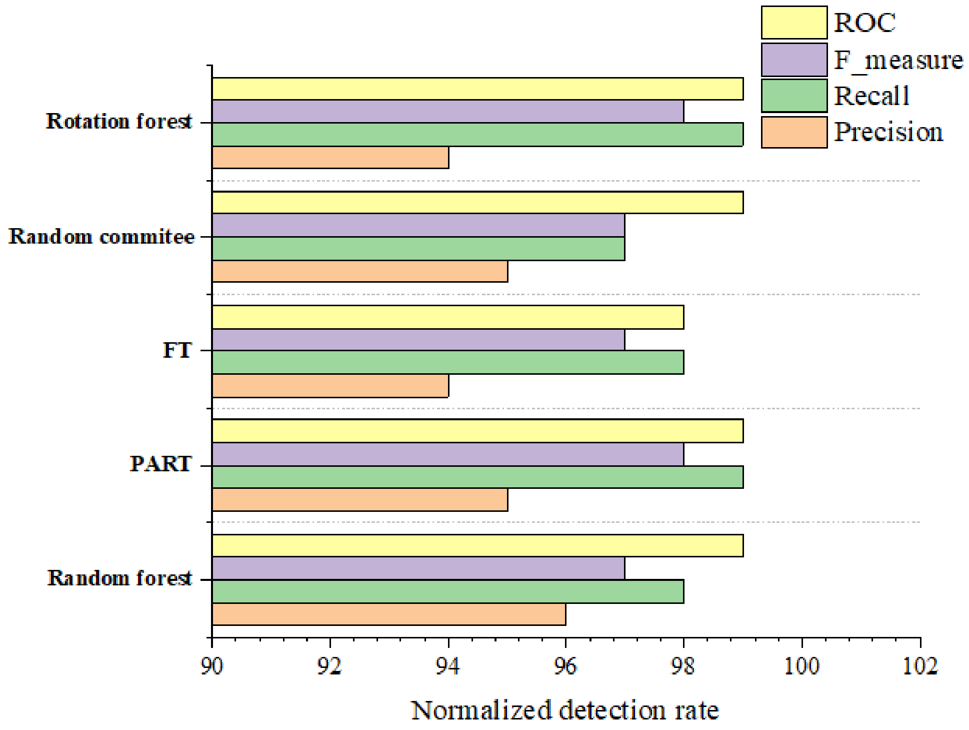
Table 1: Results of Using 67 Machine Learning Algorithms

Number of features	Precision	Recall	F-measure	ROC	Training (seconds)	Testing (seconds)	Total
45	88.23%	99.73%	96.94%	92.34%	1.55%	3.23%	3.56%
144	93.76%	100.89%	93.43%	95.67%	5.34%	23.43%	32.45%

2. Using Association Rules to Evaluate Permission Mining

We analyze the system after implementing the PRNR approach and compare the results to the model with 135 permissions. Both techniques produce almost 90 percent accuracy and F- measure, as shown in Table 3.1.

3. Using Different Machine Learning Algorithms to Evaluate Malware Detection Efficiency



igure 5: Top 5 Machine Learning Algorithms Outcomes

The detection performance with 33 licences is displayed in Figure 5, normalised by the 135 permission model's performance. The malware detection model with a complete access-list outperforms the detection algorithm with 34 privileges in terms of reliability, durability, F-measure, and ROC, but the difference is not very great. With an F-measure of more than 89 percent and other performance metrics of more than 85 percent, Table 3.3 presents the top 5 malware identification models, indicating that the detection approach is still capable of effectively identifying both benign and harmful programmes.

A system with 34 rights takes longer to complete than a version with 135 permissions, as seen in Figure 6. Approximately half the time is spent identifying malware samples using the model with the complete permission list. A dataset with fewer features, however, can save a significant amount of memory.

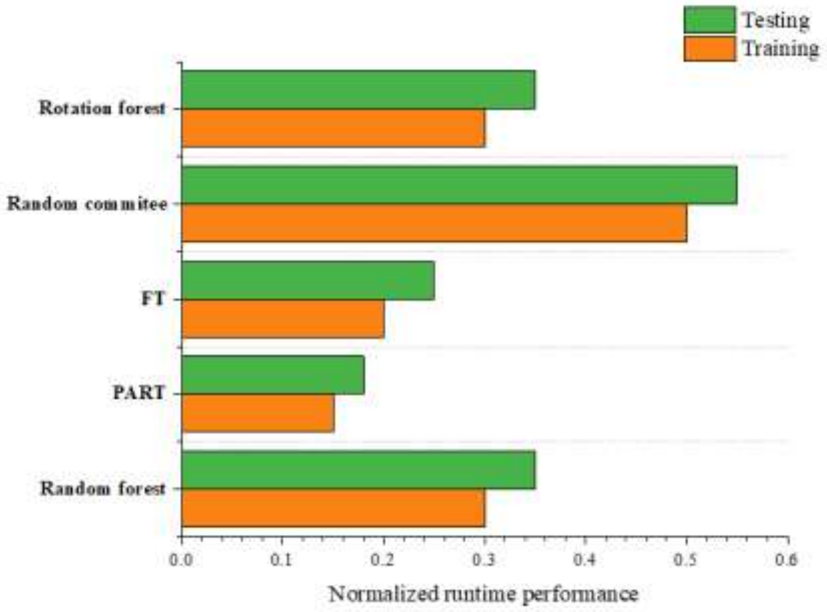


Fig. 6: Top 5 Deep Learning Algorithms' Runtime Efficiency

We also discovered that machine learning approaches based on tree structure can yield better outcomes, depending on the 66 data mining algorithms that were examined.

a. MLDP Improvement

i. MLDP's scalability

We showed that MLDP can work effectively when we used 2,765 malicious apps and a selection of 2,765 benign apps from a corpus of 342,675 benign applications. Of the 135 permissions, 34 were determined to be significant. When these 34 privileges are used, our technology detects malware with a 91.9 percent reliability rate. Our suggested approach fared better in terms of precision and f-measure than the entire authorisation set.

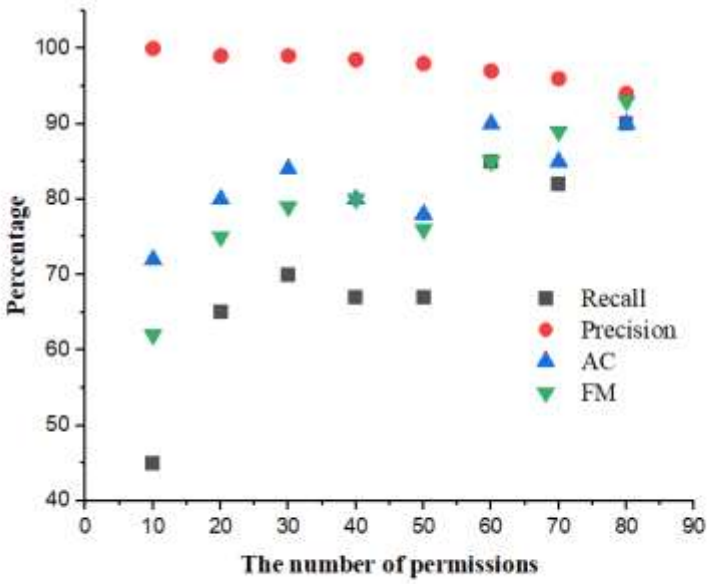


Fig. 7: 1st Step: Efficacy of PIS in Malware Analysis in PRNR

At least 89 rights are needed to create a stable system that depends on PRNR, as seen in Fig. 7. As opposed to the time we used a database containing 2,765 malicious programmes. If so, all we would need is 70 rights to establish a reliable PRNR system. This disparity is predicted as different datasets may have different PIS stopping points. As we finish the SPR, the graph immediately stabilises, as seen in Fig. 8. We only need 15 permissions in order to construct a stable system. Further details about the accuracy and f-measure are given in Table 2. The best f-measure is obtained when we employ 25 authorisation in the second PIS with SPR.

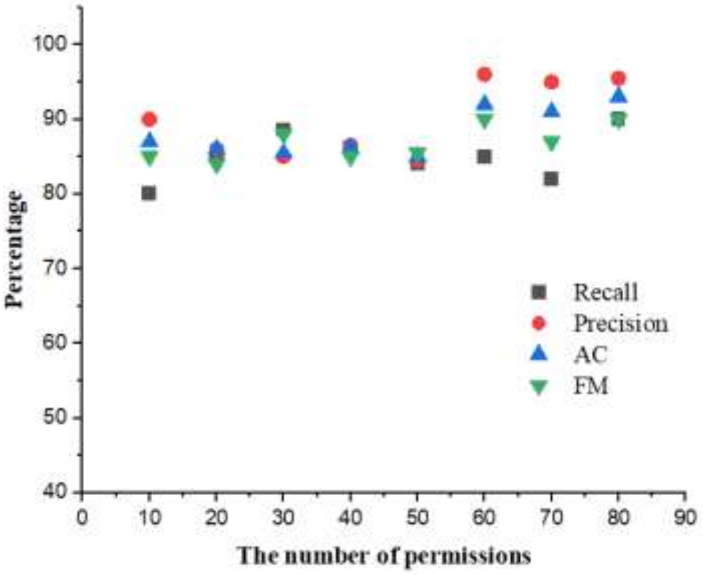


Fig. 8: 2nd Step: Efficacy of PIS in Malware Analysis in SPRTab. 2: Efficacy of PIS in SPR Malware Analysis

Tab. 4: Efficacy of PIS in SPR Malware Analysis

Number of feature	10	20	30	40	50	60
X-Value	3.48666	1.66568	1.97655	1.98998	1.88557	1.68568
FM	102.9	86.56	105.56	94.63	92.69	97.56
Recall	99.36	103.14	92.26	97.59	88.63	99.89
Precision	92.63	89.63	83.36	88.36	83.36	105.36
ACC	98.32	103.56	99.36	101.12	102.21	104.25

4. Conclusion

In this study, we show that it is possible to minimize the number of permissions that need to be reviewed in mobile virus detection while still maintaining extreme accuracy and efficiency. Every method for creating an intrusion detection system has advantages and disadvantages, as evidenced by the comparisons made among the various methods. As a result, choosing one way to deploy an intrusion detection system over the others is difficult. Network-based intrusion detection databases are valuable resources for training and testing systems. The data show that employing an SVM as the classifier may achieve over 90% accuracy, recall, reliability, and F-measure, which are comparable to those obtained using the baseline technique while requiring research that is 4-32 times less expensive than using all rights. To combat this huge cyberattack, we require a modular malware recognition solution that can swiftly identify key malicious programs. Many malware identification methods, such as system- and network methods, were devised. On the other hand, extending detection for a large number of apps continues a challenge.

Reference

1. Karimipour, Hadis, et al. "A deep and scalable unsupervised machine learning system for cyber-attack detection in large-scale smart grids." *IEEE Access* 7 (2019): 80778- 80788.
2. Alrashdi, Ibrahim, et al. "Ad-iot: Anomaly detection of iot cyberattacks in smart city using machine learning." *2019 IEEE 9th Annual Computing and CommunicationWorkshop and Conference (CCWC)*. IEEE, 2019.
3. Avatefipour, Omid, et al. "An intelligent secured framework for cyberattack detection in electric vehicles' CAN bus using machine learning." *IEEE Access* 7 (2019): 127580-127592.
4. Wang, Defu, et al. "Detection of power grid disturbances and cyber-attacks based on machine learning."
   Journal of information security and applications 46 (2019): 42-52.
5. Sultana, Nasrin, et al. "Survey on SDN based network intrusion detection system using machine learning approaches." *Peer-to-Peer Networking and Applications* 12.2 (2019): 493-501.
6. Al-Saud, Mamdooh, et al. "An intelligent data-driven model to secure intravehicle communications based on machine learning." *IEEE Transactions on Industrial Electronics* 67.6 (2019): 5112-5119.
7. Tang, Zefan, et al. "Enabling cyberattack-resilient load forecasting through adversarial machine learning." *2019 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2019.

8. Dogaru, Delia Ioana, and Ioan Dumitrache. "Cyber security of smart grids in the context of big data and machine learning." *2019 22nd International Conference on Control Systems and Computer Science (CSCS)*. IEEE, 2019.

9. Mongelli, M., M. Muselli, and E. Ferrari. "Achieving zero collision probability in vehicle platooning under cyber attacks via machine learning." *2019 4th international conference on system reliability and safety (ICSRS)*. IEEE, 2019.

10. Kadoguchi, Masashi, et al. "Exploring the dark web for cyber threat intelligence usingmachine leaning."
2019 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2019.

11. Kavousi-Fard, Abdollah, Wencong Su, and Tao Jin. "A machine-learning-based cyberattack detection model for wireless sensor networks in microgrids." *IEEE Transactions on Industrial Informatics* 17.1 (2020): 650-658.

12. Deorankar, Anil V., and Shiwani S. Thakare. "Survey on anomaly detection of (iot)- internet of things cyberattacks using machine learning." *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2020.

13. Bland, John A., et al. "Machine learning cyberattack and defense strategies." *Computers & security* 92 (2020): 101738.

14. Sarker, Iqbal H., et al. "Cybersecurity data science: an overview from machine learning perspective."
Journal of Big data 7.1 (2020): 1-29.

15. Guo, Lulu, Jin Ye, and Bowen Yang. "Cyberattack Detection for Electric Vehicles Using Physics-Guided Machine Learning." *IEEE Transactions on Transportation Electrification* 7.3 (2020):

16. Bout, Emilie, Valeria Loscri, and Antoine Gallais. "How Machine Learning changes the nature of cyberattacks on IoT networks: A survey." *IEEE Communications Surveys & Tutorials* (2021).

17. AlZubi, Ahmad Ali, Mohammed Al-Maitah, and Abdulaziz Alarifi. "Cyber-attack detection in healthcare using cyber-physical system and machine learning techniques." *Soft Computing* 25.18 (2021): 12319-12332.