Mesopotamian journal of Cybersecurity
Vol.2022, **pp**. 47–56
DOI: https://doi.org/10.58496/MJCS/2022/006; ISSN: 2958-6542
https://mesopotamian.press/journals/index.php/CyberSecurity

Review Article

Machine Learning and Data Mining Methods for Cyber Security: A Survey

Ziaul Hasan[1,*], Hassan r. Mohammad [2], Maka Jishkariani [3]

[1] *Department of Biosciences, Jamia Millia Islamia, New Delhi-110025, India*

[2] *Al-Ahliyya Amman University, Jordan*

[3] *Full Professor of Georgian Technical University, Faculty of Power Engineering, Department of Electro Energy and Electro Mechanic, Tbilisi, Georgia.*

ARTICLE INFO

Article History

Received 20 Aug 2022
Accepted 21 Oct 2022
Published 25 Nov. 2022

Keywords

Cyber-attack,
Cyber Threats,
Data Mining,
Machine learning,
Malware,
Ransomware

ABSTRACT

Data mining and machine learning (ML) methods are used more than ever in cyber security. The use of machine learning (ML) is one of the potential solutions that may be successful against zero-day attacks, starting with categorising IP traffic and filtering harmful traffic for intrusion detection. In this field, certain published systematic reviews were taken into consideration. Recent systematic reviews may incorporate older and more recent works in the topic of investigation.. Both security professionals and hackers use data mining capabilities. Applications for data mining may be used to analyze programme activity, surfing patterns, and other factors to identify potential cyber-attacks in the future. The new study uses statistical traffic features, ML, and data mining approaches. This research performs a concentrated literature review on machine learning and its usage in cyber analytics for email filtering, traffic categorization, and intrusion detection. Each approach was identified, and a summary was provided based on the relevancy and quantity of citations. Some well-known datasets are also discussed since they are a crucial component of ML techniques. On when to utilize a certain algorithm is also offered some advice. Four ML algorithms have been evaluated on MODBUS data gathered from a gas pipeline. Using ML algorithms, other assaults have been categorized, and then the effectiveness of each approach has been evaluated. This study demonstrates the use of ML and data mining for threat research and detection, focusing on malware detection with high accuracy and short detection times.

1. INTRODUCTION

The investigation of cyber security, attack types, network weaknesses, cyber risks, and components for malware discovery using data mining strategies are the fundamental foci of the examination. The outline of machine learning (ML) and data mining (DM) methods for cyber security applications is introduced in this review. The ML/DM procedures and different uses of every method to cyber interruption recognition issues are talked about. As indicated by the elements of the cyber issue to be tackled, the article offers a bunch of correlation measures for ML/DM strategies, a bunch of ideas on the best techniques to apply, and a conversation of the intricacy of different ML/DM calculations. Cyber security is the assortment of advances and techniques made to safeguard against attacks, unapproved access, change, and annihilation of PCs, organizations, developers, and data. Network security frameworks and PC (have) security frameworks make up cyber security frameworks. Every one of them has a firewall, antivirus program, and intrusion detection system, in any event (IDS). IDSs help find, determine, and distinguish data framework unlawful use, copy, change, and annihilation [1]. Attacks from outside the organization (outer interruptions) and inward interruptions are among the security slips (attacks from inside the association).Fig 1 shows the variables of Cybersecurity.

*Corresponding author. Email: zhasan.biochem@gmail.com

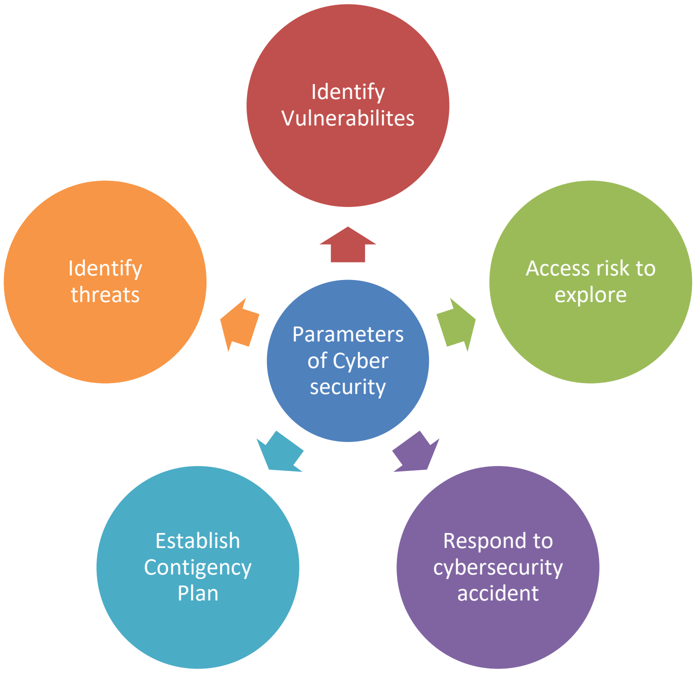
Identify Vulnerabilites
Access risk to explore
Parameters of Cyber security
Identify threats
Establish Contigency Plan
Respond to cybersecurity accident

Fig 1 Parameters of Cyber security [11]

The three primary forms of cyber analytics that underpin IDSs are hybrid, anomaly-based, and misuse-based (sometimes referred to as signature-based). Misuse-based approaches are intended to identify known attacks by using their signatures. They efficiently identify assaults of a recognized kind while producing a manageable quantity of false alarms. They need regular manual modifications to the rules and signatures in the database. Novel (zero-day) attacks cannot be found using misuse-based approaches. Anomaly-based approaches simulate the typical behavior of networks and systems and discover anomalies as departures from this pattern. They are desirable because they can identify zero-day attacks. Another benefit is that each system, application, or network has unique typical activity profiles, making it difficult for attackers to figure out which actions they may do covertly. Additionally, it is possible to create the signatures for abuse detectors using the data that anomaly-based approaches (new attacks) warn on. Because previously unknown (but valid) system behaviors may be labeled as anomalies, the fundamental drawback of anomaly-based approaches is the possibility for large false alarm rates (FARs). Hybrid methods integrate anomaly and misuse detection. They are used to increase known intrusion detection rates and lower known attack false positive (FP) rates. Few pure anomaly detection techniques were found in a thorough search of the literature; instead, most of the techniques were hybrid. As a result, anomaly detection and hybrid approaches are discussed jointly in the explanations of ML and DM methods. IDSs are also divided into network-based and host-based categories according to where they search for invasive activity. A network-based ID detects intrusions by keeping track on network device traffic. A host-based IDS keeps track of file and process activity relating to the software environment connected to a particular host. This comprehensive study focuses on ML and DM methodologies for cyber security, with a focus on their descriptions.

2. OVERVIEW IN MACHINE LEARNING AND DEEP LEARNING

A data analysis technique known as machine learning (ML) finds insights in the data without explicit programming on where to seek by automating the construction of analytical models using algorithms that learn from data that can be readily automated. A computer programme known as machine language learns from experience (E) in relation to a certain class of task (T) and performance metric (P). If the task's performance as measured by P improves with "E" Training, validation, and testing are the three stages of machine learning. The performance of the model against validation data, not the accuracy on the test data set, should be used to determine which of the alternatives the best is.

Deep learning decreases the complexity of the model by using numerous deep layers of straightforward modules. This incorporates supervised and unsupervised learning connected to the system utilizing labeled data and unlabeled data in order to obtain the desired output model. Deep learning is used for distributed computing, unlabeled, uncategorized data processing, and learning. Deep learning models are used in a variety of machine learning applications to improve data sets for voice recognition and computer vision. It is also used to address complex technological issues on a wide scale. The Techniques of the Machine Learning are:

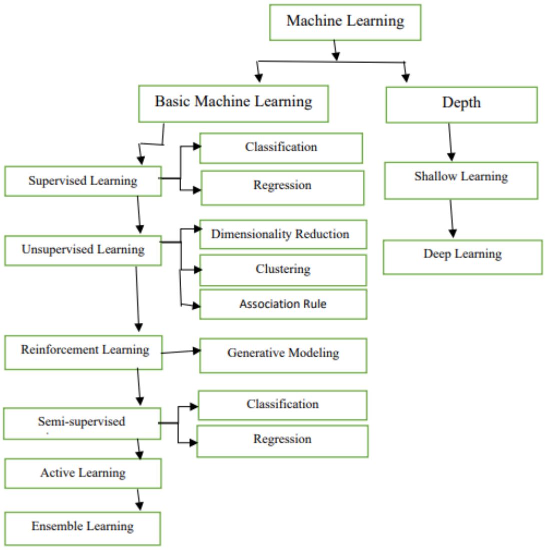

Fig 2 Classification of Machine Learning [2]

- Supervised Learning: In supervised learning, the model will learn from the training data using a pair of labelled inputs and intended outputs. A mapping function is created between input(x) and output once the dataset has been examined (y). Classification and regression challenges are the most typical application tasks
- Unsupervised Learning: Unsupervised learning uses the information and features of the data to determine the structure or models without using labels in the training dataset. There are no variables for the intended outputs and labelled inputs. Dimensionality Reduction, Clustering, and Association Rule Learning are the most typical application tasks. Unsupervised machine learning as language models will be the target of most assaults.
- Reinforcement Learning: Reinforcement learning involves interacting with the outside world and self-learning via trial-and-error behavior. It is taught to anticipate future observations by absorbing the concepts from experience. In this kind of learning, no privacy-related threats have been seen.
- Semi-Supervised Learning: This technique combines supervised and unsupervised learning to produce the required model from the training dataset using both labeled and unlabeled data. Unlabeled data are first used by the semi supervised learning algorithm for higher level interpretation, and then labeled data are used to streamline the subsequent tasks. Classification and regression are tasks.
- Active learning: Active learning involves actively choosing the training data in order to decrease the volume of labeled data while allowing for more flexibility. This affects the price and duration of labeled training data collection.
- Ensemble Learning: In ensemble learning, weak classifiers are combined and strengthened into strong classifiers by making independent predictions about each observation. Examples of ensemble learning include boosting and bagging samples. To differentiate between machine learning methods depending on how deep the identification task route is, machine learning depth is divided into shallow and deep learning.
- Shallow Learning: A common machine learning approach, shallow learning excludes several deep layers from the training dataset. As a consequence, computing that grows from many deep layers is less complicated. Therefore, shallow models have limitations in comparison to other models and are unable to identify model relationships.

3. LITERATURE REVIEW

The scholars SongnianLi, Suzana Dragicevic, (2016) in [3] made study on various geospatial theory and methods used to manage geospatial tremendous data. Given a few exceptional properties, makers thought about that standard data taking controlling methods of reasoning and frameworks are absent and the going with spaces were seen as in need for advance progress and assessment in the control. This circuits the types of progress in builds up to supervise consistent examination and to help advancing flooding data, and furthermore improving new spatial requesting techniques. The difference in speculative and systemic ways to deal with oversees trade of colossal data from illustrative and equal exploration and applications to ones that looks at pleasant and illustrative affiliation. Aithal, P. S. (2020) [4] Introduced an investigation of data-mining and simulated intelligence methods for upgraded assessment on the side of obstruction finding. Papers tending to every procedure were distinguished, checked out, and consolidated relying upon the amount of references or the consistency of a developing methodology. A couple of great high level instructive records used as a piece of simulated intelligence and data mining are represented for modernized security is shown, and a few hints on when to utilize a particular framework are advertised. This is on the grounds that data is so fundamental to simulated intelligence and data mining draws near. Li, Z., Li, X., Tang, R., & Zhang, L. (2021) [5] to dissipate the legend that individuals just contemplate cyberspace challenges as far as relationship and affiliation, we examined the worldwide cyberspace security hardships. A sum of serious areas of strength for 181 were mined from 40 objective sites involving the Apriori calculation in affiliation rules, and 56,096 pages were associated with worldwide cyberspace security. To get intensive data inclusion, this exploration additionally analyzed help, certainty, advancement, influence, and reliability. From the absolute example of 22,493 expert sites, 15,661 locales went on about terms connected with cyberspace security, making up 69.6% of the example; conversely, just 735 destinations went on about terms connected with cyberspace security from the complete example of 33,603 non-proficient locales, making up 2% of the example. The quantity of target proficient sites and non-target sites is restricted because of language requirements. Meanwhile, there aren't an adequate number of choices for intense guidelines. Web power, cyberspace security, cyberattack, cybercrime, data spillage, and data insurance are currently the really overall cyberspace security challenges. Grover, P., & Prasad, S. (2021) [6] specified significant block chain-oversaw data security challenges and carried out an exhaustive regular assessment of every one. We need to examine critical difficulties connected with current data sharing stages that rely upon trusted third parties (TTP), as well as security and security concerns connected with them, and how they might be settled utilizing block chain and data science innovation, in this article. What are the accessible apparatuses and techniques, calculations, and data science innovations for network security, data security, and data security. Our attention will be on cyber security data science (CDS), which is extensively connected with these areas as far as security data handling strategies and shrewd dynamic in certifiable applications. Another goal is to introduce the advantages and disadvantages of Block chain over data security in different authoritative areas. In general, Albums is worried about security data, utilizes machine learning methods to evaluate cyber threats, and in the end plans to improve cyber security tasks. Ul Islam, M.R. et al. (2019) [7] Software security can be summed up as developing efficient security awareness techniques to stop people from deliberately attacking software in order to steal private and sensitive data with the primary goal of achieving well-funded, destructive, and unethical goals that could harm people, countries, or the entire world. Because the injection attack pattern is so dynamic, it is very challenging to identify the pattern manually. We investigated this option since supervised machine learning for automated fraud and virus detection has been highly effective in recent years.

4. METHODS AND MATERIALS

In light of the aforementioned debate, the current study aims to investigate certain important facets of cyber security. Descriptive research is hence the form of study used in our current project.

a. Techniques of Malware Detections

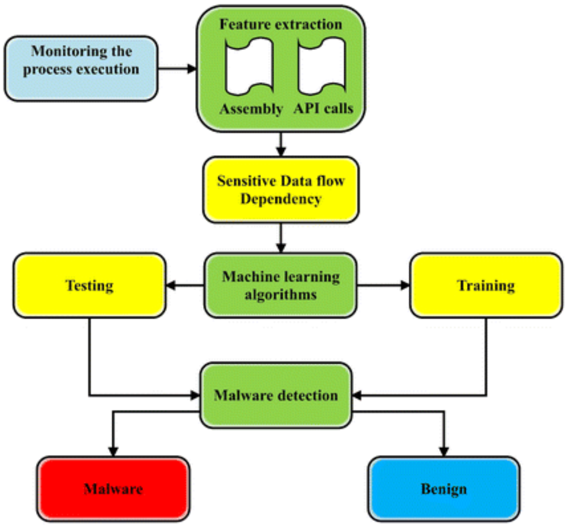
Monitoring the process execution
Feature extraction
Assembly
API calls
Sensitive Data flow Dependency
Testing
Machine learning algorithms
Training
Malware detection
Malware
Benign

Fig 3 Malware Detection Techniques

- Malware detection based on signatures: Signature databases monitor the malware effects had by earlier attacks. At the point when weak code is found, it is inspected by eliminating a particular bytes succession of code that fills in as an infection signature. In the event that it matched a current signature, the counter malware device would hail it as malware and pack the destructive code document. Here, hostile to malware software should hang tight for a mark before any gadget is attacked. [8]. While classifying a threat as malware, data mining strategies like order and relapse are utilized since it saves time and increments estimate exactness contrasted with the old methodology. This approach is easy to utilize, has broad malware data, is accessible, and is by and large acknowledged. [9] Utilizing specific obscurity and encryption strategies, a mark database might get around the risk. [8] It can't perceive the polymorphic infection that recreates data all through an enormous database. [9]

- Malware detection based on behavior: Program behavior, speed of execution, reaction time, browsing patterns, cookie information, types of attachments, and statistical features all aid in the identification of suspicious activity or harmful code. Utilizing a data mining technique, assembly characteristics and API calls are used in behavior-based detection. You may use unsupervised methods like clustering, SVM, and nearest neighbour algorithms to analyze behavior and find undetected malware. This technique aids in the detection of data flow dependencies in malicious software programmes as well as polymorphic malwares. Complex behavioral pattern detection takes more time and storage capacity. The following table shows data mining methods for detecting malware:

TABLE I TECHNIQUES OF DATA MINING FOR MALWARE DETECTION

Types of malware	**Techniques of Data Mining**	**Method of Analysis**
Polymorphic Malware Detection	K-means	Dynamic
Android Malware Detection	SVM, J48,	Naïve Bayes Dynamic
API Malware Detection	Naïve Bays, SVM, Decision Tree, Random Forest	Dynamic
N-gram Malware Detection	SVM, ANN	Dynamic
Service Oriented Mobile Malware Detection	Naïve Bayes,	Decision Tree Hybrid

Above table portrays various data mining procedures utilized for malware discovery agreeing their signature and conduct angles. To extricate concealed designs from the data static, dynamic and crossover data investigation procedures are utilized for further developing exactness of malware recognition. It is the test for cyber security specialists to choose best calculation and data investigation methods for tracking down the secret threats and give alarms to give data from additional attacks.

b. ML algorithms on MODBUS Data

This assessment's essential goal is to decide whether certain ML calculations can actually recognize cyber-attacks on MODBUS data. The ML models were made utilizing ten times cross approval. Weka [11] was utilized for this examination. Weka creates 10 particular models for the given data set in 10 crease cross approval. The weighted normal of these models is then registered and displayed as the result. The data assortment utilized was known as Telemetry Data from Gas Pipeline and was made by Mississippi Express College's Basic Foundation Insurance Center [12]. For the evaluation, only a couple of traditional classifiers were considered. The methods used were:

- Credulous Bayes, a probabilistic classifier in light of Bayes' hypothesis.
- The choice tree-based Irregular Backwoods A Ml calculation.
- One R — every part of the standard set is evaluated, and the ideal or best one is at last picked.
- An essential execution of the Choice Tree Calculation (C4.5) in J48.

c. Data Set

Weka minable arff design was used for the dataset. There were 20 qualities in all. All of the characteristics that are present in the dataset are listed in Table II below.

TABLE II ATTRIBUTES OF DATASET

Attributes	
Address	Rate of reset
Control strategy	command reaction
Perform	deadband
Pumping	period
Duration	Cycle time
solenoid	binary outcome
Set point	speed
Pressure measurement	Classification outcome
gain	system mode
Crc Rate	exact result

5. RESULTS AND DISCUSSION

The ICS dataset has recently been used by Beaver et al. [13] for ML calculation research. Be that as it may, no calculation's ROC bend was drawn, making it truly challenging to decide how well the calculations performed by and large. The plot of the false positive rate(FAR) against the test awareness is known as the receiver operating characteristics (ROC) bend. A fundamental boundary is the ROC's region under the bend. Assessing the explicitness and sensitivity is utilized. Hence, the consolidated proportion of awareness and particularity is the region under the ROC bend. This measurement might be utilized to assess the general exhibition of the ML calculations utilized in the order of the MODBUS data since the region under the ROC bend is a proportion of the general presentation of any test.

Thus, the arrangement execution of the calculations not entirely set in stone by doing an AUC examination of the ROC bend for different ML methods. The current review was finished utilizing the Weka Information stream model. In Figure 5, it is shown [14]. The four ML calculations' separate Roc bends. As per the ROC bend, the j48 calculation conveys the best generally speaking characterization results for the power framework dataset. The table beneath shows the AUCs for the four calculations.

TABLE III TRAFFIC CLASSIFICATION MODEL

algorithm	Area under the curve	Precision	Recall
OneR	0.885	0.823	0.853
Naïve Bayes	0.963	0.985	0.956
Random Forest	0.987	0.977	0.958
J 48	0.997	0.994	0.992

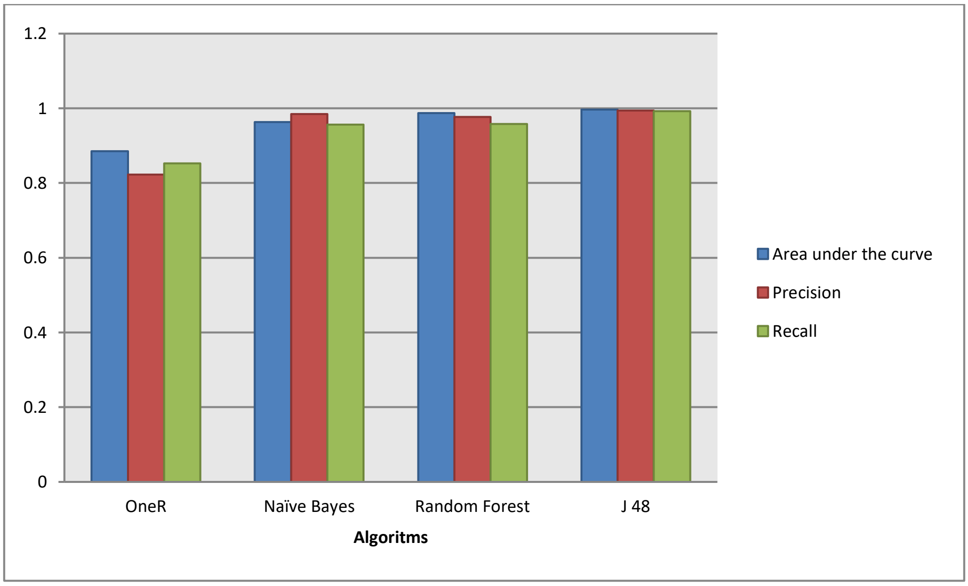

Fig 4 Traffic classification model

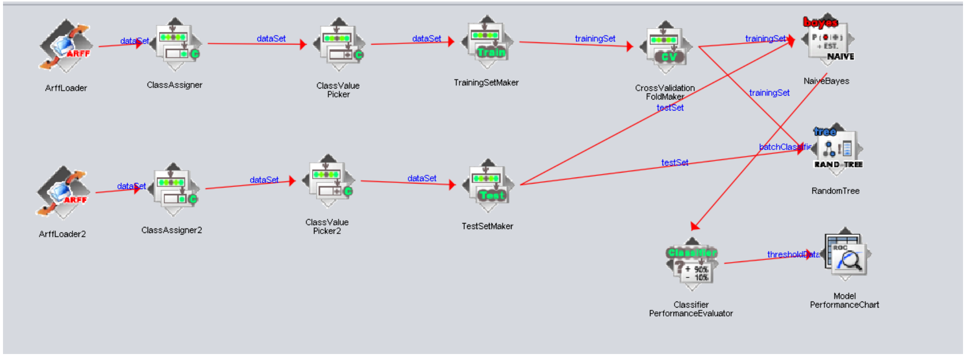

Fig 5 Weka ROC curve generator

As an issue of double grouping, this study was led. Multiclass order may likewise be used another way. Second, the weighted normal of each of the 8 classes is considered in the calculation including the AUC, Accuracy, and Review introduced in Table III. It is feasible to concentrate on each class freely, which will uncover how well the model can recognize different attack types and standard traffic [15].

The discoveries in Table III show the model's ability to classifications all traffic. It is obvious from Fig. 6 that the J48 plays out the best by and large in arrangement since the ROC bends' region under the bend values is more like 1. The adequacy of the machine learning interruption recognition framework being used in modern control frameworks is vital. Consequently, the preparation should be pretty much as effective as conceivable so that newly streaming data might be learned on time and the calculation can keep on checking data progressively [16]. Thus, while choosing a calculation, picking a couple with the best exactness and preparing time for every space of the modern control system is frequently ideal. The calculation's preparation time was determined utilizing a Python script during the K-overlay approval. Since the deliberate preparation time is how much time expected to check each overlap, in a genuine setting, the preparation time might be determined to be 1/Kth of the milliseconds showed in figure 7. The PC used for preparing highlights a Nvidia GTX 1060 devoted illustrations card and an Intel i7 6700HQ primary computer processor. Due to the 1256 Cuda centers of the Nvidia GTX 1060, it is significantly parallelized to prepare execution [17].

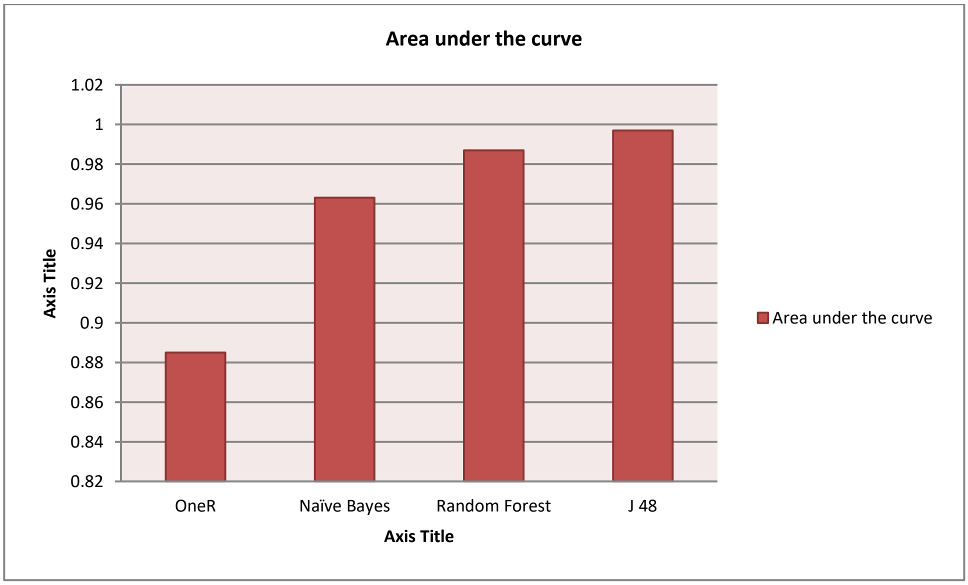

Fig 6 All 4 ML algorithms' curve area

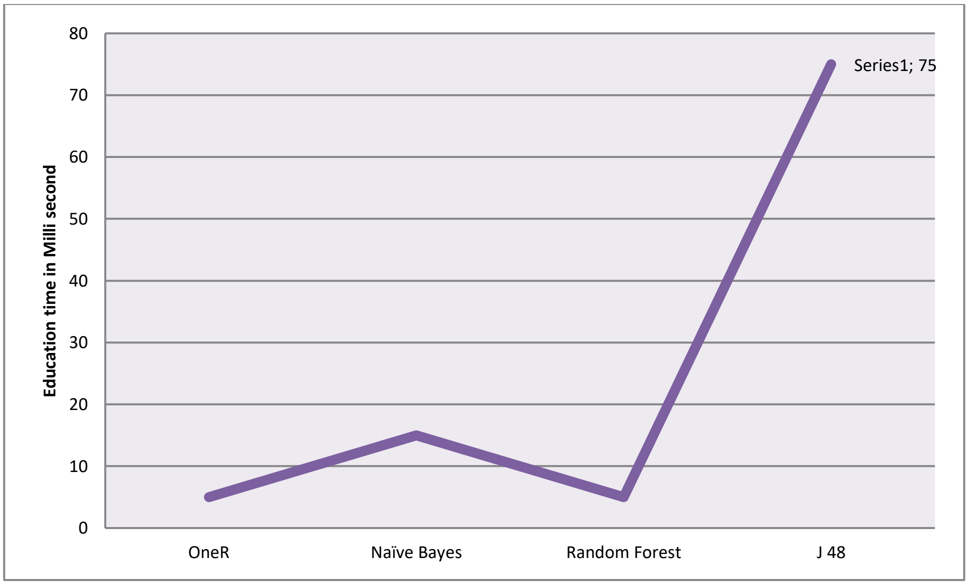

Fig 7 Training duration for every algorithm

Along these lines, the IEEE 9 transport framework when the calculations are utilized as a principal part of the interruption discovery framework, even while J48 could perform preferred as far as exactness over Irregular woodland, a little penance in precision could bring about more prominent continuous execution [18].

5.1 Malware Statistics

According to studies, malware assaults are to blame for 80% of system damage [10]. Malware is discovered to be 92% distributed through email attachments. The prevalence of mobile malware has increased by 54% during 2021. Malware mostly targeted Android smartphones in 98%. 99% of malware was downloaded through unofficial apps. Seven of the ten payloads are ransom ware [19-20]. In a single week, malware infects 18 million websites worldwide. Since 2018, 90% of financial institutions have been the victim of malware. 40% of the ransom ware victims paid the demand. Bit coin demands are made in more than 50% of ransom ware attacks.

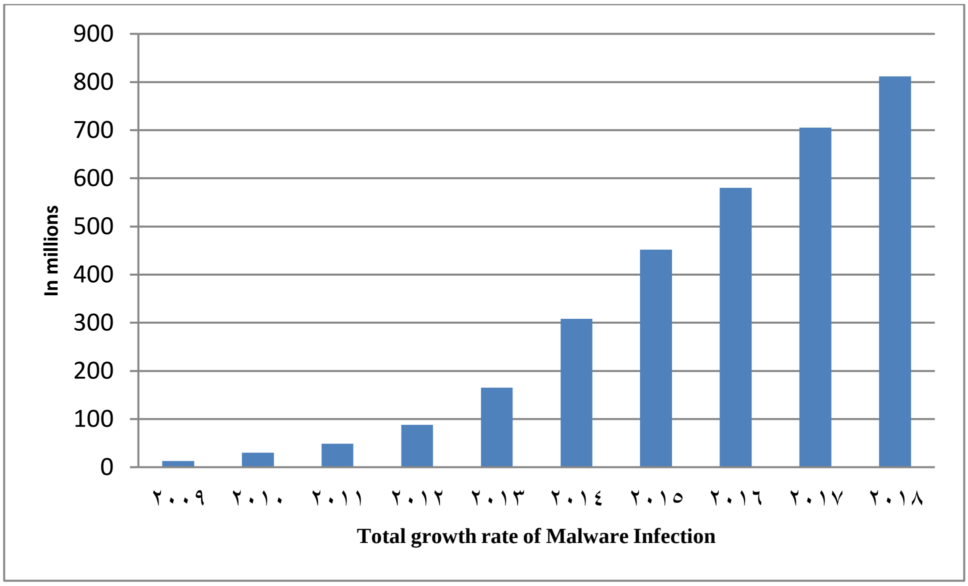

Fig 8 Total growth rate of Malware Infection

6. CONCLUSION

In this review, a careful examination led to recognize a couple of unique datasets. Following that, several ML calculations and their utilization in cyber-security were investigated. At long last, a couple of ML ideas were given for determination. A speedy assessment of an ICS data set and an assessment of the exhibition of a couple of ML calculations were finished in the paper's last option segments. Web and correspondence innovation use has expanded because of globalization [20]. The primary drivers of cybercrime incorporate data spillage, unstable Wi-Fi associations, and an absence of safety information, hardware, software, and organization weakness. Executing powerful and savvy systems for early discovery of cyber threats as a reasonable security arrangement is imperative to lessen the gamble of devastating cyber-attacks, including data breaks, ransomware attacks, and DDoS attacks. One of the difficulties for security experts is malware identification. Data mining strategies, for example, grouping, SVM, relapse, choice trees, diagram mining, and KNN calculations, might be joined against threat frameworks to help distinguish malware before it enters the framework, shielding your IT foundation from additional attacks. Insightful malware identification is made conceivable by counterfeit brain organizations, hereditary calculations, and profound learning components utilizing conduct and mark databases. Albeit the J48 calculation beats different calculations in the review's domain, more examination is expected to decide every strategy's exhibition since every calculation's presentation shifts in light of the dataset to which it is applied. Second, given its best ongoing execution for the situation viable, Irregular woodland would be a definitive decision as the principal IDS calculation.

Funding

No Funding was received for this work.

Conflicts Of Interest

Authors declare no conflicts of interest.

Acknowledgement

The authors would like to thank the anonymous reviewers for their efforts.

References

[1] A. Mukkamala, A. Sung, and A. Abraham, "Cyber security challenges: Designing efficient intrusion detection systems and antivirus tools," in Enhancing Computer Security with Smart Technology, V. R. Vemuri, Ed. New York, NY, USA: Auerbach, 2005, pp. 125–163.
[2] O. Ibitoye, "The Threat of Adversarial Attacks on Machine Learning in Network," arXiv preprint arXiv:1911.026213, 2019.
[3] S. Li et al., "Geospatial big data handling theory and methods: A review and research challenges," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 115, pp. 119-133, 2016.
[4] P. S. Aithal, "Data Mining and Machine Learning Techniques for Cyber Security Intrusion Detection," 2020.
[5] Z. Li et al., "Apriori algorithm for the data mining of global cyberspace security issues for human participatory based on association rules," Frontiers in Psychology, vol. 11, p. 582480, 2021.
[6] P. Grover and S. Prasad, "A Review on Block chain and Data Mining Based Data Security Methods," in 2021 2nd International Conference on Big Data Analytics and Practices (IBDAP), pp. 112-118, 2021.
[7] M. R. Ul Islam et al., "Automatic detection of NoSQL injection using supervised learning," in 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 2019.
[8] S. Najari and I. Lotfi, "Malware Detection Using Data Mining Techniques," International Journal of Intelligent Information Systems, vol. 3, no. 6-1, pp. 33-37, 2014.
[9] R. Choudhary and R. Saharan, "Malware Detection Using Data Mining Techniques," International Journal of Information Technology and Knowledge Management, vol. 5, no. 1, pp. 85-88, 2012.
[10] K. Rieck, T. Willems, et al., "Learning and classification of malware behavior," in 5th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin, Heidelberg: SpringerVerlag, 2008, pp. 108–112.
[11] M. Hall et al., "The WEKA data mining software: an update," ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10-18, 2009.

[12] T. H. Morris, Z. Thornton, and I. Turnipseed, "Industrial Control System Simulation and Data Logging for Intrusion Detection System Research," n.d.

[13] J. M. Beaver, R. C. Borges-Hink, and M. a. Buckner, "An Evaluation of Machine Learning Methods to Detect Malicious SCADA Communications," 2013 12th International Conference on Machine Learning and Applications, pp. 54–59, 2013.

[14] M. Norouzi, A. Souri, and M. S. Zamini, "A Data Mining Classification Approach for Behavioral Malware Detection," Journal of Computer network and Communications, vol. 2016, 2016.

[15] Yanfang, Donald Adjeroh, et al., "A Survey on Malware Detection Using Data Mining Techniques," ACM Computing Surveys, vol. 50, no. 3, p. 41, 2017.

[16] K. Rieck, T. Willems, et al., "Learning and classification of malware behavior," in 5th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin, Heidelberg: SpringerVerlag, 2008, pp. 108–112.

[17] A. Martin, H. D. Menéndez, and D. Camacho, "MOCDroid: multi-objective evolutionary classifier for Android malware detection," Soft Comput, vol. 21, pp. 7405–7415, 2016.

[18] A. Hellal and L. B. Romdhane, "Minimal contrast frequent pattern mining for malware detection," Comput Secur, vol. 62, pp. 19-32, 2016.

[19] A. Bhattacharya and R. T. Goswami, "DMDAM: data mining based detection of android malware," in Proceedings of the First International Conference on Intelligent Computing and Communication, Springer Singapore, Singapore, 2017, pp. 187–194.

[20] C. Fan, H. W. Hsiao, C. H. Chou, and Y. F. Tseng, "Malware detection systems based on API log data mining," in 2015 IEEE 39th Annual Computer Software and Applications Conference, 2015, pp. 255–260.

.