RESEARCH PAPER

MALWARE DETECTION USING MACHINE LEARNING PARADIGM

Dhairya bhardwaj
(dhairyabhardwaj007@gmail.com)

Vimal Maurya
(vimal.21scse1011714@galgotiasuniversity.edu.in)

Abhijeet kumar
(Abhijeet.21scse1010930@galgotiasuniversity.edu.in)

GALGOTIAS UNIVERSITY
Plot No.2, Sector -17 A, Yamuna Expressway, Greater Noida, Gautam Buddh Nagar, U.P., India

SCHOOL OF COMPUTING SCIENCE & ENGINEERING

ABSTRACT

Nowadays there are lots of issues faced by internet users, and one of them is viruses and malware issue. In this upcoming generation there are lots of new malware in market that cause lots of harm to internet devices and users' data. One of them is Polymorphic malware, it is a type of new malicious software that is more adaptable than any other previous versions of that malwares. It modifies its signature traits constantly to escape itself to being identified. To identify this type of malware, we use various types of machine learning algorithms. We select that algorithm that gives best and highest accuracy to be used in the system. When we are identified particular malware on the computer system or networks, and hence providing the security to computer networks by different approaches. There are lots of algorithms and techniques we can use in malware detection like DT, CNN and SVM algorithms. The results showed with our dataset and used following techniques to detect malware are :- Random Forest, Decision Tree, Ada boost and Linear Regression.

With accuracy of following techniques: -

- Decision Tree: 99.08728721477725
- Random Forest: 99.41325606664252
- Ada boost: 98.54762767113364
- Linear Regression :60.5760783194729

These outputs are correctly significant, as we know nowadays malicious soft-ware in internet world and day by day it's becoming common & complex.

INTRODUCTION

Due to increasing internet users, there is lots of sensitive information and data that we need to secure. Nowadays, cyberattacks are most significantly increasing concern in current cyber world. One of them is Malware attack. It is a type of software that develop by hackers by some set of instructions or program to harm a particular computer, system, organization and any other business. Malwares are of different types like Trojans horses, ransomware, spyware, spyware, adware, rogue software, wipers, scareware, and so on. It's a type of software that runs on others system without their permission or consent. In this study, we demonstrated the harmful malware that corrupt the files and detect them on systems, and by that we improve the security of computer system and networks. Here, we use lots of machine learning techniques to compute the differences with all methods accuracy and proposed approach. Modules of Malware detection are responsible for collecting data and analyses it for train such that it needs to determine whether it has specific malware or not, it can harm our system or not for security concern. Algorithms that have ability or trained by means of machine learning, they can improve or enhance their ability to predict the wellbeing of that how they performed previously and make proper changes further. Worldwide, cyberattacks has now become the serious concern for business, sensitive data of any organization that may steal the confidential information of particular system & harm them. Even, we notice every day we receive the fraud calls, messages and mails on our devices (i.e., smartphones, laptop etc.). Fraudsters uses lots of harmful software in attempt to get access of private networks or systems to transfer money, steal useful information and harm system. To keep safe networks and system from such fraudsters has become urgency for all business owners, private organization and government entities. For that, we require some methods and techniques of machine learning. In this research paper, our aim to discover and detailed info about detecting the malware and techniques to protect the private information from hackers by means of data mining and machine learning. We analyze signature-based features to enhance the malware detection and classification. Our experiments have shown proven results on some machine learning algorithms. New Generation malware has become increasingly common and major threat complex, as it uses constantly modification for not to be identified. It has become threat to security of modern websites.

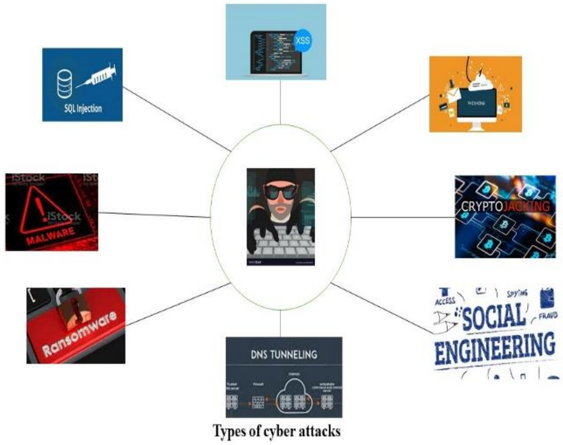


Figure 1. Types of cyberattacks.

Figure 1 below depicts the types of cyberattacks in this digital world or cyberworld. Malware is a software developed by hackers or fraudsters with aim to harm other's networks and system without their consent. It has been increasing in different sectors like medical gear, IOT devices and environmental and industrial systems.

Currently, Malware in the cyber is very hard to detect as it constantly updates its behavior. To overcome this situation and tackle this, we require wider range of defensive actions. Here, we use can use both static and dynamic learning methods to identify behavior relations between members of malware present. In static analysis, it examines harmful files without actually running them on system. But dynamic analysis first takes account of tracking data flows, checking functions work, and take action to monitor their actions. Machine learning algorithms helps in static and behavior of particular malware, to identify the structure of malware, to increasing identify complex malware that can use signature-based method. But we can't rely singly on signature-based techniques of machine learning. Nowadays we have lots of successful malware detection techniques to detect malware.

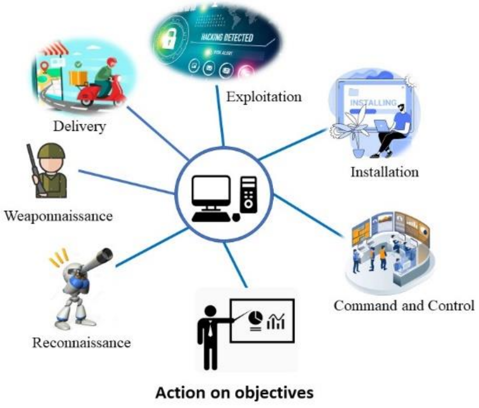


Figure 2. Martin cybers kill chain for prevention of cyber intrusion activity

Figure 2 illustrates the Martin cyber kill chain for cyberattack protection and in measure with security concern. In Feb 2020, AWS was the target of large-scale denial cyberattack. That attack on AWS was of 2.3 Tbsp. DDoS Attack which have transfer rate of 293.1 Mpps packet forwarding and request rate of 694,201. Some expert claimed that attack is largest known DDoS attack. Also remember the cyberattack on Twitter in July 2020, twitter faced cyberattacks by 3 hackers, they gained access of accounts of some Great Personalities: - Ex- President of USA Obama, Amazon's Jeff Bezos, and Tesla's CEO Elon Musk. There are thousands of attacks happen daily in this Digital world. After 2 weeks of that attack, US Justice Department filed case charge of those 3 individuals, the youngest of them was at 17 at that time. He has made one another attack at Marriott's Starwood Hotels which disclose the personal information of 500 million customers in 2018. These attacks can be geopolitical as Russia made an attack on Ukrainian electricity infrastructure in 2017. This showed Russian hackers' ability to execute the large scale cyberattacks first time. The Russian cyber military unit Sandworm launched the attack on command center that attack made possible Russian to seize the control of substation's computer systems. This attack estimated ultimately affect people between 200,000 and 300,000.

LITERATURE REVIEW

The increasing number of computers, smartphones and other Internet-enabled devices gets more cyber-attack and more common threat. A large number of malware detection methods are made to tackle the issue of malware attacks activity. Common machine learning -based malware detection methods take much processing time, but help in merging new malware. Feature engineering may be less perspective to increasing demand of machine learning algorithms, such as deep learning. In this research, we have experiment various malware detection techniques.

Researchers has developed the ways to check data samples by machine learning nad deep learning.

Armaan (2021) described and tested the accuracy of many techniques' models. As we all know that for an application to function there must have data to operate. At this time, the proliferation of malicious software program poses a sizeable chance to worldwide stability. In the 1990s, because the quantity of interconnected computer systems exploded, so did the superiority of malicious software program [23], which in the end brought about the widespread distribution of malware. Multiple protecting measures had been created in reaction to this phenomenon. Unfortunately, present day safeguards can't maintain up with current threats that malware authors have created to thwart safety programs. In current years, researchers' cognizance on malware detection studies has shifted in the direction of ML set of rules strategies. In this studies paper, we gift a protecting mechanism that evaluates 3 ML set of rules processes to malware detection and chooses the maximum suitable one.

S.No	Specs	Score
1	ref-cycles	1.296660e+07
2	stalled-cycles-backend-percent	3.017643e+06
3	bus-cycle	1.494372e+06
4	stalled-cycles-frontend-percent	9.583614e+05
5	cache-references	1.233573e+05
6	Instructions-per-cycle	6.668084e+04
7	cache-misses-percent	9.394599e+03
8	bracnhes	1.923075e+07
9	page-faults	3.900794e+03
10	Branch-misses-percent	2.306111e+03

Table:1

Chowdhury (2018) proposed a new malware detection method that uses machine learning classification technique. N-gram and API call capabilities were prior into this approach. Experimental evaluation confirmed the efficiency and dependability of this proposed technique. they are also focusing on merging all possible features to increase detection while decreasing negatives. Experimental results by Chowdhury approach are shown in table below. his approach was clearly clever.

classifier	Accuracy
Decision Tree	99.08%
Random Forest	99.41%
Ada boost	98.54%
Linear Regression	60.57%

Table:2

Table. Classifiers results comparisons.

Methods Accuracy (%) TPR (%) FPR (%)

KNN 95.02 96.17 3.42

CNN 98.76 99.22 3.97

Naïve Byes 89.71 90 13

Random Forest 92.01 95.9 6.5

SVM 96.41 98 4.63

DT 99 99.07 2.01

The technique wishes a workaround this is adaptable sufficient to address non-widespread data. To successfully control and save you destiny assaults, we need to examine malware and create new policies and styles withinside the shape of advent of malware kind. To locate styles, IT safety professionals may also use malware evaluation equipment. The availability of technology that examine malware samples and decide their degree of malignancy appreciably gain the cybersecurity sector. This equipment assists screen safety signals and save you malware attacks. If malware is dangerous, we need to put off it earlier than it transmits its contamination any further. Malware evaluation is turning into an increasing number of famous because it facilitates corporations reduce the consequences of the developing quantity of malware threats and the growing complexity of the approach's malware may be used to attack [22].

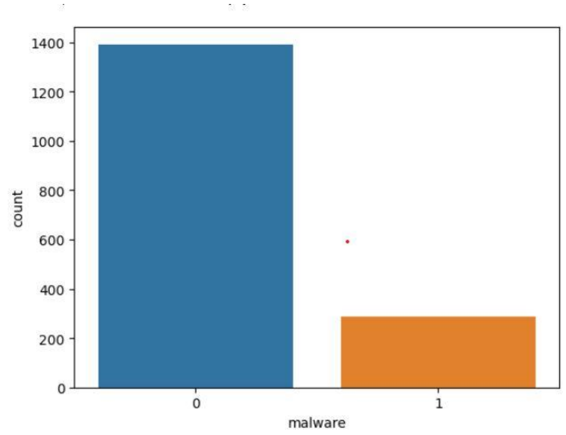


Figure.3 Malware detection graph.

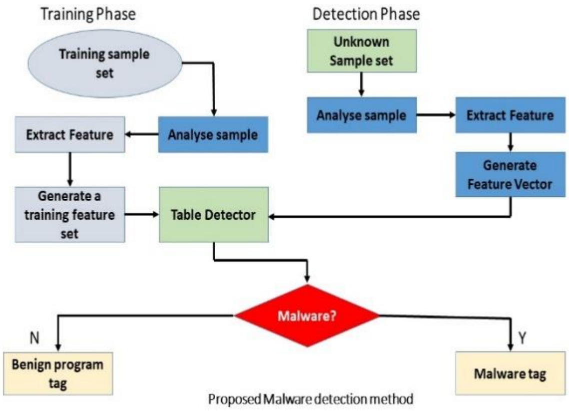
Malware continues to develop and flows at an increasing rate. Nur (2019) compared 3 classifiers techniques to analyze and check the accuracy of the ML classifier as it uses static analysis to extract features based on PE information. Malicious programs and their threats, or singly "malware" became increasingly common and internet continues to emerged. Their rapid growth over the internet has provided hackers or fraudsters with permission to a wide number of malware generation tools. Every day, malware reach increasing. Our study focused on machine learning techniques how they classify and works. It's recommended that ML systems be trained and tested to check a file has malware or not. Experimental results verifies that the random forest technique is preferred for data classify, has an accuracy of 99.41%. The main advantage is that user install a file or software as it will be checked before opening it and maintain validity.

RESEARCH PROBLEM

We can detect the malware potential using static and dynamic analysis. In Static analysis, first a virus is disassemble using reverse engineering method, that focused on breaking malware binaries to find harmful strings. Both methods have their own advantages and disadvantages; however, both are used of analyzing malware, both are best to use. and after reducing the number of dangerous features increase the accuracy of malware detection. Now the researcher has more time to collect data and analyze it. But we have a concern most of characteristics are used to detect malware whereas fewer can also work same on it. The manner of selecting which malicious functions to enforce starts off evolved with coming across feasible methods or algorithms. We want answers which could each locate malware that has in no way been seen earlier than and substantially lessen the wide variety of traits which might be presently had to do so.

METHOLOGY

This research paper introduces many steps and components of machine learning flowchart for malware detection and classification, checks the challenges and have some limitations in its workflow and uses the latest innovations and trends in the field on deep learning techniques. Proposed research method of this study is stated below. To understand the complete concept of this proposed machine learning techniques for malware detection is depicted in figure below. has full flowchart start to finish.



Proposed Malware detection method

Figure 4. Proposed ML malware detection method

- **Dataset**

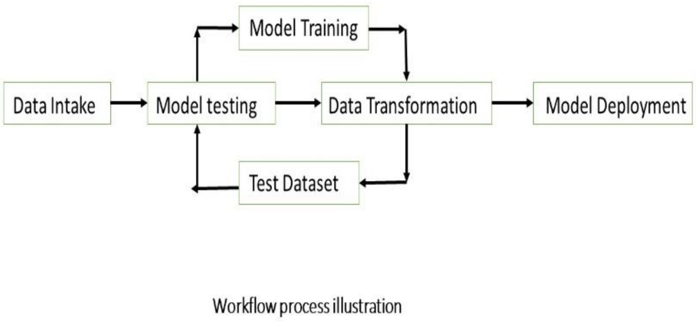


Figure 5. Workflow process illustration

- **Pre-processing**

Data were stored in the file system as a binary code and files were unprocessed. As they are prepared in advance of our research. When we up pack the data flees, we require a protected environment like virtual machine (VM). Paid software which automatically unpack the compressed executables.

- **Features Extraction**

Current generation datasets contain thousands of features. In recent years, features of datasets have grown, now it is required to innovate new machine learning algorithms as previous has become overfit. For this we develop or select fewer smaller set of features from large sets of features, it also works well to maintain accuracy. This research paper is to redefine the current dataset of both static and dynamic features, keep them which are helpful and eliminate that were not valuable for data analysis.

- **Feature Selection**

Feature selection is performed after the feature extraction process which helps in discovery of more features. Feature selection has many advantages as it enhance the accuracy, simplify the model, and reduce overfitting, it's because it recognizes new features from a pool. As we know that Researchers bused many techniques in order to identify harmful code in software. But now feature selection technique is extremely used, also in this study. This technique is very effective at picking the right features for building malware detection models.

CONCLUSION

This paper shows that academician have contemporary shown increasing interest in machine learning algorithms solution for malware identification. We purposed three protective methods to detect malware and the most accurate one. The results show that compared with other classifiers, DT (99.08%), RF (99.41%), Ad boost (98.54%) and Linear regression (60.57%) performed well in term of detection accuracy. In this experiment, we quantify the Machine learning detection accuracy of classifiers which have the highest accuracy by comparison of other Machine learning (ML) classifiers. As a result of our experimental effort Machine learning algorithms can identify the different malwares now. And we have seen that Decision Tree (DT) have the (99.41%) of accuracy which is highest in comparison of all other classifiers we have evaluated. Potentially to provide the accurate and highest accuracy of malware detection based on our cautious selected dataset have shown assurance in our experimental findings. And we have four Machine Learning Models (Decision Tree, Random Forest, Ad boost, Linear Regression) which were trained and checked their efficiency using the given dataset.

REFERENCE

1. Nikam, U.V.; Deshmuh, V.M. Performance evaluation of machine learning classifiers in malware detection. In Proceedings of the 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 23–24 April 2022; pp. 1–5. [CrossRef]
2. Akhtar, M.S.; Feng, T. IOTA based anomaly detection machine learning in mobile sensing. EAI Endorsed Trans. Create. Tech. 2022, 9, 172814. [CrossRef]
3. Sethi, K.; Kumar, R.; Sethi, L.; Bera, P.; Patra, P.K. A novel machine learning based malware detection and classification framework. In Proceedings of the 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Oxford, UK, 3–4 June 2019; pp. 1–13.
4. Abdulbasit, A.; Darem, F.A.G.; Al-Hashmi, A.A.; Abawajy, J.H.; Alanazi, S.M.; Al-Rezami, A.Y. An adaptive behavioral-based increamental batch learning malware variants detection model using concept drift detection and sequential deep learning. IEEE Access 2021, 9, 97180–97196. [CrossRef]
5. Feng, T.; Akhtar, M.S.; Zhang, J. The future of artificial intelligence in cybersecurity: A comprehensive survey. EAI Endorsed Trans. Create. Tech. 2021, 8, 170285. [CrossRef]
6. Sharma, S.; Krishna, C.R.; Sahay, S.K. Detection of advanced malware by machine learning techniques. In Proceedings of the SoCTA 2017, Jhansi, India, 22–24 December 2017.
7. Chandrakala, D.; Sait, A.; Kiruthika, J.; Nivetha, R. Detection and classification of malware. In Proceedings of the 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 8–9 October 2021; pp. 1–3. [CrossRef]
8. Zhao, K.; Zhang, D.; Su, X.; Li, W. Fest: A feature extraction and selection tool for android malware detection. In Proceedings of the 2015 IEEE Symposium on Computers and Communication (ISCC), Larnaca, Cyprus, 6–9 July 2015; pp. 714–720.

9. Akhtar, M.S.; Feng, T. Detection of sleep paralysis by using IoT based device and its relationship between sleep paralysis and sleep quality. EAI Endorsed Trans. Internet Things 2022, 8, e4. [CrossRef]
10. Gibert, D.; Mateu, C.; Planes, J.; Vicens, R. Using convolutional neural networks for classification of malware represented as images. J. Comput. Virol. Hacking Tech. 2019, 15, 15–28. [CrossRef]
11. Firdaus, A.; Anuar, N.B.; Karim, A.; Faizal, M.; Razak, A. Discovering optimal features using static analysis and a genetic search based method for Android malware detection. Front. Inf. Technol. Electron. Eng. 2018, 19, 712–736. [CrossRef]
12. Dahl, G.E.; Stokes, J.W.; Deng, L.; Yu, D.; Research, M. Large-scale Malware Classification Using Random Projections And Neural Networks. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing-1988, Vancouver, BC, Canada, 26–31 May 2013; pp. 3422–3426.
13. Akhtar, M.S.; Feng, T. An overview of the applications of artificial intelligence in cybersecurity. EAI Endorsed Trans. Create. Tech. 2021, 8, e4. [CrossRef]
14. Akhtar, M.S.; Feng, T. A systemic security and privacy review: Attacks and prevention mechanisms over IOT layers. EAI Endorsed Trans. Secur. Saf. 2022, 8, e5. [CrossRef]
15. Anderson, B.; Storlie, C.; Lane, T. "Improving Malware Classification: Bridging the Static/Dynamic Gap. In Proceedings of the 5th ACM Workshop on Security and Artificial Intelligence (AISec), Raleigh, NC, USA, 19 October 2012; pp. 3–14.
16. Varma, P.R.K.; Raj, K.P.; Raju, K.V.S. Android mobile security by detecting and classification of malware based on permissions using machine learning algorithms. In Proceedings of the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 10–11 February 2017; pp. 294–299.
17. Akhtar, M.S.; Feng, T. Comparison of classification model for the detection of cyber-attack using ensemble learning models. EAI Endorsed Trans. Scalable Inf. Syst. 2022, 9, 17329. [CrossRef]
18. Rosmansyah, W.Y.; Dabarsyah, B. Malware detection on Android smartphones using API class and machine learning. In Proceedings of the 2015 International Conference on Electrical Engineering and Informatics (ICEEI), Denpasar, Indonesia, 10–11 August 2015; pp. 294–297.
19. Tahtaci, B.; Canbay, B. Android Malware Detection Using Machine Learning. In Proceedings of the 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 15–17 October 2020; pp. 1–6.
20. Baset, M. Machine Learning for Malware Detection. Master's Dissertation, Heriot Watt University, Edinburg, Scotland, December 2016. [CrossRef]
21. Akhtar, M.S.; Feng, T. Deep learning-based framework for the detection of cyberattack using feature engineering.

Secur. Commun. Netw. 2021, 2021, 6129210. [CrossRef]

22. Altaher, A. Classification of android malware applications using feature selection and classification algorithms. VAWKUM Trans. Comput. Sci. 2016, 10, 1. [CrossRef]
23. Chowdhury, M.; Rahman, A.; Islam, R. Malware Analysis and Detection Using Data Mining and Machine Learning Classification; AISC: Chicago, IL, USA, 2017; pp. 266–274.
24. Patil, R.; Deng, W. Malware Analysis using Machine Learning and Deep Learning techniques. In Proceedings of the 2020 SoutheastCon, Raleigh, NC, USA, 28–29 March 2020; pp. 1–7.
25. Gavrilu¸t, D.; Cimpoesu, M.; Anton, D.; Ciortuz, L. Malware detection using machine learning. In Proceedings of the 2009 International Multiconference on Computer Science and Information Technology, Mragowo, Poland, 12–14 October 2009; pp. 735–741.
26. Pavithra, J.; Josephin, F.J.S. Analyzing various machine learning algorithms for the classification of malwares. IOP Conf. Ser. Mater. Sci. Eng. 2020, 993, 012099. [CrossRef]
27. Vanjire, S.; Lakshmi, M. Behavior-Based Malware Detection System Approach For Mobile Security Using Machine Learning. In Proceedings of the 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), Gandhinagar, India, 24–26 September 2021; pp. 1–4.
28. Agarkar, S.; Ghosh, S. Malware detection & classification using machine learning. In Proceedings of the 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), Gunupur Odisha, India, 16–17 December 2020; pp. 1–6.
29. Sethi, K.; Chaudhary, S.K.; Tripathy, B.K.; Bera, P. A novel malware analysis for malware detection and classification using machine learning algorithms. In Proceedings of the 10th International Conference on Security of Information and Networks, Jaipur, India, 13–15 October 2017; pp. 107–113.
30. Ahmadi, M.; Ulyanov, D.; Semenov, S.; Trofimov, M.; Giacinto, G. Novel feature ex-traction, selection and fusion for effective malware family classification. In Proceedings of the sixth ACM conference on data and application security and privacy, New Orleans, LA, USA, 9–11 March 2016; pp. 183–194.
31. Damshenas, M.; Dehghantanha, A.; Mahmoud, R. A survey on malware propagation, analysis and detec-tion. Int. J. Cyber-Secur. Digit. Forensics 2013, 2, 10–29.
32. Saad, S.; Briguglio, W.; Elmiligi, H. The curious case of machine learning in malware detection. arXiv 2019, arXiv:1905.07573.
33. Selamat, N.; Ali, F. Comparison of malware detection techniques using machine learning algorithm. Indones. J. Electr. Eng. Comput. Sci. 2019, 16, 435. [CrossRef]
34. Firdausi, I.; Lim, C.; Erwin, A.; Nugroho, A. Analysis of machine learning techniques used in behavior-based malware detection. In Proceedings of the 2010 Second International Conference on

Advances in Computing, Control, and Telecommunication Technologies, Jakarta, Indonesia, 2–3 December 2010; pp. 201–203. [CrossRef]

35. Hamid, F. Enhancing malware detection with static analysis using machine learning. Int. J. Res. Appl. Sci. Eng. Technol. 2019, 7, 38–42. [CrossRef]
36. Prabhat, K.; Gupta, G.P.; Tripathi, R. TP2SF: A trustworthy privacy-preserving secured framework for sustainable smart cities by leveraging blockchain and machine learning. J. Syst. Archit. 2021, 115, 101954.
37. Kumar, P.; Gupta, G.P.; Tripathi, R. A distributed ensemble design based intrusion detection system using fog computing to protect the internet of things networks. J. Ambient Intell. Human. Comput. 2021, 12, 9555–9572. [CrossRef]
38. Prabhat, K.; Gupta, G.P.; Tripathi, R. Design of anomaly-based intrusion detection system using fog computing for IoT network. Aut. Control Comp. Sci. 2021, 55, 137–147. [CrossRef]
39. Prabhat, K.; Tripathi, R.; Gupta, G.P. P2IDF: A Privacy-preserving based intrusion detection framework for software defined Internet of Things-Fog (SDIoT-Fog). In Proceedings of the Adjunct Proceedings of the 2021 International Conference on Distributed Computing and Networking (ICDCN '21), Nara, Japan, 5–8 January 2021; pp. 37–42. [CrossRef]
40. Kumar, P.; Gupta, G.P.; Tripathi, R. PEFL: Deep privacy-encoding-based federated learning framework for smart agriculture. IEEE Micro 2022, 42, 33–40. [CrossRef]
41. Akhtar, M.S.; Feng, T. Malware Analysis and Detection Using Machine Learning Algorithms. Symmetry 2022, 14, 2304