Prediction of Cybercrime Rate by Using Data Mining Techniques

AUTHOR

Abstract

Over the past decade, the use of the internet has grown to an astonishing extent. It has dominated almost every field of life and it acts as a channel of communication between various walks of life. However, with an amplified penetration of the internet, cybercrimes have also gained significant impetus. In order to implement their illegal activities and achieve illegal objectives, cybercriminals employ online networking. They use online networking devices to connect with any device of another online user and gain illegal profits in terms of finance, publicity or any other personal motive. The main reason behind such blatant violation of law by these cybercriminals is due to the loopholes and vulnerabilities present in the online system. They conveniently exploit these weak links and breach the privacy of online users. Despite efforts to control cybercrimes, they are increasing with every day passing. All manual methods, as well as existing technical approaches, have been ineffective in reducing cybercrimes to a satisfactory level. Therefore, this article presents an improved model for the implementation of data mining techniques for the prediction of cybercrimes. The model utilizes association rule and decision tree classifiers with Naïve Bayes algorithm and J48 algorithm and the collective knowledge gained will be invaluable in enhancing the accuracy of cybercrime prediction. The law imposing agencies should utilize such a model to fight the threat of cybercrimes.

Date of Submission: 01-08-2021 Date of Acceptance: 15-08-2021

I. Introduction

Cybercrime is the use of a computer as an instrument to perpetrate illegal aims. Currently, it is used with many different terminologies, such as e-crime, computer crime, etc. (Singh, Prasad, Narkhede, & Mehta, 2016). With an ever-increasing scope of global digitalization, cybercrimes have gained impetus and have created a sense of menace across the globe. Cybercriminals have increased their criminal activities and are targeting both private and public organisations (Prasanthi & Ishwarya, 2015). Either national or global, cybercrime activities place a negative influence on society in the form of emotional, financial and ethical dejection. What makes cybercrime so daunting and hazardous is the explicit breach of the individual as well as organizational privacy, resulting in damage to important files, datasets, software or web pages (Kumar, Koley, & Kuamr, 2015).

A number of malicious websites are set up by suspicious individuals, and they have exploited the loopholes present in cyberspace to an ominous extent.

Cybercriminals have utilized online life to their advantage to perpetuate a wide range of cybercrimes, such as phishing, spamming, malware articulation and cyberbullying. Despite the actions taken by law enforcement agencies to eradicate such malicious websites, the operators of these websites could not be stopped as they switched to other alternatives to achieve their illegal objectives (Alami & Elbeqqali, 2015). Further, cybercriminals have been able to continue their illegal activities due to certain security vulnerabilities in the existing system (Numan et al., 2020).

Moreover, social media platforms have enabled everyone to publish or access any information as per one's desire. It has revolutionized the way the world used to interact previously and has helped a great deal in transforming the world into a global village in its true sense. Moreover, online web-based life (e.g., Facebook and Twitter) has become one of the most fundamental parts of today's generation. However, it has also made its online users vulnerable to certain cybercrimes; it has served as an advantageous tool for cybercriminals to achieve their illegal and malicious purposes (Alami & Elbeqqali, 2015).

The rate of cybercrimes has been growing without any decline. An important thing to mention here is the lack of any airtight systematic and reliable source to review cybercrimes. This is because of the lack of record maintenance due to online users' reluctance to report such crimes over reservations regarding police response. Moreover, lack of awareness among internet users regarding information technology acts is another reason for the encouragement of cybercriminals. It would not be wrong to claim that raising awareness among people would suffice in reducing cybercrimes to a substantial level (Ch, Gadekallu, Abidi, & Al-Ahmari, 2020).

Since the use of computers is increasing in almost every field of life, the complexity of the cybercrimes that it will entail would be immense in both magnitude and dimension. Moreover, with an enormous amount of network data and online transaction traffic, perceiving the risks of cybercrimes would be more than a challenge. There is an urgent need to develop advanced methods and techniques that will assist in predicting cybercrime patterns in such a way that will create a path to controlling cybercrimes. Cybercrime prediction is focused on predicting the location and timing of future cybercrimes accurately by analyzing past and future data (Lekha & Prakasam, 2017).

One such technique for predicting cybercrimes is data mining. Data mining entails analyzing a large amount of data by separating the valuable examples from the data and then drawing meaning from it. It examines the available data from multiple perspectives and then provides inferences to extract invaluable information (Gangavane & Nikose, 2015). Data mining techniques can play a crucial role in overcoming the intricacies involved in cybercrimes by enhancing the efficiency and pace of cybercrime prediction.

Therefore, this paper aims at proposing a model useful in combating cybercrimes. The proposed approach is based on the use of various data mining techniques which will help in uncovering suspicious behaviors of cybercriminals. The proposed model is established to gain excess to the huge amount of cybercrime data and then make accurate predictions in accordance with the previous records of cybercriminals. However, in the present era where digitalization is at its peak, protecting such a huge pile of data from cyber attacks would also be an enormous challenge.

II. Literature Review

Ch et al. (2020) had intended to propose a useful data mining tool by making use of the machine learning technique. They argued that the proposed tool would enable them to evaluate the cybercrime rates at the state level by classifying the cybercrimes into various types. They employed a security analytics approach in combination with a data analytics approach to execute the classification of cybercrimes from the available data. Their findings revealed that the proposed model would provide accurate analysis along with fine reports. The reports showed the classification of cybercrimes with almost 99.99% accuracy.

Toapanta, Gallegos, Andrade, and Espinoza (2020) had described a model to predict cyber offenses by using data mining techniques. The main technique they utilized was the decision tree technique which involved two stages, viz. learning stage and prediction stage. In order to execute the decision tree technique, the required repository was obtained. Their findings revealed that in the first stage, the dimension 'population per community' had the highest proportion of cybercrime among all other repository dimensions. Then, they applied data mining techniques to preclude cyber offenses. Their model resulted in an explicit decrease in the risks of cyber attacks among the communities subjected to the prediction process.

Lekha and Prakasam (2018) had discussed the utilization of data mining techniques for the detection of cyber fraudulent activities. They argued that with an ever-increasing trend of online transactions and processes, cybercriminals have also increased the pace of their illegal activities and iterated the need to widen the scope of investigation regarding the counterfeit activities carried out by the cybercriminals. Moreover, they regarded it as considerably vital for law enforcement institutions to combat such fraudulent individuals. They assigned weights to the use of various data mining techniques to detect cybercrimes and prevent financial losses in real-time. They claimed that these data mining techniques can be used for cybercrime detection in a wide range of fields including, e-commerce, health insurance, among many others.

Michael (2020) has discussed the implementation of data mining techniques to predict cybercrime patterns under a knowledge-based system. The proposed system intended to pinpoint the certain risk factors and situations that were causing the augment in cybercrimes. It would help in devising a strong defensive cyber offense control strategy. They argued the need to integrate the use of a novel approach to tackle the menace of cybercrimes. To this end, they proposed the extensive use of data mining techniques as the novel approach to visualize the cyber offense data and predict its future occurrences. In data mining techniques, they assigned weights to the use of algorithms to extract logical facts out of the available dataset. Then, the extracted information will be used to analyze and visualize probable results regarding cybercrimes.

Ganesan and Mayilvahanan (2017) had analyzed the implementation of data mining techniques to predict the cybercrimes in social media. They discussed the nature and types of cybercrime that have caused a huge sense of insecurity among internet users. They argued that the increasing use of social media websites has created room for cybercriminals to execute their malicious activities and achieve their illegal objectives. They used the clustering technique i.e., k-means clustering algorithm to extract information about cybercriminals and hackers. They concluded with reiteration on the invention of new data mining techniques and methods to enhance the effectiveness of cybercrime investigation.

Prabakaran and Mitra (2018) had executed a paper survey of various data mining techniques used in cybercrime detection.

They argued that data mining techniques such as machine learning can help analyze a large set of intricate crime data and extract valuable information to predict the pattern of future cybercrimes. It would enable the reduction of risk in various fields and be implemented extensively in healthcare systems, education, manufacturing, etc. They iterated the importance of data mining techniques in bringing positive change in society as it reduces the insecurities among users through crime investigation. They discussed some of the important data mining techniques, viz. genetic algorithm, Hidden Markov Model (HMM), naive Bayesian and k-mean clustering.

Jain et al. (2016) had a model for controlling cyber crimes using various data mining techniques. They intended to invent such an approach that would enable the investigation of both digital as well as physical crimes at the same time. The proposed model was based around the integration of various computer based forensic tools in their approach which would enhance the effectiveness of cybercrime analysis and control. First, the model would report the crime followed by the adoption of certain data mining techniques to detect the cyber offense launched on a user's computer. They used the k-mean algorithm and Naïve Bayes algorithm in their model. They concluded that a cybercrime detection model based on forensic tools would be valuable for law enforcement agencies to predict future cybercrimes, leading to their control and reduction.

Hosseinkhani, Koochakzaei, Keikhaee, and Naniz (2014) had discussed the use of various data mining techniques in order to identify suspicious activities and crime hot spots and predict future crime patterns. They argued that data mining had gained significant importance in the field of criminology due to its landmark success in subduing criminals. They described that the data mining approach used to detect the future crime pattern in criminology is based around the principle of the relationship between criminal activities and the criminal. They iterated the need to develop such further crime analysis techniques based on data mining that would evaluate the highly complex crime dataset, assisting law enforcement agencies in their proper functioning.

Priya and Meenakshi (2017) had conducted a study to evaluate the utilization of data mining in detecting cybercrimes. To this end, they used C4.5 (J48) data mining technique through the help of the WEKA tool. The dataset for the study consisted of 750 uniform resource locators (URLs). They used this dataset to plot the J48 algorithm which is used for the application of the C4.5 algorithm in WEKA. Among the available 700 URLs, they chose 300 URLs in order to predict the cyber offenses with the assistance of a classifier developed by training the J48 algorithm. Their findings clearly showed that the C4.5 data mining technique used for the data analysis had high accuracy (82.6%).

Chandrakala, Rajini, Dharmarajan, and Selvam (2020) discussed the application of data mining techniques in the detection and prediction of cybercrimes. The main data mining techniques discussed by the authors were social network analysis, naïve Bayes rule, decision tree method, and clustering analysis. They also greatly focused on Denial of Service (DoS) assaults by cybercriminals. After their analysis by using the J48 data mining algorithm, they also proposed a continuous acknowledgment model based on its association with security gadgets, resulting in increased crime detection accuracy.

Kaur, Vashisht, and Saurabh (2012) had proposed a model for the development of an accurate and efficient cybercrime detection system. The basic principle was to construct such a detection system that can adapt to the fluctuating behavior in combination with various data mining techniques. Their proposed system worked in two phases: the first one included the in-depth analysis of the collected user information while the second phase was involved in the detection of cybercrime in the form of an alarm. They used the k-mean clustering algorithm in their model to obtain the desired clusters.

Kiani, Mahdavi, and Keshavarzi (2015) conducted a study to execute the classification of clustered crimes on the basis of their occurrences. They applied a theoretical model by using various data mining techniques including k-means clustering. The model was based on the real crime dataset recorded by police in England and Wales during 1990-2011. They gave special attention to the features to enhance the overall effectiveness of the proposed model. Moreover, they employed a genetic algorithm (GA) for the optimization of Outlier Detection operator parameters. To this end, they used the RapidMiner tool.

Chitra and Subashini (2013) had analyzed the implementation of various data mining techniques in detecting cybercrimes in the banking sector such as fraud detection, risk management, crime prevention. They emphasized the need to develop more data mining techniques to enhance service quality and customer satisfaction and retention by detecting fraud and illegal activities on a real-time basis.

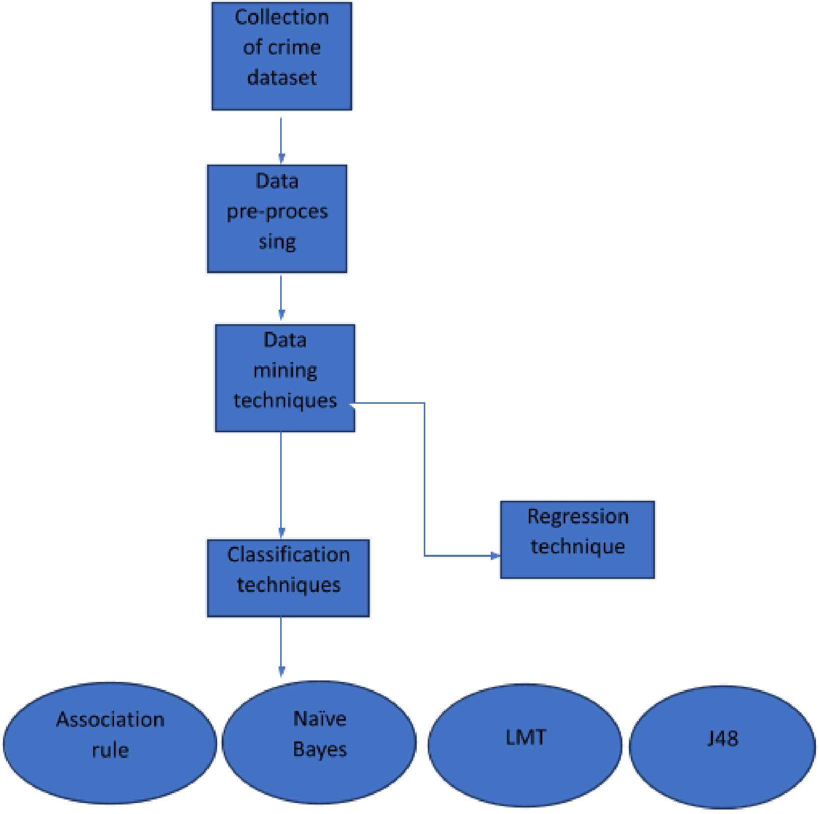
Tianfield (2017) had discussed the utilization of data mining techniques in preventing cyber crimes. The author proposed a defense framework based on data extracted regarding cyber crimes in order to enhance situational awareness regarding cybersecurity. Moreover, a multi-loop architecture was also narrated to detect cybercrimes on the basis of various data mining techniques. The author also discussed a number of data mining techniques used in crime prediction in detail.

III. Theoretical Framework

Proposed model for cybercrime prediction

Discovering and exploring cybercrimes and underlining their affiliations with cybercriminals are utilized in evaluating cybercrime penetration in society. The proposed model for cybercrime prediction involves the use of the WEKA data mining tool for various data mining techniques. For the crime prediction, various classifiers were utilized including, viz. Bayes Net, Naïve Bayes, Simple Logistic Regression, JRip, Decision Tree, J48 and LMT. After all these classifiers, the obtained results from all classifiers are compared to gauge the best performing algorithm.

Fig 1: Proposed model for cybercrime prediction



A. *Collection of cybercrime dataset*

Data has to be obtained from Kaggle which is a repository of the dataset on various domains. The dataset will comprise several cybercrime-related features, such as location, crime year, crime head, etc. Different types of cybercrimes (fraud, extortion, cyberbullying, revenge, etc.) and crime prediction will serve as the primary components of the dataset.

B. *Pre-processing of data*

The data pre-processing was performed with an objective to extract valuable information to transform the raw data into an effective format. Data pre-processing will involve the following key steps:

- Data integration: Irrelevant parts of the data are removed and missing data and values are integrated into the dataset.
- Data transformation: Data is normalized to measure the information values in a given range, i.e., -1.0 – 1.0 or 0.0 – 1.0. Then, appropriate attributes are selected to further the mining process.
- Data reduction: Data mining is done to handle a huge amount of data. However, if the data is larger than a certain limit, the analysis would become a hard challenge. Therefore, the data is reduced to assist the mining process. Data reduction can be carried out through attribute subset selection, numerosity reduction and dimensionality reduction.
- Data discretization: It is done to replace the raw numeric attribute values with interval or concept level values.

C. *Data mining techniques*

Data mining is the process of analyzing data in such a way that enables the identification of reliable patterns or interrelationships between variables by using various data mining approaches. Data mining generally aims to gather and transform helpful information for future use (Oladokun, Adebanjo, & Charles-Owaba, 2008). In the proposed model, the following data mining techniques are used:

1. Data mining classification technique

This is the process of classifying different items in the dataset into various groups. The classification model is developed as per the dataset available for training. It is generally the most used data mining technique. Classification helps in identifying in which group each data occurrence is linked (Lekha & Prakasam, 2017).

2. Data mining regression technique

Regression is a data mining technique used in a certain type of dataset to predict the variety of numeric values.

D. Prediction

Prediction of the dataset includes searching for the hidden patterns or existing knowledge of the historical data available.

E. Classifiers

The following classifiers are described in the proposed model:

1. Association rule classifier

Association rule classifiers are the if-then statements that help to show the probability of relationships between data items within large datasets. This technique produces rules for cybercrime datasets on the basis of frequent occurrences of cybercrime patterns. It involves the following subsequent measures:

- Identification of commonly occurring items in the cybercrime datasets
- The detection of patterns in program implementation and customer behaviors as association rules

2. Decision tree classifier

The decision builds classification of the cybercrime data in the form of a tree structure. It divides the cybercrime dataset into smaller subsets and an associated decision tree is established. Consequently, a tree with decision nodes and leaf nodes is developed.

F. Algorithms

The following algorithms are used in the proposed model:

1. Naïve Bayes algorithm

The naïve Bayes algorithm is based on the Bayes proposal which holds that each pair of features is independent. It is quite helpful in detecting cybercrime activities. In most real time situations, such as spam filtering this technique works quite well (Bahnsen, Aouada, Stojanovic, & Ottersten, 2016).

2. J48 algorithm

The J48 algorithm is an extended version of the C4.5 decision tree. For the proper classification of problems in cybercrime, the J48 algorithm is quite helpful as it is more precise and spikier. There are two phases in the J48 algorithm:

- Development of tree
- Validation of the build tree over available cybercrime dataset

The J48 algorithm operates on the basis of the pruning method for the development of the tree. The J48 algorithm generates the classifier output in the form of rules set and decision tree. Moreover, the rules are quite convenient and easy to recognize and employ.

3. Logistic model tree (LMT)

It is a classification model that has an associated supervised algorithm that combines logistic regression and decision tree learning. LMT uses a decision tree as leaves. It does not require any tuning parameters and is often regarded as more accurate than the C4.4 decision tree and standalone logistic regression. LMT produces a single tree containing binary splits on numeric attributes, multiway splits on nominal ones and logistic regression at the leaves. Moreover, it also ensures that only relevant attributes are included in the latter (Kravvaris, Kermanidis, & Thanou, 2012).

IV. Recommendations

Indeed, data mining techniques have improved a lot over the years and have helped a great deal in controlling and preventing cybercrime cases. However, there are still a number of loopholes that need to be amended through extensive research and strategic models based on authentic crime datasets. Following recommendations should be taken into account to improve cybercrime detection through data mining techniques.

- There should be an increased effort to devise a hybrid crime detection model. That is to say, there is a need to combine the prediction of cybercrime with the prediction of real-world offenses and crimes which will turn out to be a more accurate source of cybercrime prediction.
- Such hybrid models in combination with DP, NN and spatiotemporal correlations can significantly enhance the accuracy of crime prediction. Moreover, there should be an increased inclination towards the incorporation of more sophisticated data mining techniques in such models.

- Cybercrime prediction can be also improved by using threat management and warning correlation analytics by classifying subsets of alerts. It would enable it to employ only specific analyses required for the cybercrime prediction from the available data.
- Moreover, a close assessment of time series can also serve as a crucial tool in evaluating crime datasets as the reported crimes are registered along with the time. Furthermore, the efficiency of crime prediction models would be greatly amplified if better insights into the crime are gained in terms of sector and area, and network impact.
- There is an urgent need to combat the accessibility problem of the cybercrime dataset. This accessibility problem is basically due to the diversity of cybercrimes on multiple sites, such as Facebook, YouTube, Instagram, etc. Cyber Criminal profiling is the most feasible solution in this regard. However, it would need considerable cooperation between researchers, law enforcement institutions and government as the data used in cybercriminal profiling is often of confidential and personal nature, creating many ethical hurdles in the process.
- Cybercriminal profiling can be made viable if researchers find a way to protect the data provided by law enforcement agencies, while at the same time utilizing the data for cybercrime prediction.

V. Future Scope

Presently, the increase in cybercrime rates is creating an alarming situation for users as well as law enforcement agencies. With a dramatic augmentation in the use of the internet, suspicious people have made their way towards the commitment of crime using online life. Data mining technique has been regarded as a tool to deal with this daunting situation as it will help in the construction of such models that would enable accurate analytics regarding crime identification, detection, classification and eventually crime prevention.

Moreover, the extensive use of advanced data mining techniques such as machine learning in the future will assist in accurately finding cyber offenses that intrude into society due to certain loopholes in the security systems. Moreover, data mining techniques will help implement strict cyber offense regulations such as incarnation that would create a sense of fear among the cybercriminals preventing further spread of cybercrime. Last but not least, the data mining models will effectively reduce the time consumed in manual reporting of cybercrime, increasing the efficiency of crime prevention campaigns. It will enable police to identify the hot-spot locations vulnerable to frequent cyberattacks which will be invaluable in halting cybercrimes before their occurrence.

VI. Conclusion

The model proposed in this study produces a vital concept regarding cybercrime prediction by employing various novel data mining techniques, such as association rule and decision tree with LMT and J48 algorithms. This paper presents the improvement and extension of already proposed solutions. The proposed solution is based on the use of data mining techniques to tackle cybercrime offenses. The paper mainly focuses on enhancing the efficiency of the cybercrime detection systems in terms of execution time and improving the precision in prediction by using multiple knowledge resources. Moreover, advanced intrusion tools should be developed and employed in security systems on a standard basis. In order to eradicate cyber crimes and offenses at a rapid pace, users should be trained to deal with these cyber menaces which can be achieved by cooperation between government and law imposition organizations. Last but not least, there should be increased promotion of awareness of cyber crimes in order to make customers realize the vitality of this issue.

References

[1]. Alami, S., & Elbeqqali, O. (2015). Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts. Paper presented at the 2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA).

[2]. Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134-142.

[3]. Ch, R., Gadekallu, T. R., Abidi, M. H., & Al-Ahmari, A. (2020). Computational system to classify cyber crime offenses using machine learning. Sustainability, 12(10), 4087.

[4]. Chandrakala, T., Rajini, S. N. S., Dharmarajan, K., & Selvam, K. (2020). Development of crime and fraud prediction using data mining approaches. Technology, 11(12), 1450-1470.

[5]. Chitra, K., & Subashini, B. (2013). Data mining techniques and its applications in banking sector. International Journal of Emerging Technology and Advanced Engineering, 3(8), 219-226.

[6]. Ganesan, M., & Mayilvahanan, P. (2017). Cyber Crime Analysis in Social Media Using Data Mining Technique. International journal of pure and applied mathematics, 116(22), 413-424.

[7]. Gangavane, H., & Nikose, M. (2015). A Survey on Document Clutering for identifying Criminal. IJRITCC, 2(2), 459-463.

[8]. Hosseinkhani, J., Koochakzaei, M., Keikhaee, S., & Naniz, J. H. (2014). Detecting suspicion information on the Web using crime data mining techniques. International Journal of Advanced Computer Science and Information Technology, 3(1), 32-41.

[9]. Jain, N., Sharma, P., Anchan, R., Bhosale, A., Anchan, P., & Kalbande, D. (2016). Computerized forensic approach using data mining techniques. Paper presented at the Proceedings of the ACM Symposium on Women in Research 2016.

[10]. Kaur, M., Vashisht, S., & Saurabh, K. (2012). Adaptive algorithm for cyber crime detection. International Journal of Computer Science and Information Technologies (IJCSIT), 3(3), 4381-4384.

[11]. Kiani, R., Mahdavi, S., & Keshavarzi, A. (2015). Analysis and prediction of crimes by clustering and classification. International Journal of Advanced Research in Artificial Intelligence, 4(8), 11-17.

[12]. Kravvaris, D., Kermanidis, K. L., & Thanou, E. (2012). Success is hidden in the students' data. Paper presented at the IFIP International Conference on Artificial Intelligence Applications and Innovations.

[13]. Kumar, S., Koley, S., & Kuamr, U. (2015). Present Scenrio of Cyber Crime in INDIA and its Preventions. IJSER, 6(4), 1972-1976.

[14]. Lekha, K. C., & Prakasam, S. (2017). Data mining techniques in detecting and predicting cyber crimes in banking sector. Paper presented at the 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS).

[15]. Lekha, K. C., & Prakasam, S. (2018). Implementation of data mining techniques for cyber crime detection. International Journal of Engineering, Science and Mathematics, 7(4), 607-613.

[16]. Michael, G. (2020). Knowledge based system for predicting cyber crime patterns using data mining techniques. Journal of Critical Reviews, 7(10), 2043-2053.

[17]. Numan, M., Subhan, F., Khan, W. Z., Hakak, S., Haider, S., Reddy, G. T., Alazab, M. (2020). A systematic review on clone node detection in static wireless sensor networks. IEEE Access, 8, 65450-65461.

[18]. Oladokun, V., Adebanjo, A., & Charles-Owaba, O. (2008). Predicting students academic performance using artificial neural network: A case study of an engineering course.

[19]. Prabakaran, S., & Mitra, S. (2018). Survey of analysis of crime detection techniques using data mining and machine learning. Paper presented at the Journal of Physics: Conference Series.

[20]. Prasanthi, M. L., & Ishwarya, T. (2015). Cyber Crime Prevention & Detection. International journal of advanced research in computer and communication engineering, 4(3).

[21]. Priya, A., & Meenakshi, E. (2017). Detection of phishing websites using C4. 5 data mining algorithm. Paper presented at the 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT).

[22]. Singh, A. K., Prasad, N., Narkhede, N., & Mehta, S. (2016). Crime: Classification and pattern prediction. International Advanced Research Journal in Science, Engineering and Technology, 3(2), 41-43.

[23]. Tianfield, H. (2017). Data mining based cyber-attack detection. System simulation technology, 13(2), 90-104.

[24]. Toapanta, S. M. T., Gallegos, L. E. M., Andrade, B. E. C., & Espinoza, M. G. T. (2020). Analysis to predict cybercrime using information technology in a globalized environment. Paper presented at the 2020 3rd International Conference on Information and Computer Technologies (ICICT).

XXXXXXXX, et. al. "Prediction of Cybercrime Rate by Using Data Mining Techniques." *IOSR Journal of Computer Engineering (IOSR-JCE),* 23(4), 2021, pp. 01-07.