EAI Endorsed Transactions
on Internet of Things

Research Article EAI.EU

Analyse and Predict the Detection of the Cyber - Attack Process by Using a Machine-Learning Approach

Charanjeet Singh[1*], Ravinjit Singh[2], Shivaputra[3], Mohit Tiwari[4] and Bramah Hazela[5]

[1]Electronics and Communication Department, Deenbandhu Chhotu Ram University of Science and Technology, Murthal
[2]Faculty of Business & Management (FABM), Akademi Laut Malaysia (ALAM), Malaysia
[3]Department of Electronics and Communication Engineering, Dr. Ambedkar Institute of Technology, Bengaluru
[4]Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, Delhi
[5]Amity School of Engineering & Technology Lucknow, Amity University, Uttar Pradesh, India

Abstract

Crimes committed online rank among the most critical global concerns. Daily, they cause country and citizen economies to suffer massive financial losses. With the proliferation of cyber-attacks, cybercrime has also been on the rise. To effectively combat cybercrime, it is essential to identify its perpetrators and understand their methods. Identifying and preventing cyber-attacks are difficult tasks. To combat these concerns, however, new research has produced safety models and forecast tools grounded on artificial intelligence. Numerous methods for predicting criminal behaviour are available in the literature. While they may not be perfect, they may help in cybercrime and cyber-attack tactic prediction. To find out whether an attack happened and, if so, who was responsible, one way to look at this problem is by using real-world data. There is data about the crime, the perpetrator's demographics, the amount of property damaged, and the entry points for the assault. Potentially, by submitting applications to forensics teams, victims of cyber-attacks may get information. This study uses ML methods to analyse cyber-crime consuming two patterns and to forecast how the specified characteristics will furnish to the detection of the cyber-attack methodology and perpetrator. Based on the comparison of eight distinct machine-learning methods, one can say that their accuracy was quite comparable. The Support Vector Machine (SVM) Linear outperformed all other cyber-attack tactics in terms of accuracy. The initial model gave us a decent notion of the assaults that the victims would face. The most successful technique for detecting malevolent actors was logistic regression, according to the success rate. To anticipate who the perpetrator and victim would be, the second model compared their traits. A person's chances of being a victim of a cyber-attack decrease as their income and level of education rise. The proposed idea is expected to be used by departments dealing with cybercrime. Cyber-attack identification will also be made easier, and the fight against them will be more efficient.

Keywords: Machine learning, Cyber-attacks, Cybercrime, SVM

Received on 11 December 2023, accepted on 01 March 2024, published on 08 March 2024

doi: 10.4108/eetiot.5345

*Corresponding author. Email: charanjeet.research@gmail.com

1. Introduction

The goalmouth of machine learning is a forward forecast based on past data. Machine learning is a subfield of AI that innovates more accuracy without specific instructions—basics of Machine Learning, Python implementation of a simple machine learning algorithm. Machine learning aims to create computer programs that can adapt to new data. The usage of specialised algorithms is vital to the training and forecast processes [1, 2, and 30].

An algorithm is fed the training data and utilises that information to make predictions about new test data. In a broad sense, there are three distinct types of machine learning. Learning may be divided into three categories: supervised, unsupervised, and reinforced [3, 4, and 5]. Supervised learning requires data to be labelled by a human before it can be used by the program to learn. No labels are used in unlabelled learning. To the learning algorithm, it was a source of. This algorithm must determine how the input data should be gathered. Finally, Reinforcement learning expands via dynamic interactions with its surroundings and incorporating positive and negative feedback [7, 8, and 29].

Python allows data scientists to use ML techniques to find methods that provide valuable insights. Many algorithms may be randomly divided into two groups, supervised and unsupervised learning, and their ability to "learn" from data to make predictions. Forecasting what sort, of a set of data points fits to be called "classification."[9, 10, 28]. Ghosh et al.'s 2023 study [29] focuses on "Water Quality Assessment Through Predictive Machine Learning", highlighting the use of machine learning for analyzing and predicting water quality parameters. In "Unraveling the Heterogeneity of Lower-Grade [30] Gliomas," Rahat, Ghosh, and colleagues (2023) delve into deep learning-assisted segmentation and genomic analysis of brain MR images, offering new insights into this medical condition. Potato Leaf Disease [31] Recognition and Prediction using Convolutional Neural Networks," by Ghosh, Rahat, and team (2023), showcases the application of convolutional neural networks in accurately identifying diseases in potato leaves.

Aims, labels, and groupings are alternative names for classes. Creating a near-perfect mapping function between discrete output variables (Y) and continuous input variables (X) is a crucial part of predictive modelling in the classification area [11, 15, 26, 27]. Classification is a kind of supervised learning in statistics and machine learning; it involves teaching a computer to classify incoming data points according to established rules. It's conceivable that this data set includes non-binary information in addition to binary data (such as the subject's gender or the email's spam status) [12, 14]. There are several applications where classification problems arise, such as biometric identification, document categorization, voice recognition, and handwriting analysis [13, 16; 17–12; 18–19].

2. Methodology

Officers who focus on the specific kind of crime that a person has suffered are sought after by the public. The unit's database takes a comprehensive record of these statistics. These crimes are documented by the police in detail, including the nature, manner, year, etc. They collect data, sort it into categories, fix analyses, and create visual illustrations. When many cyberattacks are launched against a single target at the same time, only one attack is logged.

Examining the event's details, rather than the statistics, will reveal whether or not several techniques were used. Although many crimes are recorded in this method, cybercrime has become increasingly important in recent years. There has been little success in preventing cybercrime, although it has resulted in significant material and moral harm. Because few previous investigations into cybercrime have used hard data, this is the topic chosen. The goal of the suggested model is to use information about the victim to predict their possibility of being a victim of crime in the future. As a bonus, it will help law enforcement to expect and study cybercrime suspects, offenders and victims. The method also helps avoid any unintended consequences. The study's findings will allow for targeted interventions and better public awareness of potential dangers. The data collection was comprised of actual incidents of cybercrime that happened in Elaz province between 2018 and 2022[22]. It was difficult to get hands-on real information and wash it up so that it could be analysed using ML techniques. When the data set was obtained, every nuance of cybercrime was scrutinised. Data science techniques were used to cut out the unnecessary bits [23, 24].

In Fig. 1, see the total amount of damages, offences, and entry points used by attackers in the dataset. In addition, the information about these four characteristics is organised by colour. Predictions were done in Python utilizing this data from several different modules.

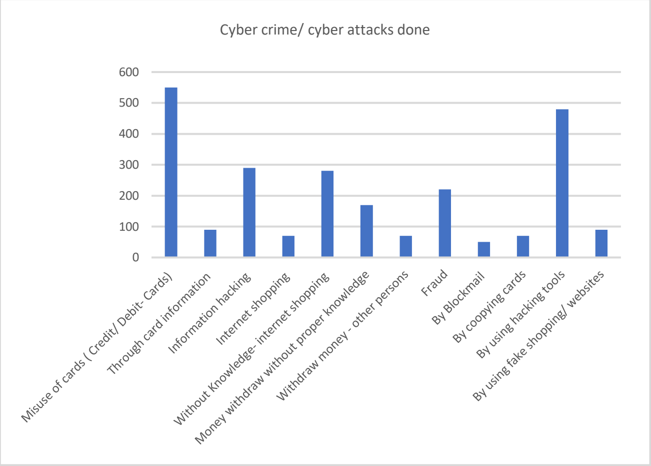


Figure 1. The datasets use a variety of approaches to generate the number of cybercrimes or cyber-attacks

Data was visualised using this Programme, which made use of the program's main libraries, including Numpy, Pandas, and Matplotlib. The key benefits of utilizing the ML method described in this study include the following: the ability to emit criminal strategies, extract complex data relationships, produce results that are impossible for humans to predict, and identify models in both unstructured and structured data. [25, 26].

Selecting relevant and interconnected characteristics from a dataset is known as feature selection. When training information for machine learning, it helps save time and space. Training times may lengthen if the characteristics are chosen incorrectly, raising the model's error rate and complicating its interpretation. Our dataset's features and elements have been established. Each item in Table 1 corresponds to an actual crime. They included these characteristics in the data used for training— maximum, median, and lowest values for our dataset's attributes Fig. 2.

Table 1. Each crime method

Item	Type of Offense
Crime	Payment processing using debit or credit cards; unauthorized access to data; hacking
Gender type	Gender: male, female, or unknown
Age	under 26 years; 27-39 age 40-50 age factor
Income factor	minimal/ Moderate/maximum
Job	employed / self-employed/Homemaker/ retired / etc.
Marriage	Sole / Pair
Education Qualification	Prime/ Higher/ Undergraduate
Harming	Shop on the internet without doing any research. An unknown individual withdrew funds
Attack	Abuse of a credit or ATM card Online financial management Interception of sensitive information using digital means Unwanted messages Buying on an unofficial website

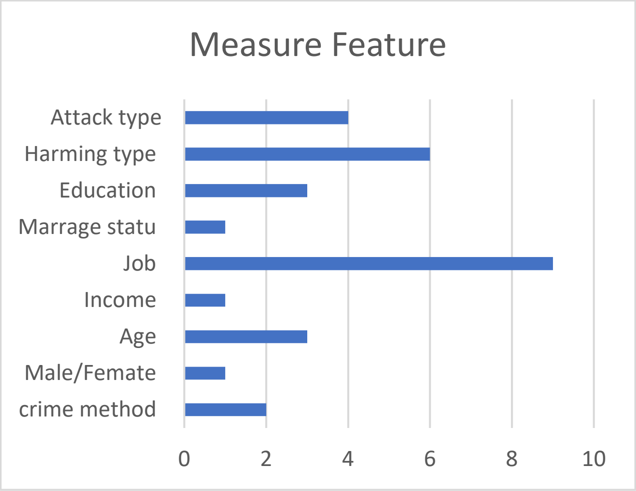


Figure 2 Measuring actual crime

When features are standardized, they are rescaled to fit a normal distribution. It would be best to do this before using any machine learning techniques. Standardized the information and assigned the values between 1 and 10 to the columns to reflect the wide range of information included. The data for algorithms was optimized using Python's Standard Scaler ()—the two-way causal connection between criminal activity and harm, violence, and means of assault. Eighty per cent was used as training data, while twenty per cent was used as test data.

In the first model, forecast the attack technique by providing details about the sufferer, the offence, the offender, the sufferer's age and gender, profession and earnings, marital, education level, and the assault.

The second model aimed to identify potential perpetrators based on demographic information such as age, gender, income, profession, marital status, education, assault type, severity of damages, and aggression method.

3. Result with discussion

This study attempts to appropriately analyse incident information to prevent lawbreaking and apprehend individuals responsible for it. The focus of this research is on using the findings from the data analysis to counter criminal activity. These findings will illuminate the investigations conducted by law enforcement and uncover any previously hidden information learning methods can analyze victim data, cybercrime methods, and recognized perpetrator status to determine if the same cybercriminal is responsible for consecutive attacks. Over the years, various techniques have been used to uncover the damages suffered by victims of cyber incidents in Elaz province. Damages incurred by each victim in the dataset were calculated by adding the total amounts for each year. The decline in such events, mainly after 2018, is widely attributed to the deterrent provided by the legislation and awareness campaigns. As shown in Fig. 3, the economic damages incurred due to cyberattacks in Elaz are staggering. The losses above highlight the need to address cyber security and attack techniques.

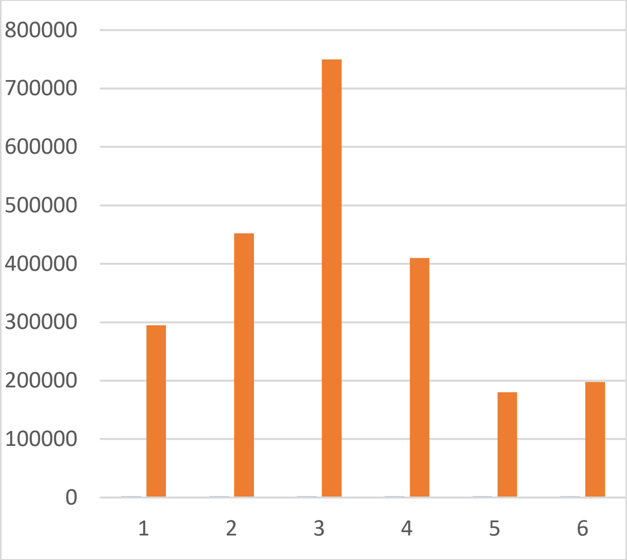
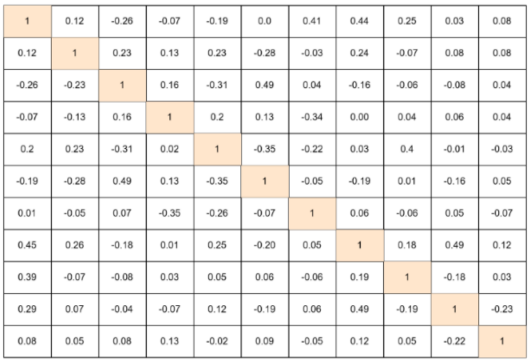


Figure 3 Economic damage of cybercrime

Outcomes from SVM (Kernel), Logistic Reversion, XGBoost, RF, SVM (Linear), KNN, DT, and NB, among others, which are shown here. Decide the Pearson correlation constant by mentioning the instance in Fig. 4. Nearly every conceivable combination of variables is shown by this correlation matrix to have strong correlations.



1	0.12	-0.26	-0.07	-0.19	0.0	0.41	0.44	0.25	0.03	0.08
0.12	1	0.23	0.13	0.23	-0.28	-0.03	0.24	-0.07	0.08	0.08
-0.26	-0.23	1	0.16	-0.31	0.49	0.04	-0.16	-0.06	-0.08	0.04
-0.07	-0.13	0.16	1	0.2	0.13	-0.34	0.00	0.04	0.06	0.04
0.2	0.23	-0.31	0.02	1	-0.35	-0.22	0.03	0.4	-0.01	-0.03
-0.19	-0.28	0.49	0.13	-0.35	1	-0.05	-0.19	0.01	-0.16	0.05
0.01	-0.05	0.07	-0.35	-0.26	-0.07	1	0.06	-0.06	0.05	-0.07
0.45	0.26	-0.18	0.01	0.25	-0.20	0.05	1	0.18	0.49	0.12
0.39	-0.07	-0.08	0.03	0.05	0.06	-0.06	0.19	1	-0.18	0.03
0.29	0.07	-0.04	-0.07	0.12	-0.19	0.06	0.49	-0.19	1	-0.23
0.08	0.05	0.08	0.13	-0.02	0.09	-0.05	0.12	0.05	-0.22	1

Figure 4. Confusing Matrix

Table 2: Model 1- Presentation of ML Method

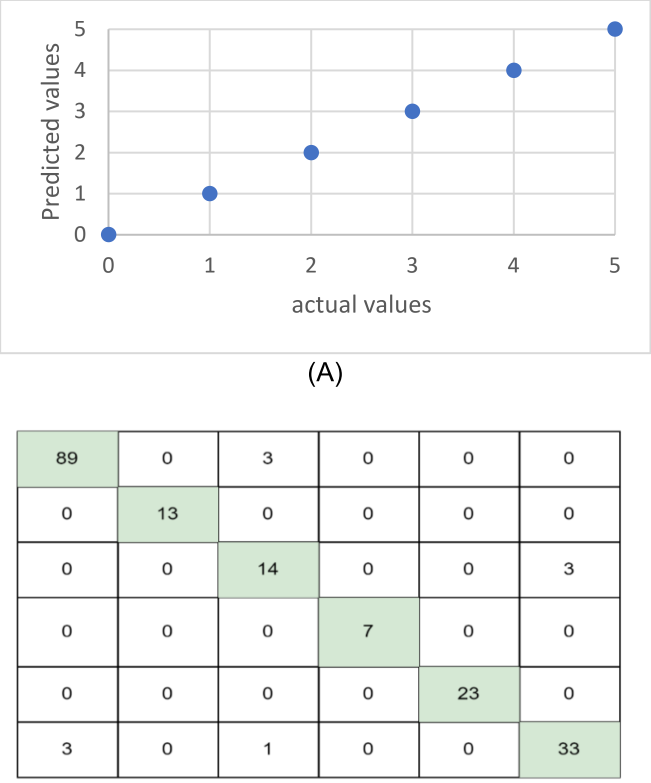
	Accuratenes s %	Exactnes s %	Recollec t %	F1-score %
LR	93.12	94.25	93.23	94.2
KNN	91.5	71.56	76.25	72.12
SVML	95.12	95.22	95.25	95.55
SVMK	92.67	92.56	92.65	92.52
NB	81.55	81.54	81.54	8216 4
DT	92.65	92.6	92.57	92.65
RF	94.89	94.88	94.84	94.68
XGBOOS T	93.45	92.86	92.25	92.66

After the dataset was trained, it was tested across all techniques to kick off the experiment. Incorporate quality control and precise standards as well. To determine the F1 score, precision, accuracy, and recall, compare the predicted values with the test data.

Table 2 shows the efficacy, recall, accuracy, and F1 score of the initial model in predicting the assault plan. There was a clear winner when it came to prediction accuracy; SVML achieved 95.55%. At the margin of error, the SVML method outperformed the others, including radio frequency, LR, the XGBoost SVMK, KNN, and NB. With an 82.54% success rate, New Brunswick ranked worst. Alternative algorithms to NB produced similar results. With Figure 5A showing the distribution graph of the observed values and Figure 5B showing the error matrix, see that the SVML approach was successful in generating the predicted values.

When the comparison of the recall, method accuracy and F1 scores, the SVML algorithm again came out on top, albeit by a narrow margin. Obtain results over 93% using any of SVMK, LR, RF, DT, or XGBoost; their respective performances were quite similar. The worst performing KNN and NB scored around 9% lower than the others. In most cases, the output from each algorithm was satisfactory. These findings demonstrated the viability of using a machine-learning strategy to foretell the vector of a cyberattack. The proposed model would allow users to forecast which crimes a specific individual will fall victim to depending on data about that individual. May also Group early warning systems.

When a person's characteristics are input into the proposed model (Table 2), it will be possible to anticipate the types of crimes to which that individual would be exposed. May also develop Group early warning systems.



(A)

89	0	3	0	0	0
0	13	0	0	0	0
0	0	14	0	0	3
0	0	0	7	0	0
0	0	0	0	23	0
3	0	1	0	0	33

(B)

Figure 5 (A) 1st method compares with real values (B) Confused matrix

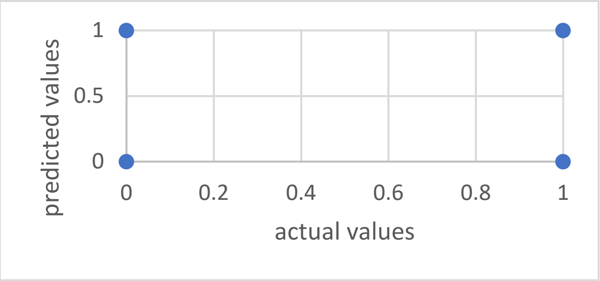
Table 3: Method 2 presentation in ML method

	Correctnes s %	Exactnes s %	Recal l %	F1scor e %
LR	66.52	61.25	61.23	60.56
KNN	65.23	57.46	57.88	57.14
SVML	65.66	66.81	65.85	64.89
SVMK	65.08	66.78	65.88	63.85
NB	63.15	58.29	58.32	56.24
DT	63.26	64.88	63.41	63.57
RF	64.25	64.55	64.26	63.21
XGBOOS T	65.33	66.22	67.32	65.44

Table 3 shows the F1 scores, recall, accuracy, and precision of the second forecast algorithms. The precision was achieved using algorithms like LR (65.52%), SVMK (1.39%), SVML (0.927%), KNN (1.54%), XGBoost (2.43%), RF (3.35%), and DT (3.35%).

Every algorithm was very close to NB's performance, even though NB was the worst. Figure 6B displays the chart of the actual values and the probable values obtained by the SVML technique, though Figure 6A shows the fault matrix.

Despite NB's poor performance, the other algorithms came very close to matching it. Both the real and predicted values created by the SVML approach are shown in Figure 6A's distribution graph, and Figure 6B's error matrix.



(A)

39	41
24	72

(B)

Figure 6 (A) Contrast fault matrices (B) Confusion matrix

The dataset size is a constraint of the study since it contains actual data. The availability of temporal data allows for the estimation of time series, although this data is essential. Similarly, if the technical details of the attacks were recorded by the authorities, precise estimates may help in identifying the perpetrator.

4. Conclusion

This study proposes a strategy for detecting and preventing cyber-attacks using machine learning algorithms with historical data on such assaults. The system predicts the demographics of potential victims and the kind of attacks they may face. Machine learning strategies are sufficiently compelling. The linear SVM approach is the most effective.

About 61% of the time, the method can properly predict the attacker who will launch a cyber-attack. May use other forms of artificial intelligence to attempt to improve this figure, believe that bringing attention to malware and social engineering assaults is essential. There was an inverse relationship between the victim's degree of education and money and the likelihood of a cyber-attack. The research's overarching goal is to help law implementation agencies become more proactive in the battle against cybercrime by providing them with better, more efficient tools. By evaluating the characteristics of the assault victims that surfaced in this investigation, may develop warning methods and new training systems for those with comparable qualities.

References

1. Bilen, Abdulkadir & Özer, Ahmet. (2021). Cyber-attack method and perpetrator prediction using machine learning algorithms. PeerJ Computer Science. 7. e475. 10.7717/peerj-cs.475.
2. Al-majed, Rasha & Ibrahim, Amer & Abualkishik, Abedallah & Mourad, Nahia & Almansour, Faris. (2022). Using machine learning algorithm for detection of cyber-attacks in cyber physical systems. Periodicals of Engineering and Natural Sciences (PEN). 10. 261. 10.21533/pen.v10i3.3035.
3. Mazhar, T.; Irfan, H.M.; Khan, S.; Haq, I.; Ullah, I.; Iqbal, M.; Hamam, H. Analysis of Cyber Security Attacks and Its Solutions for the Smart grid Using Machine Learning and Blockchain Methods. Future Internet 2023, 15, 83. https://doi.org/10.3390/fi15020083
4. Sarker, I.H. Machine Learning for Intelligent Data Analysis and Automation in Cybersecurity: Current and Future Prospects. Ann. Data. Sci. (2022). https://doi.org/10.1007/s40745-022-00444-2
5. A. Alshehri, N. Khan, A. Alowayr and M. Yahya Alghamdi, "Cyberattack detection framework using machine learning and user behavior analytics," Computer Systems Science and Engineering, vol. 44, no.2, pp. 1679–1689, 2023.
6. Amjad Rehman, Tanzila Saba, Muhammad Zeeshan Khan, Robertas Damaševičius, Saeed Ali Bahaj, "Internet-of-Things-Based Suspicious Activity Recognition Using Multimodalities of Computer Vision for Smart City Security", Security and Communication Networks, vol. 2022, Article ID 8383461, 12 pages, 2022. https://doi.org/10.1155/2022/8383461
7. Liu Qiang, Qu Xiaoli, Wang Dake, Abbas Jaffar, Mubeen Riaqa, Product Market Competition and Firm Performance: Business Survival Through Innovation and Entrepreneurial Orientation Amid COVID-19 Financial Crisis, Frontiers in Psychology, 12 , 2022, ISSN-1664-1078, 10.3389/fpsyg.2021.790923.
8. URL=https://www.frontiersin.org/articles/10.3389/fpsyg.2021.790923
9. Ibor, A.E., Oladeji, F.A., Okunoye, O.B. et al. Conceptualisation of Cyberattack prediction with deep learning. Cybersecur 3, 14 (2020). https://doi.org/10.1186/s42400-020-00053-7
10. Yirui Wu, Dabao Wei, Jun Feng, "Network Attacks Detection Methods Based on Deep Learning Techniques: A Survey", Security and Communication Networks, vol. 2020, Article ID 8872923, 17 pages, 2020. https://doi.org/10.1155/2020/8872923
11. Delplace, Antoine, Sheryl Hermoso, and Kristofer Anandita. "Cyber-attack detection thanks to machine learning algorithms." arXiv preprint arXiv: 2001.06309 (2020).

12. McCarthy A, Ghadafi E, Andriotis P and Legg P. (2023). Defending against adversarial machine learning attacks using hierarchical learning. Journal of Information Security and Applications. 72: C.
13. Ahsan, M.; Nygard, K.E.; Gomes, R.; Chowdhury, M.M.; Rifat, N.; Connolly, J.F. Machine Learning Techniques in Cybersecurity. Encyclopedia. Available online: https://encyclopedia.pub/entry/25675 (accessed on 30 April 2023).
14. Kenfack, P.D.B., Mbakop, F.K. and Eyong-Ebai, E. (2021) Implementation of Machine Learning Method for the Detection and Prevention of Attack in Supervised Network. Open Access Library Journal, 8, 1-25. doi: 10.4236/oalib.1108000.
15. AlZubi, Ahmad Ali, Mohammed Al-Maitah, and Abdulaziz Alarifi. "Cyber-attack detection in healthcare using cyber-physical system and machine learning techniques." Soft Computing 25.18 (2021): 12319-12332.
16. Zhao L, Zhu D, Shafik W, et al. Artificial intelligence analysis in cyber domain: A review. International Journal of Distributed Sensor Networks. 2022; 18(4). doi:10.1177/15501329221084882
17. Narayan, Valliammal, and Barani Shaju. "Malware and Anomaly Detection Using Machine Learning and Deep Learning Methods." Research Anthology on Machine Learning Techniques, Methods, and Applications, edited by Information Resources Management Association, IGI Global, 2022, pp. 149-176. https://doi.org/10.4018/978-1-6684-6291-1.ch010
18. Ahmad Naim Irfan, Suriayati Chuprat, Mohd Naz'ri Mahrin, Aswami Ariffin. (2022) Taxonomy of Cyber Threat Intelligence Framework. 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), pages 1295-1300.
19. Aksu, Dogukan, and M. Ali Aydin. "Detecting port scan attempts with comparative analysis of deep learning and support vector machine algorithms." 2018 International congress on big data, deep learning and fighting cyber terrorism (IBIGDELFT). IEEE, 2018.
20. Khuphiran, Panida, et al. "Performance comparison of machine learning models for DDoS attacks detection." 2018 22nd International Computer Science and Engineering Conference (ICSEC). IEEE, 2018.
21. Arshey, M., and KS Angel Viji. "Thwarting cyber-crime and phishing attacks with machine learning: a study." 2021 7th international conference on advanced computing and communication systems (ICACCS). Vol. 1. IEEE, 2021.
22. Shivlal Mewada, Anil Saroliya, N. Chandramouli, T. Rajasanthosh Kumar, M. Lakshmi, S. Suma Christal Mary, Mani Jayakumar, "Smart Diagnostic Expert System for Defect in Forging Process by Using Machine Learning Process", Journal of Nanomaterials, vol. 2022, Article ID 2567194, 8 pages, 2022. https://doi.org/10.1155/2022/2567194
23. Rege, Manjeet, and Raymond Blanch K. Mbah. "Machine learning for cyber defense and attack." Data Analytics 2018 (2018): 83.
24. P. Patro, R. Azhagumurugan, R. Sathya, K. Kumar, T. R. Kumar and M. V. S. Babu, "A hybrid approach estimates the real-time health state of a bearing by accelerated degradation tests, Machine learning," 2021 Second International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India, 2021, pp. 1-9, doi: 10.1109/ICSTCEE54422.2021.9708591.
25. Choudhary, Atul S., Pankaj P. Choudhary, and Shrikant Salve. "A Study on Various Cyber Attacks and a Proposed Intelligent System for Monitoring Such Attacks." 2018 3rd International Conference on Inventive Computation Technologies (ICICT). IEEE, 2018.
26. Kumari, Maya. "Application of Machine Learning and Deep Learning in Cybercrime Prevention—A Study." Int. J. Trend Res. Dev (2019): 1-4.
27. Saharkhizan, Mahdis, et al. "An ensemble of deep recurrent neural networks for detecting IoT cyber-attacks using network traffic." IEEE Internet of Things Journal 7.9 (2020): 8852-8859.
28. Swaminathan, Aravind, et al. "Prediction of Cyber-attacks and Criminality Using Machine Learning Algorithms." 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT). IEEE, 2022.
29. Ghosh, H., Tusher, M.A., Rahat, I.S., Khasim, S., Mohanty, S.N. (2023). Water Quality Assessment Through Predictive Machine Learning. In: Intelligent Computing and Networking. IC-ICN 2023. Lecture Notes in Networks and Systems, vol 699. Springer, Singapore. https://doi.org/10.1007/978-981-99-3177-4_6
30. Rahat IS, Ghosh H, Shaik K, Khasim S, Rajaram G. Unraveling the Heterogeneity of Lower-Grade Gliomas: Deep Learning-Assisted Flair Segmentation and Genomic Analysis of Brain MR Images. EAI Endorsed Trans Perv Health Tech [Internet]. 2023 Sep. 29 [cited 2023 Oct. 2];9. https://doi.org/10.4108/eetpht.9.4016
31. Ghosh H, Rahat IS, Shaik K, Khasim S, Yesubabu M. Potato Leaf Disease Recognition and Prediction using Convolutional Neural Networks. EAI Endorsed Scal Inf Syst [Internet]. 2023 Sep. 21 https://doi.org/10.4108/eetsis.3937