Asian Journal of Research in Computer Science

Volume 16, Issue 4, Page 453-464, 2023; Article no.AJRCOS.111124
ISSN: 2581-8260

Harnessing Machine Learning for Effective Cyber security Classifiers

Tamanna Jena [a*], Achyut Shankar [b] and Adyesha Singhdeo [c]

[a] Fairleigh Dickinson University, Canada.
[b] University of Warwick, United Kingdom.
[c] University of British Columbia, Canada.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2023/v16i4405

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/111124

Method Article

Received: 16/10/2023
Accepted: 23/12/2023
Published: 29/12/2023

ABSTRACT

Machine learning has emerged as a transformative force, innovating diverse industries through its capacity to infuse meaningful insights from large datasets. It plays a pivotal role in powering data analysis, discover pattern matching, identifying hidden or evolving risks in securing systems. The ability of categorizing and behavior analysis is central to its efficacy in cybersecurity. This paper highlights the importance of machine learning in landscape of cyber threats. In this paper, we have identified few machine learning algorithms to categorize huge dataset. The complexities of identifying hidden risks increases by many folds, when the input data is voluminous. Evaluating and contemplating the underlying meaning of data is time-consuming and can be missed easily. We compared different types of machine learning algorithms. Each machine learning algorithm has its strength and weakness. It is found that, the TressJ48 algorithm is proficient in classifying the large dataset, better than Naive Bayes and Decision Stump algorithms. The efficient classifier helps to generate insight, which can be further used to make decisions in terms of cybersecurity.

**Corresponding author: E-mail: tamannasinghdeo@gmail.com;*

Keywords: Machine learning; security; cybersecurity; detection system; classification.

1. INTRODUCTION

The proliferation of smart devices like smart phones, tablets, IoT devices, and other connected technologies like Intrusion detection system (IDS), Intrusion Prevention System (IPS) has been an eminent trend. This is otherwise called Fourth Industrial Revolution, commonly refereed as Industry 4.0 [1]. In the last 5 years the digital devices increased from 8.3 billion to 30 billion. Statista suggested that there will be an exponential rise of digital devices ranging to 75.44 billion by the year 2025 [2]. With the increasing ubiquity of IoT devices, the number of devices to be used in cyber-attacks increases [3,4]. The usage of interconnection of digital devices generates a huge volume of data. The exponential increase in connected devices and the extent of cybersecurity threats makes it evident to keep the cybersecurity practices secure. The technical report of Gartner 2023, outlines the strategic technology trends [5]. It highlights that the observable data is the most precious monetizable asset for any business in the upcoming technological era. If the data and metadata can be used as input in AI-based models to extract business capabilities, then businesses can use the obtained knowledge to gain competitive advantage and security from peers. Additionally, it is predicted that by 2026, the organizations that operationalize AI for transparency, trust and security were likely to see a 50 % improvement in terms of business goals and user acceptance.

This paper aims to highlight the importance and impact of cybersecurity practices in terms of the selected machine learning algorithms. Organizations of size small, medium and large are using cybersecurity approaches like Intrusion Detection System (IDS) [6,7], Intrusion Prevention System (IPS) [8-10], penetration testing to find vulnerabilities in their system and make it more robust. This paper mainly focuses that efficiency of IDS, IPS and penetration system [11-14] depends on the machine learning algorithm used in these systems. The introduction of this paper outlines the importance of machine learning algorithms [15-21] in cybersecurity measures. Section 2, discusses the about of various types of machine learning algorithms which are used in cybersecurity context. It also entails the difference between machine learning and cognitive intelligence. Section 3, depicts the related work on the subject matter and the reason behind the research path. Section 4 highlights the machine learning based classification model. It discusses its input, data flow, and output. It mainly emphasizes on identification of the vector space model, which provides framework for analyzing textual information and section 5, explains the methodology of the proposed classification model, experimental set-up and result. Section 6 includes a conclusion.

2. MACHINE LEARNING

The brain and natural intelligence extract information are categorized at the level of data, information, knowledge and intelligence. In this paper, we are trying to follow the same path. The data level represents the raw data. At the information level, we clean the data and define a scope. At the knowledge level, we try to get insight. Finally, at the intelligence level, we try to apply the gained knowledge in the system to make it more robust and efficient.

Data and information processing have been studied for a long time. However, with the emergence of Big Data and powerful data processing tools, researchers and academicians are non-stop working towards application-oriented data processing. The research progress in theories, mathematical approaches, and systematic studies in cognitive informatics and machine learning computing are yet to be mainstream. The basic approach is to invent cognitive computers, cognitive robots, and cognitive systems that extend human learning ability, wisdom and creativity [22-25]. The cognitive ability of system can enhance with usage of machine learning. Machine learning models are categorized into supervised and unsupervised learning. It further divided into many methods like SVM (Support Vector Machine), ANN (Artificial Neural Network), MLP (Multi-Layer Perceptron). Every technique has its strength and weakness. With more devices and more businesses moving to information technology, there has been a steep rise in cybersecurity threats in the last decade. In today's day and age, every organization of all sizes are directly or indirectly impacted by data leak and cyber threats. Researchers have highlighted the state of the art of cyber-attacks and suggested a security framework based on industries to tackle attacks [26-28]. As per Cisco Systems, there has been a jump of 25% in

cybersecurity attacks in almost more than 60 % of organizations worldwide since 2020 [29]. In cybersecurity machine learning can be beneficial to understand the weak spot, and identifying the application detection systems [30-34]. Applications like cognitive robots, cognitive learning engines, cognitive internets, cognitive translators, cognitive control systems, and cognitive automobiles are a few functional aspects of cognitive computing [35].

3. RELATED WORK

Machine learning has been instrumental in cybersecurity [36-40]. The capabilities which make machine learning importance are anomaly detection [41-45]. Behavioral analysis [46,47], predictive analysis [48,49], real-time analysis [50,51,52], network security monitoring [53-57]. The intrusion detection model has been used for many years to detect break-ins, penetration, and computer-related attacks [58-60]. It was found that anomaly detection in predefined signatures, network traffic and content hidden in single network packets were enhancing the efficacy of the Intrusion Detection System (IDS) [61]. Additionally, data processing enhances the accuracy and capability of network intrusion detection systems (NIDS) [62,41]. Data preprocessing helps in uncovering novel attacks, misconfiguration, and even network failures [63]. Some researchers used the feature selection method to classify the important and impactful features from irrelevant ones. It was found that an entropy-based multi-step outlier-based approach was proving to be beneficial for detecting anomalies in network-wide traffic and a tree-based clustering technique to generate to identify anomalies [64]. Although anomaly detection on network traffic was helping in detecting attacks and network failures, it also suffers from certain drawbacks like stale datasets to work, and the algorithms were inefficient in learning new models. Hence, machine learning algorithms become popular in anomaly detection in IDS [65]. Extracting patterns from cybersecurity data and building a data-based or data-driven model is the way forward. Researchers are using the multi-layered framework to find efficient cybersecurity modelling [24]. Some researchers used genetic algorithms to find the optimization of learning algorithms [66,67]. Some researchers analyze the effectiveness of machine learning classification for accurately predicting user behavior [68]. They consider algorithms like ZeroR, Naïve Bayes, Decision Trees, Random Forests, Support Vector Machines and Logistic Regression classifiers. Some researchers used advanced Naïve Bayed i.e., Hidden Naïve Bayes Model as they found it works better than its traditional form in terms of higher accuracy [69].

4. MACHINE LEARNING BASED CLASSIFIER MODEL

Our Machine Learning based classifier model consists of various roles: attackers, users and security analysts.

1. Attack vector or attacker: Individuals who launch attack with malicious intention are called attacker. If the motivation of any action, or pattern is found to be malicious for others or inclined to extract information to gain competitive advantage, then the user is flagged as an attack vector. On the basis of the outcome of the machine learning on the data and metadata of the attack vector, its profile is created. The Table 1 outlines the categories of the attackers. The important decision factors in identifying the category of attack vectors are type of attack, extent of the impact, behavior pattern, cultural characteristics and transactions with other attackers
2. User: Any individual who uses connected devices and the internet can be considered as a user. However, on the basis of the scope of the experiment, the user is also categorized into various types. The important decision factors in identifying the category of user are knowledge-based, extent of use, recurring applications, and commercial/personal use.
3. Security Analysts: Any individual who uses different data collection tools, and sensors to collect data, and apply machine learning tools, and algorithms on the normalized data and metadata to obtain knowledge about efficient automation, pattern of attackers, and pattern of users is termed a security analyst. The security analyst evaluates every action of the users, attackers, environment, peripheral factors which are either extrinsic or intrinsic in nature. The security analyst will come up with a hypothesis, which will be backed up by analytical and statistical reasoning. This entire process is termed a machine learning pipeline.

Table 1. Categories of cyber users

Categories of users	Motivations	Extent of Impact
Script Kiddies	Curiosity	Low
Hacktivists	Defacement of individual or group for political reasons	High
Cyber punks	Exploring and engaging malicious attacks	Medium
Coders	Write/Use automated tools	Low
Insiders	Displeased employer	High
Cyber terrorists	Spreading fear and instability	High
Hackers	Leaking data	High
Pen testers	Finding/identifying vulnerabilities of a system	Low

The Machine learning based classifier model includes the following processes:

1. Collection of data
2. Cleaning of data
3. Generation of the vector space model
4. Generation of hypothesis
5. Research on hypothesis
6. Knowledge on the basis of hypothesis

After collecting the data from a reliable resource, the data is processed. We collected various datasets, like DNS dataset, Darknet dataset, IoT datasets, Malware datasets, Intrusion Detection datasets (IDs). Then we amalgamated all these datasets so that a diversity is maintained in the dataset. Moreover, it resembles like a real-life network traffic. Post amalgamation, we processed the data further. It mainly involves normalization of data, formatting the data in a certain order or format, and getting rid of irrelevant data or missing data. A hypothesis is formed on the basis of the available evidence as the starting point for investigation. If the findings align with the hypothesis, then the hypothesis is confirmed. Here, our hypothesis is TreesJ48 is the machine learning algorithm which classifies the dataset better that Naïve Bayes and Decision tree machine learning algorithm. Our next step after processing of data is to apply the identified machine learning algorithms on it. Otherwise, the hypothesis is discarded and on the basis of the finding, a new hypothesis is formed which will be tested next.

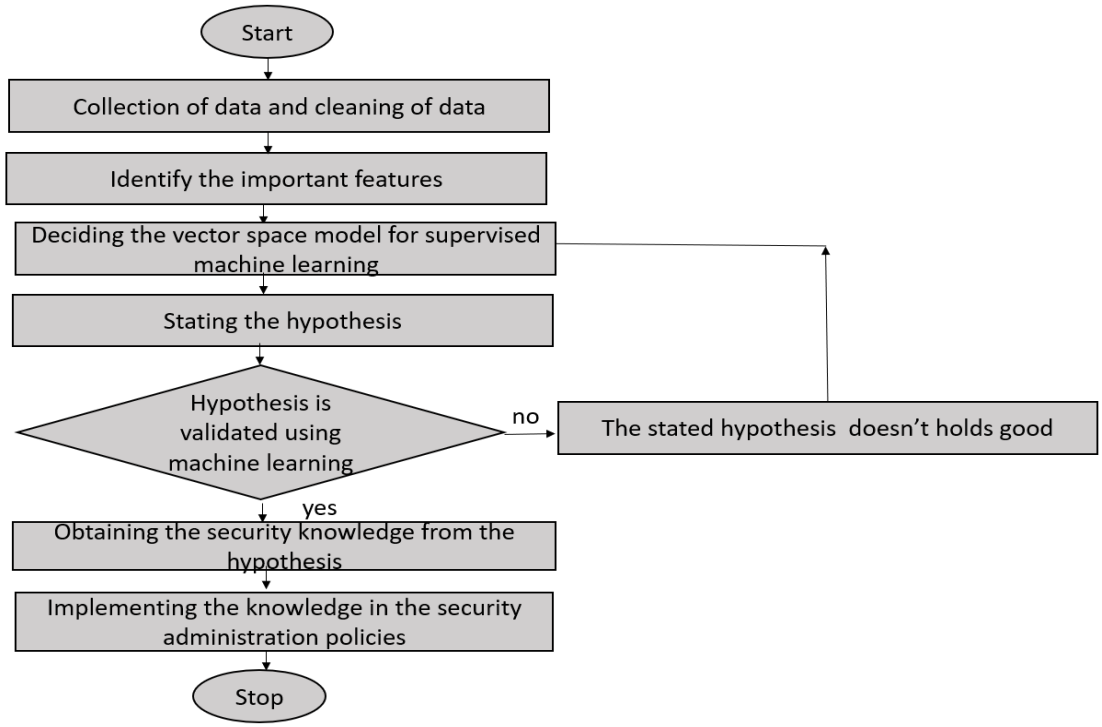

Fig. 1. Effect of different flowchart of machine learning based classifier model

5. METHODOLOGY

We experimented with a dataset by the Canadian Institute for Cybersecurity. The datasets are of diverse types, including DNS datasets, IDS datasets and malware datasets. The dataset which we used most for our experiment is collected from CIC Darknet 2020 [70], CIC-Bell-DNS 2021 [71] and CIC-Bell-DNS-EXP-2021 [31]. The former dataset consists of audio-stream, browsing, chat, email, video stream, VOIP [70]. The later datasets consist of benign, malware, phishing and spam datasets. These datasets were collected in 2020-2021 in collaboration with CIC and Bell [71,72]. The datasets consist of benign, DNS, darknet,malware, spam, phishing datasets. After collecting datasets, it was found that some data were balanced (60:40%; benign: malicious). Whereas some were unbalanced (90:10%; benign: malicious). In the experimental set up, we collected different types of datasets mentioned above and amalgamated the datasets to ensure the dataset is diverse and looks like a representative dataset of real-world traffic. On the amalgamated dataset, we normalized few data and used the function to datetime mapped few data into time form to enhance the readability of the data. We got the duration of the attacks/transactions. After preparation of the data, we used machine learning algorithms to categorize. Our null hypothesis is, if the machine learning tool will be able to categorize the dataset into different types like audio streaming, video, email and others from the raw data then it will help in finding behavioral patterns and identifying similar attack vectors. The scope of this paper is to find out the efficient machine learning algorithms which can categorize large datasets into different categories, on the basis of their types. The proposed alternate hypothesis for our experiment is finding an efficient machine learning algorithm that can categorize large raw datasets into different types on the basis of their size, and type of data, hence it will further help in identifying the behavioral patterns of cyber users.

Table 2 outlines the types of classifiers we have identified as the output. The input dataset after processing entails of more than 185 features. We decided to make the data specific for better understanding. Hence, 67 features to run tests like classifications using different types of algorithms. It shaped the vector space model. Fig. 2 outlines the flow diagram of the Relations of data and knowledge in the machine learning based classifier Model. As the data obtained is in raw form, we identified 67 essential features as the column of the X matrix. Y is another matrix which consists of the type of traffic category explained in Table 2.

Table 2. Categories of the medium's types in the dataset

Traffic category	Applications	Descriptions
Audio-streaming	Vimeo, Spotify and Youtube	It identifies audio application that requires continues stream of data
Browsing	Firefox and Chrome	Traffic generated by users using HTTP and HTTPS
Chat	ICQ, AIM, Skype, Facebook and Hangouts	It identifies instant messaging applications, used in facebook, Hangouts, Skype
Email	SMTPS, POP3S and IMAPS	It identifies a traffic where clients configured to communicate through SMTP/S, POS3/SSLIMAP/SSL
P2P	Torrent and Transmission (BitTorrent)	It identifies file-sharing protocols, it mainly uses Vuze applications
Transfer	Skype, FTP over SSH (SFTP) and FTP over SSL (FTPS)	It identifies traffic applications whose main purpose is to send or receive documents or files
Video-Stream	Vimeo and YouTube	It identifies applications that requires a steady stream of video data
VOIP	Facebook, Skype and Hangouts voice calls	It identifies applications where voice-calls using Facebook, Skype and Hangout

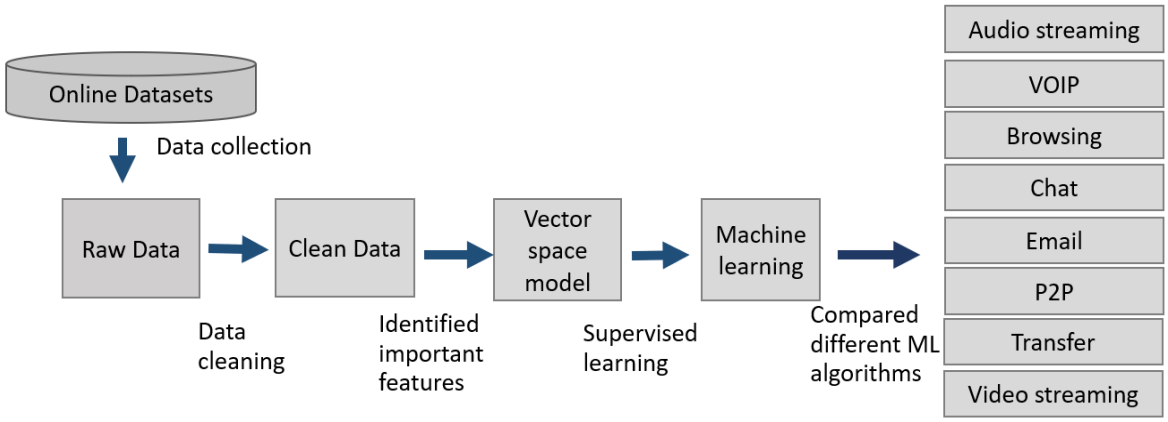

Fig. 2. Flow of data in in machine learning based classifier model

5.1 Experimental Setup

Machine learning is a form of applied statistics which primarily focuses on estimation and prediction. It provides the ability to learn to produce the desired prediction without involving a rigorous amount of programming. It is divided into supervised learning, unsupervised learning and semi-supervised learning [21]. In the former type of machine learning, the learning function maps input to output. As the results are known, the algorithms are corrected until the performance reaches an acceptable performance level. Therefore, the application of supervised learning is high in every domain involved in the prospect of artificial intelligence. Some examples of supervised learning algorithms are the Decision tree, TreesJ48 algorithm, Naive Bayes algorithm, and Decision Stump. The most popular unsupervised learning algorithms are K-mean clustering and Hierarchical Clustering.

The outcome of the machine learning-based classifier model allows organizations to get insight into the existing risks, vulnerabilities, possible attacks, hidden attacks, patterns of attackers and gaps [72a]. Based on the knowledge obtained, the organizations can project the policies, conditions and strategies in the near future. The outcomes are categorized primarily into two levels of security insights. The first/ lower level is to find the technical solution, which can be further used in automation or responsive action. The second level/higher level of awareness helps to establish modified strategic decisions. The latter mainly involves human intervention. It is risky and needs a visionary approach.

The dataset used is a large table containing 141530 instances. Where each instance is depicted in a row and the number of columns was 185. The attributes along the column were IP address, Source IP, Destination IP, Packets per second, Type of protocol, and number of packets/sec. For the vector model, we cleaned the initial dataset and identified 65 features which were specific and relevant to the instances in categorizing the dataset. The dataset has eight types of mediums ie, audio-streaming, browsing, VOIP, email, P2P, video-transfer, and chat outlined in Table 2. We changed the type of the column to nominal. For implementation of machine learning algorithms experiments, we used Weka. It is a collection of machine learning algorithms for data mining tasks. The tools embedded can be used for data preparation, classification, regression, clustering and many more association rules. It performs well to decide what information is relevant most. One of the best advantages of using Weka is the implementation of multiple algorithms is comparatively easy and the results obtained are intuitive [73]. We normalized a few data of the dataset like bandwidth, total forward packets, and total backward packets. We shortlisted 3 machine learning algorithms for this experiment, i.e, Decision Stump, TreesJ48, Naive Bayes [66,69,74,75].

J48 Trees is a classification algorithm which produces decision trees based on information theory [76-80]. Here J stands for Java. It is basically a statistical classifier and an open-source Java implementation algorithm. Its strength is it requires less effort for data preparation during pre-processing. It doesn't require the normalization of data. It doesn't suffer from overfitting. The main motivation behind choosing this algorithm is, that it ignores the missing values and overcomes the overfit. If there is overfitting in the classification then it self-

prunes the node and subbranches of the overfit node. J48 Decision Tree is a univariate decision tree.

A Decision Stump is a machine-learning model consisting of a one-level decision tree. It is a decision tree with one internal node which is immediately connected to the terminal nodes i.e., leaves. Here the decision is made on the value of just a single input feature, and a tree model is formed using a hierarchy of branches. The path from the root node through internal nodes to a leaf node represents a classification decision rule.

Naive Bayes machine learning algorithm is a Bayesian Learning algorithm. It is mainly popular in Natural Language Processing. The model makes an assumption tag of the text and tries to classify the texts with numerous classes. This algorithm is simple to implement and works well with both discrete and continuous data. It was selected as it can handle an enormous size of datasets. Though its performance is average it can be used for exclusively numerical value.

During implementation, we tried many different setting options like cross validation, training set, and percentage set. We decided with cross-validation to be 10. In n-cross validation, the dataset is divided into n equal-sized folds or subsets. The model is then trained and then evaluated n-times. The model used each time a different fold as the validation set and the remaining 1-fold as the training set.

The weka tool kit produces the confusion matrix [73] after the selection of the algorithms. In the case of the correct classification, the numbers from the top-left to bottom-right need to be bigger numbers than the rest of the matrix i.e, when the confusion matrix predominantly looks like a diagonal matrix.

In the case of the algorithm, which categorizes data sets efficiently, the confusion matrix will look like a diagonal matrix. Fig. 3 shows the confusion matrix obtained using the treesJ48 algorithm. Confusion matrix left top quadrant consists of true positive, top-right consists of false positive. Bottom left quadrant contains false negative and bottom right makes true negative. So, if the confusion matrix looks like a diagonal matrix then its efficiency as a classifier is excellent otherwise if the confusion matrix is a poor classifier. Fig. 4 shows the confusion matrix obtained using the Decision Stump algorithm and Fig. 5 shows the confusion matrix obtained using the Naive Bayes algorithm. It was found that TreesJ48 is able to categorize the dataset efficiently. Table 3 validates the result of confusion matrices and the efficiencies of the algorithms. Hence the simulation proves that algorithms were able to categorize the dataset differently. The efficient

Table 3. Comparisons between machine learning algorithms

Type of Classifier	Run time per model	Percentage of correctly classified instances	Root Mean Squared Error	Total no. of instances	Cross-validation 10 folds
Decision Stump	9.62 secs	51.81%	0.2414	141530	10
TreeJ48	29.97 secs	95.08%	0.0741	141530	10
Naive Bayes	0.08 secs	45.00%	0.2669	141530	10

```
    a     b     c     d     e     f     g     h     i     j     k   <-- classified as
 1106   164     9     6    75     0    87    37     0     0     0 |     a = AUDIO-STREAMING
   13 32194     5     1   440     0    26   120     7     0     2 |     b = Browsing
   31    44 11234     7    84     0    21    28    26     0     3 |     c = Chat
    4    14    47  6052    17     0     5     6     0     0     0 |     d = Email
   19  1008    96     7  9693    15    62   161    37     0     0 |     e = File-Transfer
    1     6     1     0    20    56     0     0     0     0     0 |     f = File-transfer
    0   160     1     2    27     0 48318     4     8     0     0 |     g = P2P
   26   734   253     7   211     0    12  8191    31     0    21 |     h = Video-Streaming
    0     4   383     0     5     0     2   232 15954     0     0 |     i = Audio-Streaming
    0     0    44     0     0     0     0     0     0   237     0 |     j = Video-streaming
    0     0   310     0     2     0     0    40     5     0  3209 |     k = VOIP
```

Fig. 3. Confusion Matrix: implementing treesj48 machine learning algorithm

a	b	c	d	e	f	g	h	i	j	k	<-- classified as
0	0	0	0	0	0	1484	0	0	0	0	a = AUDIO-STREAMING
0	0	0	0	0	0	32545	0	263	0	0	b = Browsing
0	0	0	0	0	0	413	0	11065	0	0	c = Chat
0	0	0	0	0	0	492	0	5653	0	0	d = Email
0	0	0	0	0	0	6693	0	4405	0	0	e = File-Transfer
0	0	0	0	0	0	47	0	37	0	0	f = File-transfer
0	0	0	0	0	0	48300	0	220	0	0	g = P2P
0	0	0	0	0	0	3382	0	6104	0	0	h = Video-Streaming
0	0	0	0	0	0	0	0	16580	0	0	i = Audio-Streaming

Fig. 4. Confusion matrix: implementing decision stump machine learning algorithm

a	b	c	d	e	f	g	h	i	j	k	<-- classified as
9	724	25	12	13	50	597	32	21	0	1	a = AUDIO-STREAMING
31	25930	472	149	268	643	4787	112	359	4	53	b = Browsing
8	1546	5443	355	153	750	1951	60	197	6	1009	c = Chat
4	1241	1750	416	64	542	1221	20	93	1	793	d = Email
10	6204	338	63	154	437	3407	109	332	3	41	e = File-Transfer
0	19	4	0	2	6	43	1	2	0	7	f = File-transfer
22	21748	1134	172	258	759	23795	169	415	5	43	g = P2P
27	5487	148	55	90	1078	2189	122	273	6	11	h = Video-Streaming
4	2976	263	32	222	7852	4021	191	989	15	15	i = Audio-Streaming
0	127	14	2	6	28	71	11	18	4	0	j = Video-streaming
1	423	757	76	32	470	537	17	48	0	1205	k = VOIP

Fig. 5. Confusion matrix: implementing naive bayes machine learning algorithm

algorithm can be used for further findings in the dataset. It was discovered during the simulation that, theTreesJ48 algorithm was able to identify the significant predictive values in the dataset. It was proficient in classifying significant from insignificant predictive values and not landing on biased end results.

6. CONCLUSION

This article advocates for the adoption of a Machine Learning-based Classifier model in cybersecurity, underscoring its multitude of advantages. The strengths of this proposed model encompass scalability, real-time analysis, entity behavior analysis, network security monitoring, and effective anomaly detection. The focus is on classifying extensive datasets, comprising information from diverse connected devices such as IDS, smartphones, tablets, smart TVs, and routers. Real-time classification of these datasets holds significant value in making timely and informed cybersecurity decisions. The proposed model incorporates three distinct machine learning algorithms: Decision Stump, Naïve Bayes and TreesJ48. TreesJ48 has emerged as the standout performer for the given dataset, achieving an impressive 96% accuracy rate. It showcased efficiency in handling both continuous and non-continuous data, affirming its robustness in diverse classification tasks and exemplifying its potential to elevate cybersecurity practices.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Schiller E, Aidoo A, Fuhrer J, Stahl J, Ziörjen M, Stiller B. Landscape of IoT security. Computer Science Review. 2022 May 1;44:100467.
2. Revenues from the artificial intelligence (ai) market worldwide, from 2016 to 2025.

Available:https://www.statista.com/statistics/607716/worldwideartificialintelligence-market-revenues/
3. Pacheco J, Ibarra D, Vijay A, Hariri S. IoT security framework for smart water system, in: IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), IEEE, Piscataway, NJ, US. 2017;1285–1292.
4. M. Vega, Internet of things statistics, facts & predictions (2020's update); 2020. Accessed: 2020-10-17. [Online]. Available:https://review42.com/ internet-of-things-stats/. [5] G.D. Maayan, The IoT rundown
5. Gartner. Press Releases; 2017. Available:https://www.gartner.com Accessed 20th June 2023.
6. Khraisat A, Gondal I, Vamplew P, Kamruzzaman J. Survey Of Intrusion Detection Systems: Techniques, Datasets And Challenges. Cybersecurity. 2019 Dec;2(1):1-22.
7. Saranya T, Sridevi S, Deisy C, Chung Td, Khan Ma. Performance analysis of machine learning algorithms in intrusion detection system: A Review. Procedia Computer Science. 2020 Jan 1;171:1251-60.
8. Azeez Na, Bada Tm, Misra S, Adewumi A, Van Der Vyver C, Ahuja R. Intrusion Detection And Prevention Systems: An Updated Review. Data Management, Analytics And Innovation: Proceedings of Icdmai.2019;1(2020):685-96.
9. Aldaej A. Enhancing Cyber Security In Modern Internet Of Things (Iot) Using Intrusion Prevention Algorithm For Iot (Ipai). Ieee. Access. 2019 Jan 20.
10. Vinayakumar R, Alazab M, Soman Kp, Poornachandran P, Al-Nemrat A, Venkatraman S. Deep Learning Approach For Intelligent Intrusion Detection System. Ieee Access. 2019 Apr 3;7:41525-50.
11. Nixon Ik. Standard Penetration Test State-Of-The-Art Report. Inpenetration Testing. Routledge. 2021 Feb 25;1:3-22.
12. Heiding F, Süren E, Olegård J, Lagerström R. Penetration Testing Of Connected Households. Computers & Security. 2023 Mar 1;126:103067.
13. Rahman A, Williams L. A Bird's Eye View of knowledge needs related to penetration testing. Inproceedings of the 6th Annual Symposium on Hot Topics in the Science of Security. 2019 Apr 1;1-2.
14. Martínez Torres J, Iglesias Comesaña C, García-Nieto Pj. Machine learning techniques applied to cybersecurity. International Journal of Machine Learning and Cybernetics. 2019 Oct;10:2823-36.
15. Sarker Ih. Machine Learning: Algorithms, Real-World Applications and Research Directions. Sn Computer Science. 2021 May;2(3):160.
16. Alloghani M, Al-Jumeily D, Mustafina J, Hussain A, Aljaaf Aj. A Systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and Unsupervised Learning for Data Science. 2020:3-21.
17. Mahesh B. Machine Learning Algorithms-A Review. International Journal of Science and Research (Ijsr).[Internet]. 2020 Jan; 9(1):381-6.
18. Vabalas A, Gowen E, Poliakoff E, Casson Aj. Machine learning algorithm validation with a Limited Sample Size. Plos One. 2019 Nov 7;14(11):E0224365**.**
19. Jhaveri Rh, Revathi A, Ramana K, Raut R, Dhanaraj Rk. A Review on machine learning strategies for real-world engineering applications. Mobile Information Systems. 2022 Aug 28;2022.
20. Mohammed M, Khan Mb, Bashier Eb. Machine Learning: Algorithms and Applications. Crc Press; 2016 Aug 19.
21. Singh SP, Piras G, Viriyasitavat W, Kariri E, Yadav K, Dhiman G, Vimal S, Khan SB. Cyber Security and 5G-assisted Industrial Internet of things using Novel Artificial Adaption based Evolutionary Algorithm. Mobile Networks and Applications. 2023 Aug 9:1-7.
22. Andrade Ro, Yoo Sg. Cognitive Security: A Comprehensive Study of Cognitive Science in Cybersecurity. Journal of Information Security and Applications. 2019 Oct 1;48:102352.
23. Buczak, AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials. 2015;18(2):1153–1176.
24. Sarker Ih, Kayes As, Badsha S, Alqahtani H, Watters P, Ng A. Cybersecurity Data Science: An Overview From Machine Learning Perspective. Journal Of Big Data. 2020 Dec;7:1-29.
25. Kantarcioglu M, Xi B. Adversarial Data Mining: Big Data Meets Cyber Security. Inproceedings of the 2016 ACM Sigsac Conference on Computer and

Communications Security. 2016 Oct 24; 1866-1867.
26. Reddy Yh, Ali A, Kumar Pv, Srinivas Mh, Netra K, Achari Vj, Varaprasad R. A Comprehensive Survey of internet of things applications, Threats, and Security Issues. South Asian Res J Eng Tech. 2022;4(4):63-77.
27. Makhdoom I, Abolhasan M, Lipman J, Liu Rp, Ni W. Anatomy of threats to the internet Of Things. Ieee Communications Surveys & Tutorials. 2018 Oct 11;21(2): 1636-75.
28. Kuzminykh I, Ghita B, Such Jm. The Challenges with internet of things security for business. Ininternational Conference on Next Generation Wired/Wireless Networking. Cham: Springer International Publishing. 2021 Aug 26;46-58.
29. Svyrydenko D, Możgin W. Hacktivism of the anonymous group as a fighting tool in the context of Russia's War Against Ukraine. Future Human Image. 2022 Jan 1;17:39-46.
30. Sammut C, Webb Gi, Editors. Encyclopedia of machine learning. Springer Science & Business Media; 2011 Mar 28.
31. Alqahtani H, Sarker Ih, Kalim A, Minhaz Hossain Sm, Ikhlaq S, Hossain S. Cyber Intrusion Detection Using Machine Learning Classification Techniques. Incomputing Science, Communication and Security: First International Conference, Coms2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1. Springer Singapore. 2020;121-131.
32. Shaukat K, Luo S, Varadharajan V, Hameed Ia, Xu M. A Survey on machine learning techniques for cyber security in the last decade. Ieee Access. 2020 Dec 2;8:222310-54.
33. Thomas T, Vijayaraghavan Ap, Emmanuel S. Machine Learning Approaches In Cyber Security Analytics. Singapore: Springer; 2020 Nov.
34. Mahdavifar S, Maleki N, Lashkari Ah, Broda M, Razavi Ah. Classifying Malicious Domains Using Dns Traffic Analysis. In2021 IEEE Intl conf on dependable, autonomic and secure computing, Intl conf on pervasive intelligence and computing, Intl conf on cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (Dasc/Picom/ Cbdcom/Cyberscitech). IEEE. 2021 Oct 25;60-67.
35. Rizvi S, Scanlon M, Mcgibney J, Sheppard J. Application of Artificial Intelligence To Network Forensics: Survey, Challenges And Future Directions. Ieee Access. 2022 Oct 13;10:110362-84.
36. Dasgupta D, Akhtar Z, Sen S. Machine learning in cybersecurity: A comprehensive survey. The Journal of Defense Modeling and Simulation. Dasgupta. 2022 Jan;19(1): 57-106.
37. Sarker IH. Machine learning for intelligent data analysis and automation in cybersecurity: current and future prospects. Annals of Data Science. 2023 Dec;10(6):1473-98.
38. Khan MN, Ara J, Yesmin S, Abedin MZ. Machine learning approaches in cybersecurity. In Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2021. Singapore: Springer Nature Singapore. 2022 Feb 1;345-357.
39. Bharadiya J. Machine Learning in Cybersecurity: Techniques and Challenges. European Journal of Technology. 2023 Jun 2;7(2):1-4.
40. Annamalai C. Combinatorial and Multinomial Coefficients and its Computing Techniques for Machine Learning and Cybersecurity. The Journal of Engineering and Exact Sciences. 2022 Sep 29;8(8): 14713-01i.
41. Berghout T, Benbouzid M, Muyeen SM. Machine learning for cybersecurity in smart grids: A comprehensive review-based study on methods, solutions, and prospects. International Journal of Critical Infrastructure Protection. 2022 Jul 1; 100547.
42. Rege M, Mbah RB. Machine learning for cyber defense and attack. Data Analytics. 2018 Nov 18;2018:83.
43. Ravinder M, Kulkarni V. A Review on Cyber Security and Anomaly Detection Perspectives of Smart Grid. In2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE. 2023 Jan 23;692-697.
44. Sánchez-Zas C, Larriva-Novo X, Villagrá VA, Rodrigo MS, Moreno JI. Design and Evaluation of Unsupervised Machine Learning Models for Anomaly Detection in Streaming Cybersecurity Logs. Mathematics. 2022 Oct 31;10(21):4043.
45. Apruzzese G, Laskov P, Montes de Oca E, Mallouli W, Brdalo Rapa L, Grammatopoulos AV, Di Franco F. The role of machine learning in cybersecurity.

Digital Threats: Research and Practice. 2023 Mar 7;4(1):1-38.
46. Aiyanyo ID, Samuel H, Lim H. A systematic review of defensive and offensive cybersecurity with machine learning. Applied Sciences. 2020 Aug 22;10(17):5811.
47. Abushark YB, Irshad Khan A, Alsolami F, Almalawi A, Mottahir Alam M, Agrawal A, Kumar R, Ahmad Khan R. Cyber security analysis and evaluation for intrusion detection systems. Comput. Mater. Contin. 2022 Jan 1;72:1765-83.
48. Chio C, Freeman D. Machine learning and security: Protecting systems with data and algorithms. O'Reilly Media, Inc; 2018 Jan 26.
49. Wazid M, Das AK, Chamola V, Park Y. Uniting cyber security and machine learning: Advantages, challenges and future research. ICT Express. 2022 Sep 1;8(3):313-21.
50. Macas M, Wu C, Fuertes W. A survey on deep learning for cybersecurity: Progress, challenges, and opportunities. Computer Networks. 2022 Jul 20;212:109032.
51. Halbouni A, Gunawan TS, Habaebi MH, Halbouni M, Kartiwi M, Ahmad R. Machine learning and deep learning approaches for cybersecurity: A review. IEEE Access. 2022 Feb 11;10:19572-85.
52. Chiba Z, Alaoui MS, Abghour N, Moussaid K. Automatic building of a powerful IDS for the cloud based on deep neural network by using a novel combination of simulated annealing algorithm and improved self-adaptive genetic algorithm. International Journal of Communication Networks and Information Security. 2022 Apr 1;14(1):93-117.
53. Handa A, Sharma A, Shukla SK. Machine learning in cybersecurity: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019 Jul;9(4):e1306.
54. Abbas G, Mehmood A, Carsten M, Epiphaniou G, Lloret J. Safety, Security and Privacy in Machine Learning Based Internet of Things. Journal of Sensor and Actuator Networks. 2022 Jul 29;11(3):38.
55. Khan AR, Kashif M, Jhaveri RH, Raut R, Saba T, Bahaj SA. Deep learning for intrusion detection and security of Internet of things (IoT): current analysis, challenges, and possible solutions. Security and Communication Networks. 2022 Jul 9;2022.
56. Rath M, Mishra S. Advanced-level security in network and real-time applications using machine learning approaches. In Research Anthology on Machine Learning Techniques, Methods, and Applications. IGI Global. 2022;664-680.
57. Ali A, Septyanto AW, Chaudhary I, Al Hamadi H, Alzoubi HM, Khan ZF. Applied Artificial Intelligence as Event Horizon Of Cyber Security. In 2022 International Conference on Business Analytics for Technology and Security (ICBATS). IEEE. 2022 Feb 16;1-7.
58. Khraisat A, Gondal I, Vamplew P, Kamruzzaman J. Survey of intrusion detection systems: Techniques, datasets and challenges. Cybersecurity. 2019 Dec; 2(1):1-22.
59. Denning De. An Intrusion-Detection Model. Ieee Transactions On Software Engineering. 1987 Feb(2):222-32.
60. Blank S. Cyber War and Information War A La Russe. Understanding Cyber Conflict: Fourteen Analogies. 2017 Oct:1-8.
61. Krügel C, Toth T, Kirda E. Service Specific Anomaly Detection for network intrusion detection. inproceedings of the 2002 Acm Symposium on Applied Computing. 2002 Mar 11;201-208.
62. Davis Jj, Clark Aj. Data Preprocessing for Anomaly based Network Intrusion Detection: A Review. Computers & Security. 2011 Sep 1;30(6-7):353-75.
63. Iglesias F, Zseby T. Analysis of network traffic features for anomaly detection. Machine Learning. 2015 Oct;101:59-84.
64. Bhuyan Mh, Bhattacharyya Dk, Kalita Jk. A Multi-Step Outlier-Based Anomaly Detection Approach to Network-Wide Traffic. Information Sciences. 2016 Jun 20;348:243-71.
65. Zhong Y, Chen W, Wang Z, Chen Y, Wang K, Li Y, Yin X, Shi X, Yang J, Li K. Helad: A Novel Network Anomaly Detection Model Based on Heterogeneous Ensemble Learning. Computer Networks. 2020 Mar 14;169:107049.
66. Aslahi-Shahri Bm, Rahmani R, Chizari M, Maralani A, Eslami M, Golkar Mj, Ebrahimi A. A Hybrid Method Consisting of Ga And Svm for Intrusion Detection System. Neural Computing and Applications. 2016 Aug;27:1669-76.
67. Lu KD, Wu ZG. Genetic algorithm-based cumulative sum method for jamming attack detection of cyber-physical power systems.

IEEE Transactions on Instrumentation and Measurement. 2022 Jun 27;71:1-0.
68. Sarker Ih, Kayes As, Watters P. Effectiveness Analysis of Machine Learning Classification Models For Predicting Personalized Context-Aware Smartphone Usage. Journal of Big Data. 2019 Dec;6(1):1-28.
69. Koc L, Mazzuchi Ta, Sarkani S. A Network Intrusion Detection System Based on a Hidden Naïve Bayes Multiclass Classifier. Expert Systems with Applications. 2012 Dec 15;39(18):13492-500.
70. Habibi Lashkari A, Kaur G, Rahali A. Didarknet: A Contemporary Approach to Detect and Characterize the Darknet Traffic Using Deep Image Learning. In2020 The 10th International Conference on Communication And Network Security 2020 Nov 27,1-13.
71. Mahdavifar S, Hanafy Salem A, Victor P, Razavi Ah, Garzon M, Hellberg N, Lashkari Ah. Lightweight Hybrid Detection of data exfiltration using Dns Based on Machine Learning. In2021 the 11th International Conference on Communication and Network Security. 2021 Dec 3;80-86.
72. Yavanoglu O, Aydos M. A Review On Cyber Security Datasets For Machine Learning Algorithms. In2017 Ieee International Conference on Big Data (Big Data). IEEE. 2017 Dec 11;2186-2193.
72a. Singhdeo TJ, Reeja SR, Bhavsar A, Satapathy S. Penetration Testing of Web Server Using Metasploit Framework and DVWA. InInternational Conference on Frontiers of Intelligent Computing: Theory and Applications. Singapore: Springer Nature Singapore. 2023 Apr 11;189-199.
73. Markov Z, Russell I. An Introduction to the Weka Data Mining System. Acm Sigcse Bulletin. 2006 Jun 26;38(3):367-8.
74. Taheri S, Mammadov M. Learning the Naive Bayes Classifier with Optimization Models. International Journal of Applied Mathematics and Computer Science. 2013 Dec 1;23(4):787-95.
75. Razdan S, Gupta H, Seth A. Performance Analysis Of Network Intrusion Detection Systems Using J48 And Naive Bayes Algorithms. In 2021 6th International Conference For Convergence In Technology (I2ct). IEEE. 2021 Apr 2;1-7.
76. Bhargava N, Sharma G, Bhargava R, Mathuria M. Decision Tree Analysis on J48 Algorithm for Data Mining. Proceedings of International Journal of Advanced Research In Computer Science and Software Engineering. 2013 Jun;3(6).
77. Charbuty B, Abdulazeez A. Classification Based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends. 2021 Mar 24;2(01):20-8.
78. Naseem R, Khan B, Ahmad A, Almogren A, Jabeen S, Hayat B, Shah Ma. Investigating Tree Family Machine Learning Techniques for a Predictive System to Unveil Software Defects. Complexity. 2020 Nov 30;2020:1-21.
79. Ahishakiye E, Taremwa D, Omulo Eo, Niyonzima I. Crime Prediction Using Decision Tree (J48) Classification Algorithm. International Journal of Computer and Information Technology. 2017 May;6(3):188-95.
80. Machine Learning. In2020 4th International Conference on Intelligent Computing and Control Systems (Iciccs), IEEE. 2020 May 13;1264-1270.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/111124