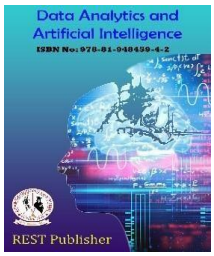
Data Analytics and Artificial Intelligence
Vol: 4(1), 2024
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/daai/
DOI: https://doi.org/10.46632/daai/4/1/5



Intelligent Threat Prediction in Cybersecurity Using Linear Regression and Random Forest Regression an Ensemble Learning Approach

Praveen Kumar Bandaru
Staff engineer, Pentangle Tech Services LLC, Farmington Hills, Michigan, United States.
Corresponding Author Email: pkbandaru607@gmail.com

Abstract: *However, due to the complexity of the attacks, it has been observed that traditional models like signature and rule-based are not efficient enough in providing protection for today's IT systems. In this study, the prediction of threat scores for cyber security with regard to critical levels of network parameters like packet rate, conflict score, and logins is examined using LR and RFR. For both training and testing purposes, the number of samples considered was 100. From the statistical analysis, it can be seen that there are very high variations in packet rates, moderate-high variations in conflict scores, and threat scores, indicating the dynamic nature of networks. It can also be observed that correlation analysis shows there is a high positive correlation between conflict scores and threat scores, but there is no impact of login on the risk factor. But the Random Forest Regression model performed well in comparison to others. For the Random Forest Regression model, the value of R^2 score was 0.9526 for the training set and 0.7503 for the testing set. It is evident from this that the Random Forest Regression model has shown superior results due to its high efficiency. Evaluation of both models was carried out on the basis of some other parameters like MSE, RMSE, MAE, Max Error, MSLE, and Median Absolute Error. It is clear that in comparison to the second model, the Random Forest model has shown superior results in the training set. Thus, both models performed similarly for test data sets.*

Keywords: *cybersecurity, threat score prediction, random forest regression, machine learning, anomaly detection*

1. INTRODUCTION

As noted by the North Atlantic Treaty Organization (NATO), the internet has been identified as a vital national resource for countries, while at the same time being part of infrastructure that acts as a major catalyst of socio-economic development. [1] The manipulation of public opinion through the use of cyborgs, trolls, bots, fake news, memes, and other techniques like deepfakes is part of social engineering. Other examples of social engineering include the manipulation of the masses through technology. [2] Some of the problems that have occurred are hacking into the information that is stored regarding the credit/debit card numbers used in making payments, violation of the intellectual property rights of Google, and the leaking of user information. [3] Along with new technology, there are several advantages that are also derived. However, due to the evolution of new technology, a number of people for different reasons are attracted by this technology. [4] Now comes the part where it is necessary to stress the importance of the usage of technologies and information in today's age to fulfill our duties. Nonetheless, it must be noted that there exist some threats due to which the use of technology and information poses dangers, but not everybody is aware of those threats, even hackers included. [5] But rather than that, there may be some strengths and weaknesses that would emerge since although such technologies play an important role in the success of any organization, misuse of them will certainly cause many risks to the operations of any organization. This can be seen even further from the absence of laws on how these technologies can be utilized, since in doing so, organizations become oblivious to what is really important. [6] Regarding the definition of cybersecurity, it would be right to say that all efforts made to prevent attacks or any other activity which can be classified as a crime can be termed as cybersecurity because there

Copyright@ REST Publisher

are some aspects of uncertainty associated with such an attack. [7] All the above-mentioned points are equally important in making the analysis of the easier because they aid in finding out the loopholes in the security system and comparing it to other systems. It is not that difficult to predict an attack from these loopholes found in the security system. [8] Due to the financial crisis, IT infrastructure has been compelled to revise its business models and, in some cases, even reduce expenditure. Although such reduction may not affect productivity significantly, the same cannot be said regarding cybersecurity, since the effects here may be more serious. [9] In all ages of human existence, honesty was never a guarantee. In the material world, we depend on years of experience to shield us from evil influences, guided by legal statutes that determine what is right or wrong. We also have devised numerous ways through technology to safeguard our properties and personal data. [10] Nowadays, the current era is marked by developments in technology, particularly in technological developments coming out of the digital age. [11] It can be said that the term cybersecurity refers to anything that involves people, processes, and technologies that aim to protect things that are digital. With Moore's Law, there is an observation that after two years, the number of items that can be fitted inside a computer chip double. The same thing is expected for technology pertaining to cybersecurity, where there will be doubling. The hackers too have updated their tools and have not spent money doing so or spent very little money in less than six months. [12] However, there is another epoch in cybersecurity. In the future, research should be concentrated on developing a prediction system that allows anticipating the cyber-attack and its consequences in advance, thereby getting rid of reactive strategies in defense. [13] Over the past decade, developments in artificial intelligence and machine learning have been the driving force behind the digital revolution. Machine learning has shifted from supervised to unsupervised learning. Moreover, machine learning has become even more complex in methods like partial, reinforcement, and deep learning. Generative artificial intelligence is currently the latest development in this field. [14] Furthermore, the fast advancements in ICT have facilitated the deployment of smooth integration between hardware and software, indicating the transition from technology-based systems to cyber-physical systems through which the emergence of technologies such as Industry 4.0 is achieved. [15] In contrast, with more and more cybersecurity incidents being reported worldwide, there is a need to secure these digital systems. It is becoming increasingly evident that processes, procedures, and systems alone cannot suffice. [16] The metaverse, for its part, represents the next phase in the development of the internet, which reflects certain features of the physical universe while fully immersing people within it. The metaverse can be described as three-dimensional virtual worlds where people interact via digital avatars that represent them. [17] Cybersecurity means the protection of computers, networks, software, and information systems against cyberattacks and unauthorized access. [18] Internet technologies have been developed with the help of different technologies like software-defined networking, big data, fog computing, and many more. However, due to these different developments of technology, many cyber security threats have come into play which pose a threat to the infrastructure. [19] From the introduction given above, it becomes evident that there is a very important need for space technologies in current times as they are used in many fields like communication, weather prediction, navigation, and security, amongst others. The use of satellites might create some issues in cyberspace as well. [20]

2. MATERIALS AND METHODS

Packet rate: Packet rate is used to describe the speed with which the data packets are processed within a network component such as a switch or router; it is usually expressed in packets per second.
Discrepancy Score: Discrepancy score is used to describe the numeric value that lies between 0 and 1 or 0 and 100 depending on whether there exists some difference from what is expected.
login attempts: Failed login attempts could lead to numerous security risks, including brute-force attacks. These are major security incidents that require identification.
threat score: The threat score is defined as the numerical measure of the security incident, file, or component concerning its risk level. The threat score normally ranges from 0 to 100.

Optimization Techniques

Linear Regression: Linear Regression is one of the basic techniques used in machine learning and statistics and attempts to identify the presence of a linear association between the response variable and predictor variables.
Random Forest Regression: The random forest regression is an advanced and useful machine learning algorithm, which makes predictions using the mean prediction values computed by decision trees within a random forest.

Copyright@ REST Publisher

3. ANALYSIS AND DISCUSSION

The distributions are quite varied in this data set, beginning with different kinds of distributions for eight numerical data. In situations where things seem fine on the surface in terms of numbers, some are increasing too rapidly, while others have even turned out to be negative numbers. Zero figures have also been quite abundant in the data set.

TABLE 1. Descriptive Statistics

	packet rate	anomaly score	login attempts	threat score
count	100	100	100	100
mean	4754.78936	0.497832	28.06	0.819129
std	2945.14517	0.293111	13.447857	0.255996
min	154.668959	0.006952	2	0.220094
25%	2012.68753	0.242005	19.75	0.635226
50%	4695.0103	0.505625	29.5	0.768141
75%	7329.01088	0.766184	38	1.013099
max	9870.18067	0.98565	49	1.435696

Table 1. Descriptive statistics have been obtained for 100 data points in each of the four variables. A very high standard deviation indicates that there is considerable variance in the number of packets per second. However, there seems to be an even distribution of conflict scores. Login scores are confined to a certain range, while the threat scores are consistently high.

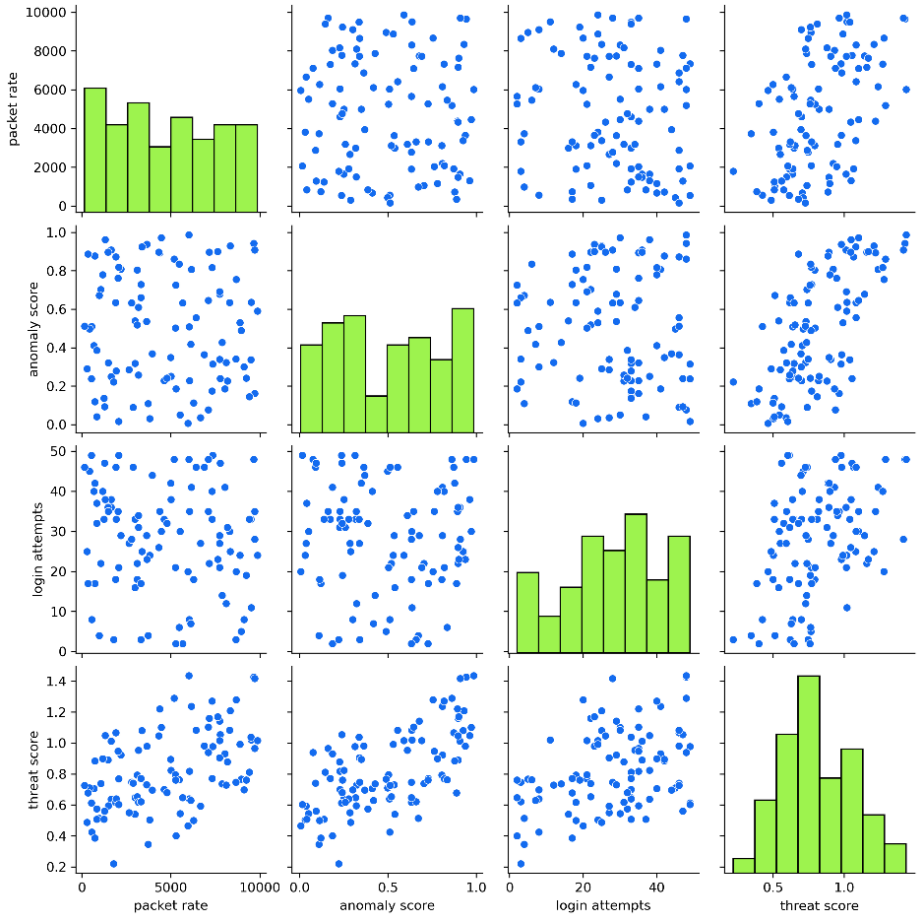


FIGURE 1. Scatter Plot of the Various Cybersecurity

FIGURE 1. The scatterplot matrix shows the connections between packet rate, conflict score, number of logins, and threat score. There is medium dispersion among the variables, and a strong positive correlation exists between conflict score and threat score. The dispersion of points represents variability in behavior, but clustering and outliers signify irregularities and variations in cyber security activities.

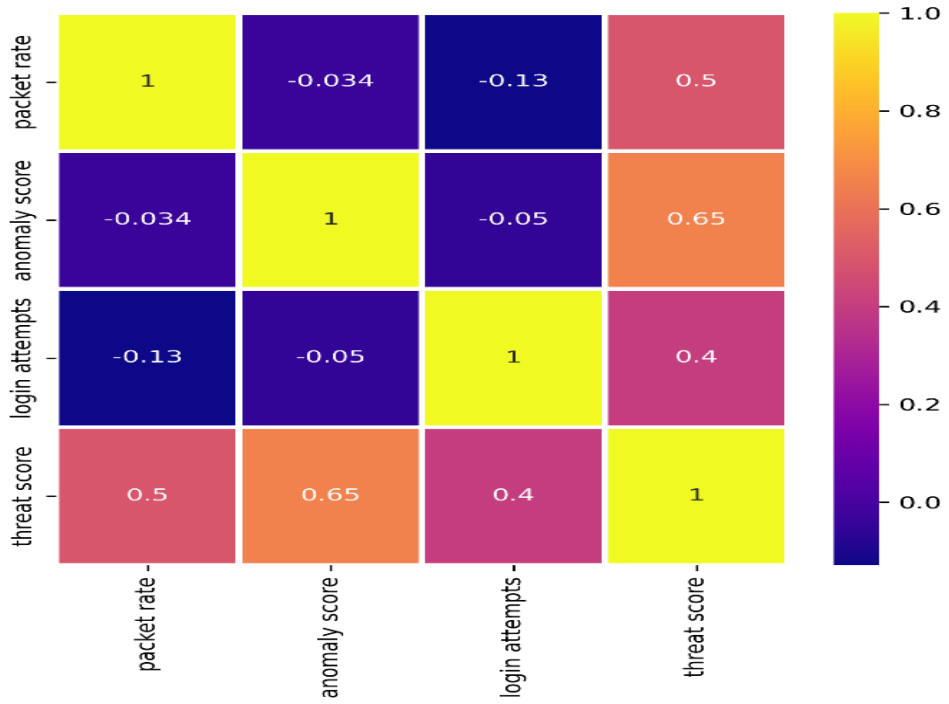


FIGURE 2. Correlation Heat Map on Process Parameters and Effects

FIGURE 2. From the heat map shown above, one can notice that packet rate, conflict score, login count, and threat score have some sort of relation amongst each other. From the heat map, one can see that the relation between threat score, conflict score, and packet rate is extremely high, while on the other hand, the relation between login count and threat score is moderately high.

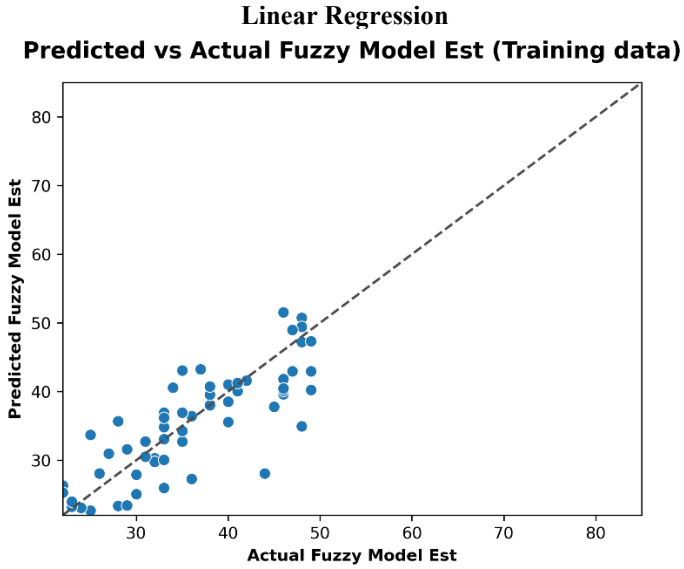


FIGURE 3. Linear Regression (Training data)

Copyright@ REST Publisher

FIGURE 3. As shown in the graph above, it is easy to know the level of the precision of the predictions made using the data given. The points plotted on the straight line depict the precision of the predictions compared to what happens in reality. It may so happen that the predictions may not be too precise and thus have points away from the main pattern, but the prediction method is very precise.

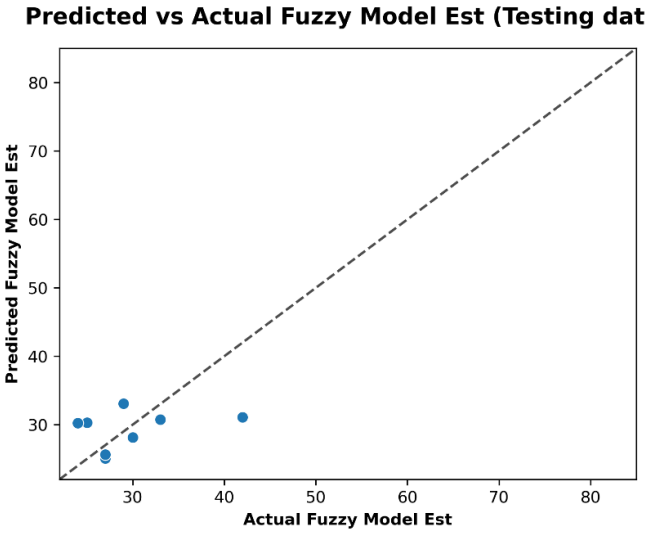


FIGURE 4. Linear Regression (Testing data)

FIGURE 4. From the model of linear regression analysis, the scatter plot showing the actual values and those estimated by the linear regression model has been created. There could be an increase in the number of the scatter points because of the inability of the linear regression model to produce precise results.

TABLE 2. Performance Metrics of Linear Regression (Training Data and Testing Data)

Data	Symbol	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Train	LR	0.8619	0.8619	25.2830	5.0282	3.8041	16.6871	0.0886	2.7634
Test	LR	0.7506	0.7558	32.0346	5.6599	4.4723	10.8840	0.1810	3.1821

TABLE 2. The outcome of linear regression analysis is shown through the scatter plot showing the relationship between the predicted and actual data for the test dataset. One can observe that the dispersion of the data points is larger than the benchmark line, meaning that the accuracy of the model is poorer than the training dataset.

Random Forest Regression

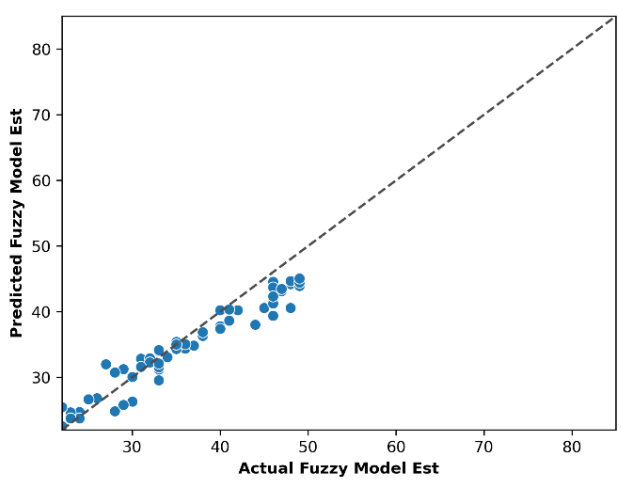


FIGURE 5. Random Forest Regression (Training data)

FIGURE 5. The graph presented above demonstrates the pattern that follows the random forest regression model based on the training data set used for prediction and actuality. It can be seen that the data points lie almost along the diagonal line, which indicates that the accuracy of the regression model is highly precise with minimal errors.

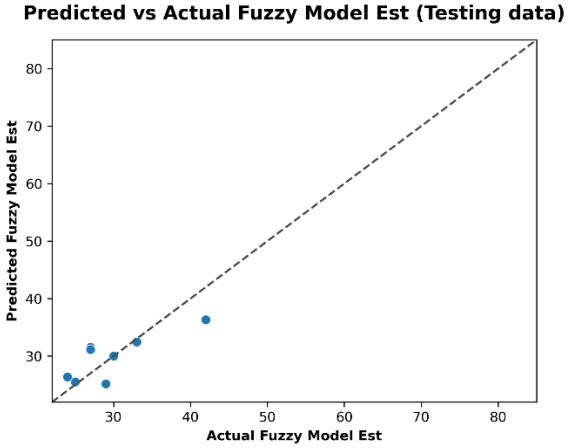


FIGURE 6. Random Forest Regression (Testing data)

FIGURE 6. The performance of Regression using Random Forest algorithm in test data is shown by making comparison between predicted and actual values. The points lie very close to the line y=x, showing high accuracy of the regression model. Although, the dispersion is a little bit more compared to training data indicating low accuracy, the overall trend can be recognized quite well.

TABLE 3. Performance Metrics of Random Forest Regression (Training Data and Testing Data)

Data	Symbol	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Train	RFR	0.9526	0.9526	8.6783	2.9459	2.3529	7.4100	0.0590	1.7525
Test	RFR	0.7503	0.7919	32.0770	5.6637	4.3090	11.0600	0.2855	3.9900

TABLE 3. The above demonstrates the outcome of the performance of the random forest regression model both during training and testing. It is evident from the above that the performance of the model in training resulted in a highly accurate model as seen from high R^2 value as well as lower values of error; hence, the model is a well-fitted one. However, the results from the test depict higher errors and poor performance.

4. CONCLUSION

Due to the increasing complexity and frequency of cyberattacks, there is a need for a solution that can effectively predict future attacks. To achieve this objective, we applied two models to evaluate their ability to predict future cyberattacks. These models include Linear Regression and Random Forest Regression. The two models were used to predict cybersecurity attacks using key factors such as packet rate, conflict score, and login count. It is important to highlight that both models are quite effective when it comes to predictions. For example, Linear Regression delivered accurate predictions, as indicated by the R^2 values of 0.8619 and 0.7506 for training and testing datasets, respectively. This indicates that the LR model can have difficulty modeling the nonlinear characteristics of real-world network traffic. On the other hand, the Random Forest Regression model outperformed LR with an R-squared value of 0.9526 on training data and 0.7503 on test data. The ensemble learning architecture of the model helps to learn complex nonlinear patterns and predict accurately. Other measures like squared error, root mean squared error, mean absolute error, maximum error, mean squared logarithmic error, and median absolute error further confirmed the better performance of RFR, particularly on training data. The above observations clearly illustrate the significance of utilizing advanced machine learning algorithms in the context of cybersecurity, since accurate prediction of the threat becomes essential. It can be clearly observed that the difference and threat scores have high correlation, which again

Copyright@ REST Publisher

validates the reliability of both for predicting the organizational threats. For future work, the researcher could consider using a larger dataset and include more behavioral features and other advanced machine learning techniques such as LSTM and neural networks, among others. Using these methods for real-time IDS may lead to the development of a smart approach to cybersecurity.

REFERENCE

[1]. Martínez Torres, Javier, Carla Iglesias Comesana, and Paulino J. García-Nieto. "Machine learning techniques applied to cybersecurity." International Journal of Machine Learning and Cybernetics 10, no. 10 (2019): 2823-2836.

[2]. Carley, Kathleen M. "Social cybersecurity: an emerging science." Computational and mathematical organization theory 26, no. 4 (2020): 365-381.

[3]. Martínez Torres, Javier, Carla Iglesias Comesana, and Paulino J. García-Nieto. "Machine learning techniques applied to cybersecurity." International Journal of Machine Learning and Cybernetics 10, no. 10 (2019): 2823-2836.

[4]. Bogoviz, Aleksei V., Vladimir S. Osipov, Marina K. Chistyakova, and Maxim Y. Borisov. "Comparative analysis of formation of industry 4.0 in developed and developing countries." In Industry 4.0: industrial revolution of the 21st century, pp. 155-164. Cham: Springer International Publishing, 2018.

[5]. Das, Rammanohar, and Raghav Sandhane. "Artificial intelligence in cyber security." In Journal of Physics: Conference Series, vol. 1964, no. 4, p. 042072. IOP Publishing, 2021.

[6]. Syafrizal, Melwin, Siti R. Selamat, and Nurul A. Zakaria. "Analysis of cybersecurity standard and framework components." International Journal of Communication Networks and Information Security 12, no. 3 (2020): 417-432.

[7]. Scala, Natalie M., Allison C. Reilly, Paul L. Goethals, and Michel Cukier. "Risk and the five hard problems of cybersecurity." Risk Analysis 39, no. 10 (2019): 2119-2126.

[8]. Wilson, Kelce S., and Müge Ayse Kiy. "Some fundamental cybersecurity concepts." IEEE access 2 (2014): 116-124.

[9]. Chronopoulos, Michail, Emmanouil Panaousis, and Jens Grossklags. "An options approach to cybersecurity investment." IEEE Access 6 (2017): 12175-12186.

[10]. Nielsen, Claus Ballegaard, Peter Gorm Larsen, John Fitzgerald, Jim Woodcock, and Jan Peleska. "Systems of systems engineering: Basic concepts, model-based techniques, and research directions." ACM Computing Surveys (CSUR) 48, no. 2 (2015): 1-41.

[11]. Hassan, Azeez Olanipekun, Sarah Kuzankah Ewuga, Adekunle Abiola Abdul, Temitayo Oluwaseun Abrahams, Monisola Oladeinde, and Samuel Onimisi Dawodu. "Cybersecurity in banking: a global perspective with a focus on Nigerian practices." Computer Science & IT Research Journal 5, no. 1 (2024): 41-59.

[12]. Jarrahi, Mohammad Hossein, Sarah Kenyon, Ashley Brown, Chelsea Donahue, and Chris Wicher. "Artificial intelligence: A strategy to harness its power through organizational learning." Journal of Business Strategy 44, no. 3 (2023): 126-135.

[13]. Kibria, Mirza Golam, Kien Nguyen, Gabriel Porto Villardi, Ou Zhao, Kentaro Ishizu, and Fumihide Kojima. "Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks." IEEE access 6 (2018): 32328-32338.

[14]. Gupta, Maanak, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. "From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy." IEEE access 11 (2023): 80218-80245.

[15]. Jiang, Yuning, Manfred A. Jeusfeld, Jianguo Ding, and Elin Sandahl. "Model-Based Cybersecurity Analysis: Y. Jiang et al." Business & Information Systems Engineering 65, no. 6 (2023): 643-676.

[16]. Howard, Newton, and Erik Cambria. "Intention awareness: improving upon situation awareness in human-centric environments." Human-centric Computing and Information Sciences 3, no. 1 (2013): 9.

[17]. Awadallah, Abeer, Khouloud Eledlebi, Mohamed Jamal Zemerly, Deepak Puthal, Ernesto Damiani, Kamal Taha, Tae-Yeon Kim et al. "Artificial intelligence-based cybersecurity for the metaverse: Research challenges and opportunities." IEEE Communications Surveys & Tutorials 27, no. 2 (2024): 1008-1052.

[18]. Sharifani, Koosha, and Mahyar Amini. "Machine learning and deep learning: A review of methods and applications." World Information Technology and Engineering Journal 10, no. 07 (2023): 3897-3904.

[19]. Salem, Aya H., Safaa M. Azzam, Omar E. Emam, and Amr A. Abohany. "Advancing cybersecurity: a comprehensive review of AI-driven detection techniques." Journal of Big Data 11, no. 1 (2024): 105.

[20]. Mohammed, Anwar. "Protecting space assets: Cybersecurity challenges and solutions for the final frontier." Baltic Journal of Engineering and Technology 2, no. 1 (2023): 55-61.

Copyright@ REST Publisher