2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)
University of Malaya, Kuala Lumpur, Malaysia. Sep 24-26, 2021

Machine Learning based Detection of Cyber Crime Hub Analysis using Twitter Data

Vinod Mahor
Computer Science
SVVV IPS College of Technology and Management
Gwalior,India
vinodengg.mt@gmail.com

Romil Rawat
Computer Science,SVIIT
SVVV
Indore,India
rawat.romil@gmail.com

Shrikant Telang
Information Technology ,SVIIT
SVVV
Indore,India
shreekanttelang@gmail.com

Bhagwati Garg
Union Bank of India,
Branch
Gwalior, India
gargpratap@gmail.com

Debajyoti Mukhopadhyay
Dept of Computer Science Engineering
Bennett University
Greater Noida, India
debajyoti.mukhopadhyay@gmail.com
ORCID ID:
0000-0002-8071-4091

Prajyot Palimkar
The Neotia University
School Of Engineering & Applied Sciences
Robotics Engineering
The Neotia University

***Abstract*—Online Social Network Platform is the hub of criminal activities .Among them twitter is the most popular one containing millions of media contents and thus can be used for crime analysis. In this research, a system for detecting Cyber Crime Hub in India, specifically Jamtara, is created. The design shows the criminal alert zones based on the name of crime incident reported. The proposed approach employs sigmoid kernel, support vector machine (SVM) classification. According to this analysis, Rate of Accuracy is 97.12 Percent (libSVM) , which is greater by 3.12 percent(mySVM) as evaluation results of the proposed system in comparison to Data collected from Government site for Cyber crime analytics the system shows the statistics of cyber crime cases.**

Keywords—Cyber Crime Hub; Jamtara; Text Classification; Support Vector Machine; Cyber Attack

I. INTRODUCTION

The rise in localized gangs, technological access, local support, training and motivation, Crime Hub [1-5] violations even fraud are some of the factors that cause Cyber Crime Hub formation. Technology has reached the group level in this age of modernization, and many people now post on social media. The Twitter platform that allows users to send textual and picture messages [6-8] openly accessed by users.

Between 2014(100 million) and 2021(412 million), the number of Twitter users worldwide increased. the total number of tweets sent in India in 2019 was 34 million and can be used as data derived from the volume of registered Twitter users in India multiplied by the volume of data collected. Using textual analysis to detect Cyber Crime Hub, the classification method can be used to organize data. Using SVM Linear method with the mySVM library having the best degree of precision in comparison to classification approaches (k-NN and Naive Bayes)[9-12].

The rest of the paper is organized as follows. Section II. shows about related work, section III. describes about research method, section IV. shows about result and finally section v. concludes the paper.

II. RELATED WORK

The The NLP[6] algorithm is used to extract and classify Crime Hub data, dividing into points and links. convergence are represented by points, and crossing intersections are represented by links. This approach correctly classifies 3168 Crime Hub tweets values into point division(Accuracy of 76.85 percent) and 467 tweets values into link division(Accuracy of 94.76 percent).

A framework based on incident-related[13-15] tweets, such as car fraud , and, number of re-tweets, hash-tag and URL connect is implemented. Tokenization[8] is used to clip and extract the term using the n-gram process, geo spot, text position of the expert 's direction, and tweet time. The information is then categorized using NLP Analysis, which has an accuracy rate of 80 percent.

A scheme for detecting injuries using Twitter was suggested by Research. This study employs a pre-processing, or prefiltering, technique by removal of Stop terms. Correcting misspellings, POS filtering, slang substitution Replacement of temporal and spatial mentions, followed by extraction using n-grams(Char and Word), TF-IDF[9],features(Syntactic, Spatial/Temporal, FeGeLOD), SVM, NB(Nave-Bayes), and the Ripper grouping. According to the results of the evaluation, the SVM classification system has the maximum accuracy rating of 91 percent.

Tokenization, halt , trailing, and feature depiction , and TF-IDF extraction are used in this analysis[10], and categorized using the SVM tool,k-NN, Naive Bayes. The aim of this grouping is to assign a class mark tweet that is correlated with Crime Hub events or not. This research resulted in a device that can be used for real-time tracking of many locations on Indian states, allowing for the detection of Crime Hub fraud using the SVM classification system having the best precision of the three classification methods, at 97.28 percent.

A survey conducted to monitor the regular Crime Hub in Indian States[11]. The result of this study is Crime Hub Watch, a device that can give valuable information to CHM (Crime Hub Management Center) in real time [16-19]. Crime

2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON) | 978-1-7281-9951-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/GUCON50781.2021.9573736

Hub Watch may be used to gather intelligence about possible hazards, fraud that could arise, the provision of public transportation and in this study, three classification approaches, namely k-NN, BN(Bayes-Network), and DT(Decision Tree), used for decision planning [20-22]. Stop expression, Lemma, POS Tagger, and other pre-processing techniques are used before the data is labeled, Tokenization and pattern recognition. IDF will be used to quantify the keyword performance. The Bayes Network, which has an accuracy rate of 98.9 percent, is the most accurate among three systems [23-26].

Study was conducted to determine the potential causes of Crime Hub irregularities by finding Textual context at social media linked to each fault in time and place. The dataset from Indian government Jharkhand Police[27] to test the method has been used for research. The system would look for discriminatory keywords that can be combined with others; if none are found, the system would move on to the next step and Keywords have been deleted. The anomaly sensor is then used to classify it.

A method to assist hackers[13] in avoiding Cyber Crime Hub, Incident, and hub maintenance by displaying statistics about the shortest techniques to the attacker's destination is defined. (Intelligent Transportation System) was the name given to the system by the researcher. To detect real-time Crime Hub in this analysis, data from Twitter was pre-processed as textual context details by Ontology-Based techniques for Detection and Tokenization, after which value is measured using IDF extraction and then categorized. According to the findings, the SVM approach having the highest accuracy(98.5 percent). It has 91.1 percent accuracy when looking at the state of Crime Hub delays, and 86.3 percent accuracy when looking at the causes of Crime Hub jams.

In the study, data from a facebook pages were used to assess customer satisfaction with an online money laundering agent in India. Three classification methods is used to analyze the data: SVM, k-NN, and Nave Bayes. When comparing SVM and Nave Bayes, the results reveal that k-NN has the best f-score [13].

III. Research Method

The first steps are to analyze the available data, which are performed on text data that has been categorized using ML (Machine Learning). The 2nd step to gather dataset for training ,taken from Twitter using the defined keywords "Asking_OTP_PIN ," "Lottery ," "account_blocked," and "immediately transfer__fee ,". In India particularly in criminal areas, The term "Asking_OTP_PIN " refers to a condition in which fraud person acts like a bank employee and ask for OTP and PIN, received at user mobile. The term "Lottery" refers to the circumstances in which a fraud person claims an agent about lottery winning amount and ask for processing fees. The term "account_blocked " refers to situations under which a fraud person informs about the credentials of bank account or card has been blocked and for reactivating it , immediate amount needs to be submitted into agent account.

The 3rd procedure is to erase redundant details and that could be done with Microsoft Excel, and any data that has the same text will be deleted, so that the content obtained would be new. The text data obtained would be automatically labeled using X-Means Clustering in the 4th stage. A computer programme Clustering is broken down into four sections clusters (0 , 1, 2, and 3). Since data may be clustered into several words of similar sentence in text content, clustering is separated into four sections. Each clustering contains the same information. It will be divided into two groups as a result of clustering: "Asking_OTP_PIN " and "Not_to_share_OTP_PIN ." The "Asking_OTP_PIN " class denotes Crime Hub conditions having high frequency of crime occurrence. The "Not_to_share_OTP_PIN " designation denotes Crime Hub conditions that exceed limited unauthenticated by fraudulent activities. Pre-processing data is the 5th stage, Tokenization generates tokens that are separated by a single character value. The first step is to perform Tokenization Space[28].

After removing tokens containing URL connections and '@' characters, non-letter tokenization would be done. Following that, any token with a length of less than two would be excluded. Lowercase can be applied to the token and containing the Indian languages listing of stop word will be deleted, And used to exclude vocabulary from the classification process that isn't needed. The final step in the pre-processing process is to produce an n-gram.

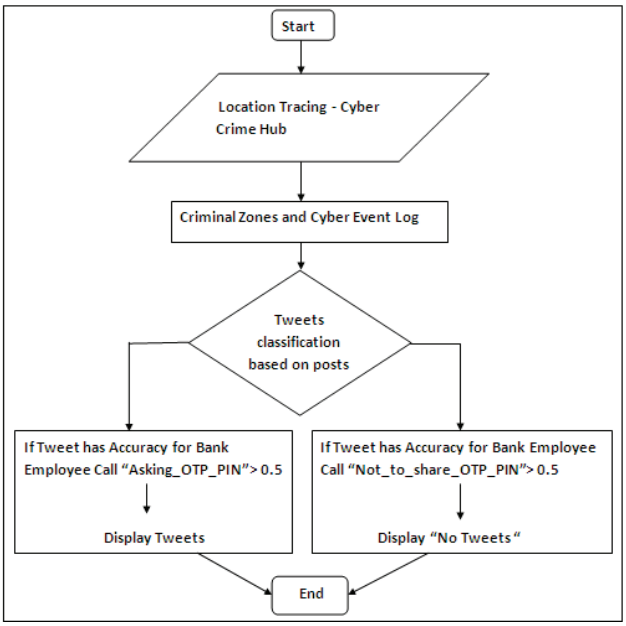

Fig. 1. Application Flowchart

The method of creating a words-listing taken from multiple tokens dependent on counting of n-grams is known as to generate n-gram. If there are two n-grams, two tokens are collected and will be combined to form a series of words. By the number of grammes, one word will be moved to the front. The tokens will be transformed into vector form in the 6th stage, and each word will be given a weight depending on the number of times it appears and the TF-IDF system weight words The 7th stage is to test the classification model's accuracy using the cross-validation method, which will be repeated up to ten times. The classification model can predict whether the Confusion Matrix is true or false based on several rows of test results. The 9th stage entails deciding the most accurate description possible. The results of the experiments will then be incorporated into the programme as the final stage. The flowchart given in Fig-1 for the programme shows the processing framework.

According to the programme architecture of Fig-1, there are original position inputs, input destination zones, and statistics in Figure 1. Within the classification frame, a tweet would be produced depending on the name of the suspicious location passed, based on the classification having degree of precision of the probability of Cyber Crime Hub exceeding 50% relative to the techniques passed. If there are no tweets that cause Cyber Crime Hub of more than 50%, and If no-tweets Happens related to Crime Hub , "No Message," will be shown and if there are, the system will produce the time that the tweet was written, the user's username, the accuracy outcome from the "Asking_OTP_PIN " class.

IV. RESULT

A total of 17834 tweets were successfully recovered between March 31, 2020 and March 8, 2021. Then, using Microsoft Excel, the text information similarity can be extracted from the results, The data is transformed into 9812 tweets. Since Rapidminer can processes only 5000 bytes of dataset, consisting the header, only 5127 tweets are used. Then, using the x-means clustering algorithm, the data would be numbered. Clustering can be bifurcated into four parts: Cluster (0) illustrates 1573 tweets of Cyber Crime Hub , Cluster(1) illustrates a very congested state of 2893 tweets, and Cluster(2) represents a very congested condition of 2987 tweets, Cluster 2 illustrates a 1987-tweet-long Crime Hub jam, while Cluster(3) illustrates an 93-tweet-long Crime Hub jam.

"Asking_OTP_PIN " and "Not_to_share_OTP_PIN " will be the two classes of clustering. Cluster 2 has a lot of criteria, like "Not_to_share_OTP_PIN " class types, while cluster 0 and cluster 1 having a lot of criteria that aren't crowded. Cluster 3 is labeled "Asking_OTP_PIN " since it represents congested Crime Hub situations. The class data would be equated "Not_to_share_OTP_PIN " class data in order to change it. Cluster 0 will receive 823 tweets, 1-cluster will receive 792 tweets, and 3-cluster will receive 91 tweets, bringing the total to 1987 tweets. The analysis process will be model-led using Rapid-miner Software from the 1st step to the 9th step in order to provide the findings.

A. N-Gram Classification Comparative Chart

TABLE I. CLASSIFICATION COMPARATIVE CHART (.N-GRAM)

Classification (SVM)	N Gram			
	n=3	n=2	n=1	Outcome (Average)
Linear	96,71 %	96,18 %	96,82 %	96,57
Polynomial	83,40 %	86,64 %	88,21 %	86.03
Sigmoid	96,23 %	96,32 %	96,23 %	95.26

According to Table 1, Linear (SVM) classifier with n=1 (96.82 percent) produces the best degree of precision, while Polynomial (SVM) classified with n=3 (83.40 percent) produces the lowest accuracy level.

B. Available Results Comparative Chart

TABLE II. AVAILABLE RESULTS COMPARATIVE CHART

Approaches	Classification (SVM)	N- Gram n=3	N- Gram n=2	N- Gram n=1	Outcome (Average)
	SVMdot (mySVM)	94,17 %	94,26 %	92,12 %	93.51
	k-NN	92,86 %	92,42 %	89,03 %	91.43
	NB	89,34 %	87,64 %	82,57 %	86.51
Proposed Technique (libSVM)	Linear	96,71 %	96,18 %	96,82 %	96,57
	Polynomial	83,40 %	86,64 %	88,21 %	86.03
	Sigmoid	96,23 %	96,32 %	96,23 %	95.26

According to Table 2, the classifying outcome when n-gram value taken (n = 1) that yield the best degree of precision are Linear(SVM) in the suggested system (96.82 percent), while NB(Nave-Bayes) in the previous study (82.57 percent) produces the lowest degree of precision. Linear (SVM) in the suggested system of 96.18 percent achieves the maximum degree of accuracy at n-gram (n = 2),In the proposed procedure, Polynomial(SVM) has the lowest accuracy level of 86.64 percent. The suggested method's Sigmoid(SVM) produces the maximum accuracy level of 96.23 percent by taking n-gram value (n = 3), while the current research's Polynomial(SVM) produces the lowest accuracy level of 83.40 percent. In accordance with the categorization process's aggregate value, the Sigmoid SVM in the proposed method produces the highest accuracy level with an average of 97.12 percent, while the Polynomial (SVM) in the suggested method produces the lowest accuracy level with an average of 86.03 percent. The proposed approach had the highest degree of accuracy, with an average of 97.12 percent, while the previous study has an average of 92.12 percent, taking k-NN values (k = 1) with aggregate of 93.51 percent, and NB (Nave Bayes) with an outcome of 86.51 percent. Figure 2 shows statistics of cyber crime at Jamtara.

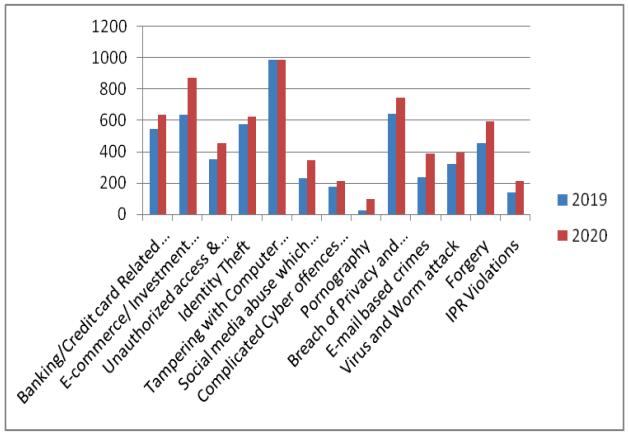

Fig. 2. Statistics of cyber crime at Jamtara.

C. Comparison Results

The system comparison of Cyber Crime Hub detection with Statistics is done to verify the accuracy. According to Table 3, the work was done According to the research findings, there are testing variations in the situations between the programme and Statistics.

TABLE III. COMPARISON RESULT

Crime Categories	Registered Cases		Reason	Locals Support	Education	Most Active International Gang
	2019	2020				
Banking Malware and crimes	>546	>635	Daily Earning of More than 3,000 INR ,local Business and Friend circle Support , Always works in Gangs ,Commission of Amount to all Agents , Practical Training without Education and Degree for Cyber Crime Conduction	Easy Availability of SIM Cards and Document Forging Printout	School and College Dropouts	South Asia Countries (Pakistan and Nigeria)
E-commerce/Business Frauds	>634	>874				
Unauthorized Access	>354	>453				
Unauthorized transactions	>576	>622				
Documents Tampering	>987	>989				
Social media abuse	>231	>344				
Smart phones Cyber offences	>174	>212				
Pornographic contents	>23	>97				
Computer related crimes Breach	>645	>745				
Phishing/Spamming	>234	>387				
Virus and Worm attack	>321	>392				
Forgery	>456	>593				
IPR Violations	>142	>213				
Total Registered Cases	> 6000	>11,000				

V. CONCLUSION

To create a model that can calculate the zone of criminal gang based on the findings of this analysis, multiple steps are implemented including (collecting data from twitter, clustering, pre-processing, and last the classification) with the highest accuracy score is 97.12 for sigmoid kernels. As opposed to the implementation of n-gram, the classifying method of Linear kernel having the best degree of precision, which is equal to 96.82 percent in n-gram (n = 1), and the classifying method Linear kernel has the best degree of precision, which is equal to 96.12 percent in n-gram (n = 2).In n-gram (n = 3), the classifying method with the Sigmoid kernel has the best degree of precision (96.26%), and in n-gram (n = 3) the SVM classification method with the Sigmoid kernel has the best degree of precision (96.71%). As opposed to other experiments using the sigmoid kernel to detect criminal gang on Twitter , this one comes out on top. The best accuracy scores on average are 97.12 percent. There are also many abbreviated words in text material for further study or production, the spell correction and modification techniques could be use in pre-processing steps to improve the accuracy of textual content classification, and can use photos and videos in posts to improve the accuracy of criminal gang detection, as well as incorporating social media details other than Twitter.

REFERENCES

[1] Rajput, B. (2020). Exploring the Phenomenon of Cyber Economic Crime. In Cyber Economic Crime in India (pp. 53-78). Springer, Cham.

[2] Banoo, S. (2020). Evaluating Personal Data Protection Bill, 2019: An Appraisal of Inception of India's Privacy Legislation. Supremo Amicus (ISSN NO. 2456-9704), 18.

[3] Banoo, S. (2016). INCEPTION OF INDIA'S PRIVACY LEGISLATION. Regulation (EU), 679, 679.

[4] Zulfikar, M. T. (2019). Detection traffic congestion based on Twitter data using machine learning. Procedia Computer Science, 157, 118-124.

[5] Jang-Jaccard, J., & Nepal, S. (2014). A survey of emerging threats in cybersecurity. Journal of Computer and System Sciences, 80(5), 973-993.

[6] Valluripally, S., Sukheja, D., Ohri, K., & Singh, S. K. (2019, May). IoT Based Smart Luggage Monitor Alarm System. In International Conference on Internet of Things and Connected Technologies (pp. 294-302). Springer, Cham.

[7] Zamojski, W., Mazurkiewicz, J., Sugier, J., Walkowiak, T., & Kacprzyk, J. (Eds.). (2019). Engineering in Dependability of Computer Systems and Networks: Proceedings of the Fourteenth International Conference on Dependability of Computer Systems DepCoS-RELCOMEX, July 1–5, 2019, Brunów, Poland (Vol. 987). Springer.

[8] Kakkar, A. (2020). A survey on secure communication techniques for 5G wireless heterogeneous networks. Information Fusion, 62, 89-109.

[9] Zhou, Z., Chen, X., Zhang, Y., & Mumtaz, S. (2020). Blockchain-empowered secure spectrum sharing for 5G heterogeneous networks. IEEE Network, 34(1), 24-31.

[10] Kadoguchi, M., Kobayashi, H., Hayashi, S., Otsuka, A., & Hashimoto, M. (2020, November). Deep Self-Supervised Clustering of the Dark Web for Cyber Threat Intelligence. In 2020 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 1-6). IEEE.

[11] Malhotra, P., Singh, Y., Anand, P., Bangotra, D. K., Singh, P. K., & Hong, W. C. (2021). Internet of Things: Evolution, Concerns and Security Challenges. Sensors, 21(5), 1809.

[12] Kaur, S., & Randhawa, S. (2020). Dark Web: A Web of Crimes. Wireless Personal Communications, 112(4), 2131-2158.

[13] Roddy, A. L., & Holt, T. J. (2020). An Assessment of Hitmen and Contracted Violence Providers Operating Online. Deviant Behavior, 1-13.

[14] Martin, J., Munksgaard, R., Coomber, R., Demant, J., & Barratt, M. J. (2020). Selling drugs on darkweb cryptomarkets: differentiated pathways, risks and rewards. The British Journal of Criminology, 60(3), 559-578.

[15] Sinha T., Chowdhury T., Shaw R.N., Ghosh A. (2022) Analysis and Prediction of COVID-19 Confirmed Cases Using Deep Learning Models: A Comparative Study. In: Bianchini M., Piuri V., Das S., Shaw R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, vol 218. Springer, Singapore. https://doi.org/10.1007/978-981-16-2164-2_18

[16] Palimkar P., Shaw R.N., Ghosh A. (2022) Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach. In: Bianchini M., Piuri V., Das S., Shaw R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, vol 218. Springer, Singapore. https://doi.org/10.1007/978-981-16-2164-2_19

[17] Chakraborty A., Chatterjee S., Majumder K., Shaw R.N., Ghosh A. (2022) A Comparative Study of Myocardial Infarction Detection from ECG Data Using Machine Learning. In: Bianchini M., Piuri V., Das S., Shaw R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, vol 218. Springer, Singapore. https://doi.org/10.1007/978-981-16-2164-2_21

[18] Rawat R., Mahor V., Chirgaiya S., Shaw R.N., Ghosh A. (2021) Analysis of Darknet Traffic for Criminal Activities Detection Using TF-IDF and Light Gradient Boosted Machine Learning Algorithm. In: Mekhilef S., Favorskaya M., Pandey R.K., Shaw R.N. (eds) Innovations in Electrical and Electronic Engineering. Lecture Notes in Electrical Engineering, vol 756. Springer, Singapore. https://doi.org/10.1007/978-981-16-0749-3_53

[19] Rajawat A.S., Rawat R., Mahor V., Shaw R.N., Ghosh A. (2021) Suspicious Big Text Data Analysis for Prediction—On Darkweb User Activity Using Computational Intelligence Model. In: Mekhilef S., Favorskaya M., Pandey R.K., Shaw R.N. (eds) Innovations in Electrical and Electronic Engineering. Lecture Notes in Electrical Engineering, vol 756. Springer, Singapore. https://doi.org/10.1007/978-981-16-0749-3_58

[20] Rajawat A.S., Rawat R., Barhanpurkar K., Shaw R.N., Ghosh A. (2021) Vulnerability Analysis at Industrial Internet of Things Platform on Dark Web Network Using Computational Intelligence. In: Bansal J.C., Paprzycki M., Bianchini M., Das S. (eds) Computationally Intelligent Systems and their Applications. Studies in Computational Intelligence, vol 950. Springer, Singapore. https://doi.org/10.1007/978-981-16-0407-2_4

[21] Rajawat A.S., Rawat R., Barhanpurkar K., Shaw R.N., Ghosh A. (2021) Sleep Apnea Detection Using Contact-Based and Non-Contact-Based Using Deep Learning Methods. In: Bansal J.C., Paprzycki M., Bianchini M., Das S. (eds) Computationally Intelligent Systems and their Applications. Studies in Computational Intelligence, vol 950. Springer, Singapore. https://doi.org/10.1007/978-981-16-0407-2_7

[22] Rawat R., Mahor V., Chirgaiya S., Shaw R.N., Ghosh A. (2021) Sentiment Analysis at Online Social Network for Cyber-Malicious Post Reviews Using Machine Learning Techniques. In: Bansal J.C., Paprzycki M., Bianchini M., Das S. (eds) Computationally Intelligent Systems and their Applications. Studies in Computational Intelligence, vol 950. Springer, Singapore. https://doi.org/10.1007/978-981-16-0407-2_9

[23] Rajawat A.S., Rawat R., Shaw R.N., Ghosh A. (2021) Cyber Physical System Fraud Analysis by Mobile Robot. In: Bianchini M., Simic M., Ghosh A., Shaw R.N. (eds) Machine Learning for Robotics Applications. Studies in Computational Intelligence, vol 960. Springer, Singapore. https://doi.org/10.1007/978-981-16-0598-7_4

[24] Kumar M., Shenbagaraman V.M., Shaw R.N., Ghosh A. (2021) Digital Transformation in Smart Manufacturing with Industrial Robot Through Predictive Data Analysis. In: Bianchini M., Simic M., Ghosh A., Shaw R.N. (eds) Machine Learning for Robotics Applications. Studies in Computational Intelligence, vol 960. Springer, Singapore. https://doi.org/10.1007/978-981-16-0598-7_8

[25] Rawat R., Rajawat A.S., Mahor V., Shaw R.N., Ghosh A. (2021) Surveillance Robot in Cyber Intelligence for Vulnerability Detection. In: Bianchini M., Simic M., Ghosh A., Shaw R.N. (eds) Machine Learning for Robotics Applications. Studies in Computational Intelligence, vol 960. Springer, Singapore. https://doi.org/10.1007/978-981-16-0598-7_9

[26] Malsa N., Vyas V., Gautam J., Shaw R.N., Ghosh A. (2021) Framework and Smart Contract for Blockchain Enabled Certificate Verification System Using Robotics. In: Bianchini M., Simic M., Ghosh A., Shaw R.N. (eds) Machine Learning for Robotics Applications. Studies in Computational Intelligence, vol 960. Springer, Singapore. https://doi.org/10.1007/978-981-16-0598-7_10

[27] https://www.jhpolice.gov.in/cyber-crime-ps

[28] Choi, D., Ko, B., Kim, H., & Kim, P. (2014). Text analysis for detecting terrorism-related articles on the web. Journal of Network and Computer Applications, 38, 16-21.