International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

CYBER CRIME ESTIMATION AND IDENTIFICATION: A HYBRID KEY INDICATOR USING MACHINE LEARNING AND DATA ANALYTICS

Mrs. Peddaboina Yamuna[*1], Saideep Yamsani[*2], Tejaswini Kachagandi[*3],Vishal Chokkarapu.[*4], Sagarika Posina[*5]

[*1] Assistant Professor, Department Of CSE, ACE Engineering College, Hyderabad, Telangana, India.

[*2,3,4,5]Student, Department Of CSE, ACE Engineering College, Hyderabad, Telangana, India.

ABSTRACT

Social media text analytics involves extracting information from text sources, including social media, through text analysis. Social media plays a significant role in addressing the problem of crime, which is a major issue in society today. The occurrence of crime negatively impacts both quality of life and economic growth. By analyzing historical data, we can identify crime patterns and make predictions about future incidents. However, there are certain unreported and unsolved crimes due to insufficient evidence, making the task of locating criminals challenging. One way to address this is by monitoring social media for criminal activity, as users sometimes share information about their surroundings on these platforms. In this research paper, we propose a Machine Learning approach to detect crimes, analyze their types, and investigate/trace the evidence of the attacker. Initially, we retrieve text messages that contain predefined keywords related to crimes. Then, we pre-process the data and employ a Support Vector Machine-based filtering approach to eliminate noise. Random Forest is subsequently used for classification. Finally, in the last stage, we analyze and categorize the crime types. By using the categorized types, we identify the evidence associated with the attacker.

Keywords: Cyber Crime, Cyber Attacks, Security, SVM, Malware, Data Analytics, Machine Learning.

I. INTRODUCTION

In today's modern world, we frequently encounter various crimes such as spamming, computer virus distribution, harassment, and cyber-stalking. While these crimes may not involve financial loss, they can be equally detrimental by causing data loss, information breaches, and unauthorized access to our computers. Therefore, it is crucial to prioritize cyber security. Cyber security refers to the process of safeguarding information stored on computers, electronic devices, and other hardware and software against unauthorized access, disclosure, modification, interruption, or destruction. It also helps protect computer systems from potential threats such as viruses, worms, bugs, and other harmful elements. Additionally, it plays a role in network monitoring and defending against various security threats. To protect our data from theft and other security issues, it is essential to employ computer security solutions to some extent. Crime is a social problem that negatively impacts various aspects of our society. It is vital for both local authorities and individuals to be able to identify high-crime areas and stay informed about the most recent crimes in specific regions. In busy environments, people are particularly concerned about improving safety and building trust among neighbors. Crime prevalence is a significant issue worldwide, especially in urban areas. While social crimes have received considerable attention, social media has only been minimally utilized in studies related to crimes and criminal behavior. Therefore, this study aims to employ the Random Forest algorithm, a machine learning technique, to effectively detect crimes based on features extracted from social media datasets using data mining concepts. Social media serves as the primary data source, and the main objective is to identify all concealed data sources and make predictions about outcomes. Furthermore, the study investigates the identity and evidence of intruders to assist digital forensic investigators in their work.

A. Statement of the Problem

There exists a range of criminal activities, encompassing offenses such as robbery, murder, rape, assault, battery, false imprisonment, kidnapping, and homicide. With the escalation of crime rates, there is an increasing urgency to expedite case resolutions. The prevalence of crime has been on the rise, emphasizing the crucial role of law enforcement agencies in managing and mitigating criminal activity.

B. The Aim and Objectives of the Study

The objective of this project is to forecast criminal activities and examine their origins by utilizing the features present in the dataset. The dataset itself is sourced from official websites. By employing Machine Learning algorithms and Python as the primary tools, we aim to predict the specific type of crime that is likely to transpire in a given area. Additionally, we intend to trace and uncover substantial details and evidential information pertaining to the perpetrator involved in the crime.

The following are the objectives of the study:

- The primary goal of the project is to forecast and analyse the crime rate that will occur in the future. Based on this information, officials can take charge and attempt to reduce crime.
- To uniquely identify each cyber user using their Internet Protocol (IP) address and location data, and then to store data pertaining to their identities and activities while online..
- The system will investigate how to convert crime information into a regression problem, which will aid detectives in solving crimes more quickly.
- To protect cyberspace information and data from criminal activities and tendencies.

C. Significance of the Study

- The findings of this study will be useful to cybercrime investigators by assisting them in determining the identity and location of a cyber criminal. This will make further investigation and possible prosecution of the criminals easier.
- The findings of this research will also assist the authorities from various nations in developing legislation to combat cybercrime and prosecute cyber criminals.

D. Scope of the Study

The project is based solely on detecting and predicting crime rates in messages exchanged between two people, as well as identifying the attacker. Following training with appropriate words, the model can predict and estimate crime rates based on the percentage of offensive words. If the percentage is low, the model recommends that users upgrade their firewall and virus protection. If the percentage is high, the computer information (IP Address) and Geo-location are determined.

BACKGROUND

Deep Learning, Machine Learning, and Computer Vision techniques have revolutionized surveillance methods, providing us with new perspectives. The implementation of the Cyber User Identification and Crime Detection System (CUICDS) developed in this research study can discourage cyber users from engaging in criminal activities. Furthermore, this study introduces a novel area of exploration for network security researchers through the cyber user identification approach. The proposed system involves data collection, data preprocessing, constructing predictive models, training and testing datasets, and comparing algorithms. The main objective of this study is to showcase the effectiveness and accuracy of a Machine Learning algorithm in predicting and identifying violent crimes.

II. EXISTING SYSTEM

Existing approaches include k-NN, RF, and Bayes models. Similar to supervised systems, unsupervised systems for classification often use techniques like K-means clustering, Gaussian mixture models, and fuzzy clustering. both fuzzy clustering and C-means clustering. Different training and testing data are subjected to variation in order to analyze the performance. The criminal is eventually located, and the unsupervised Gaussian mixture model shows improved performance when using the detection method. decreasing accuracy and performance qualities. Therefore, it is only helpful if the information about the location of the stake out is accurate. With the help of this information, we can see how continually improving technology has created yet another clever surveillance method.

Limitations:

1. **Less Accuracy in Prediction:** The lack of performance and unreliable results. With 76.56% percentage of accuracy is achieved in detecting the criminal.
2. **Slower in High Dimensional Spaces:** When worked with raw data and non- labels, the linearity decreases with increase in size. Resulting in, unclear and less efficiency.
3. **Increase in Memory Usage:** Due to the less effectiveness of high conditionality spaces, memory storage is high.

III. PROPOSED SYSTEM

This research paper suggests a key-based cyber estimation and identification system to overcome the shortcomings of the current system. This system aims to offer improved performance and accuracy measurements along with cutting-edge features to track the intruder's IP address and Geo-location.

Our system performs the process of tracing the origin of the attack in addition to analyzing and detecting the crime. delivering results that are precise, effective, and clear. This can give you a fresh viewpoint on how to carry out surveillance.

The proposed Cyber Crime Estimation and Identification system offers several advantages over the existing system and other applications. These advantages can be summarized as follows:

1. **Efficient and Accuracy Framework:** Unlike the traditional security measures employed by KNN, CNN, RF and other methods, Our proposed system utilizes a Machine Learning Algorithms like Random Forest Algorithm, Decision Tree and Support Vector Machine. It generates dynamic outcomes while working together, ensuring the probability of belonging to the class, the use of kernel tricks and the creation of hyper-rectangles in the input space to solve non-linear problems.
2. **Enhanced Performance Measures**: SVM RF can highly accurate, generalize better and are interpret-able and SVM (called RF-SVM) to effectively predict gene expression data with very high dimensions. Decision trees are better for categorical data and it deals co-linearity by supporting SVM to work better.
3. **Low Memory Consumption:** Proposed protocol uses less amount of memory storage to maintain operational and functionality parameter.

SVM vs k-NN Clasifier

In SVM, the proper training phase is required, whereas in KNN, data categorization is based on the distance metric. SVM guarantees that the divided data are separated in the best possible way because it is the ideal type. SVM is used to divide binary data into one of two classes, whereas KNN is typically used as a multi-class classifier. The one-vs-one and one-vs-all approaches are used for a multi-class SVM. The one-vs-one principle is used to train n(n 1)/2 SVMs, i.e., one SVM for each pair of classes. The data type is established by the majority output from the combined SVM output, which is fed a pattern that is unfamiliar to the entity. Multi-class categorization is where this method is most commonly applied. The Genuine data are the users 1–32, 49, 51, 53–55, 57–96, and 98–100. The rest are the crime data.

SVMs appear to be computationally intensive because, after the data have been trained, the model can still be used to predict classes even when new unlabeled data are present. However, in the case of KNN, the distance metric is calculated each time fresh unlabeled data are discovered. However, in SVMs, the R parameter must also be fixed, whereas in KNN, only the K parameter must be fixed and the distance metric must be suitable for classification.

Regularization term must be chosen together with the kernel parameters if the classes are linearly inseparable.

TP, TN, FN, and FP classification for attribute shown in Table 1 below.

UID	**Group**		**Attribute 1**					
	Cluster classification based on average	**GD = 0 CD = 1**	**New class SVM classifier: Attribute 1**	**GD = 0 CD = 1**	**TP 0**	**TN 1**	**FP 11**	**FN 10**
1	Genuine User	0	“Crime User”	1	FALSE	TRUE	FALSE	FALSE
2	Genuine User	0	“Crime User”	1	FALSE	TRUE	FALSE	FALSE
3	Genuine User	0	“Crime User”	1	FALSE	TRUE	FALSE	FALSE
4	Genuine User	0	“Crime User”	1	FALSE	TRUE	FALSE	FALSE
5	Genuine User	0	“Crime User”	1	FALSE	TRUE	FALSE	FALSE
6	Genuine User	0	“Crime User”	1	FALSE	TRUE	FALSE	FALSE
7	Genuine User	0	“Genuine User”	0	TRUE	FALSE	FALSE	FALSE

The Table-1 shows that, in contrast to KNN, SVMs show improved accuracy when comparing the accuracy of both classifiers. This can be filtered to obtain and record the data.

ALGORITHMS

Decision Tree Algorithm:

Decision tree are a technique from data mining that categorize new pieces of information into a number of predefined categories. This algorithm analysis has the potential to support an intrusion detection team with the many challenges of defending a network. Due to the ever-increasing volume of data decision trees have the potential to save time for security experts and assist in the analysis of malicious data. It also supplement penetration testing efforts and creating rules to detect malicious activities.

Random Forest Algorithm:

The Random Forest (RF) algorithm is employed for the identification of diverse attack types. By leveraging this algorithm, the system aims to enhance its accuracy through the generation of optimal decisions. Additionally, the Random Forest algorithm offers advantages such as minimizing classification errors, resulting in a more robust and precise system overall.

Accuracy = Samples correctly classfied in test data/ Number of samples in test data

Decision Rate = TP/TP+TN

False Alarm Rate = FP/TN+FP

The performance measures mentioned earlier are obtained from the confusion matrix, which is represented below:

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Where,

TN → True Negative

FN → False Negative.

FP → False Positive

TP → True Positive

SVM Algorithm:

The Support Vector Machine (SVM) is a machine learning technique that can complement the performance of an Intrusion Detection System (IDS), serving as an additional layer of detection to minimize false alarms or as an alternative detection approach. In our study, we compare the effectiveness of our IDS with one-class and two-class SVMs, both in linear and non-linear forms. The findings demonstrate that the linear two-class SVM achieves highly accurate results, while the linear one-class SVM produces comparable outcomes without the need for training datasets associated with malicious data. Additionally, our IDS can benefit from incorporating machine learning techniques to enhance its accuracy when analyzing datasets with diverse features.

For predicting future stock crime rate values, we employ a support vector regression model to categorize the test data. A training set is used to train and optimize the model. Real-time intrusion detection is crucial, and one of the advantages of utilizing SVM for IDS is its speed. SVMs can handle complex classifications independently of the feature space dimensionality, enabling them to learn a broader range of patterns and scale effectively. Furthermore, SVMs can dynamically update the training patterns when encountering new patterns during classification. By employing SVM, we can enhance the accuracy of the IDS. The classifier generated through this technique proves valuable in predicting between two possible outcomes, distinguishing between malicious and non-malicious network traffic.

Performance Evaluation:

The offender is successfully identified using the unsupervised method, specifically the Gaussian mixture model, which exhibits improved performance in the detection process. The detection accuracy for identifying the criminal reaches 76.56%. In the supervised method, when employing the SVM classifier for classification, an accuracy of 89% is achieved. Various performance metrics are computed for multiple attributes, including true positive (TP), false positive (FP), true negative (TN), false negative (FN), false alarm rate (FAR), detection rate (DR), accuracy (ACC), recall, precision, specificity, and sensitivity.

IV. PERFORMANCE EVALUATION

The primary objective of an SVM is to find an optimal hyperplane that effectively separates the training data, maximizing the margin while minimizing complexity and the risk of overfitting. SVM implementation is straightforward, requiring a relatively small training dataset, and it is well-suited for analyzing large datasets. Once the optimal hyperplane is constructed, the SVM classification process is efficient and fast.

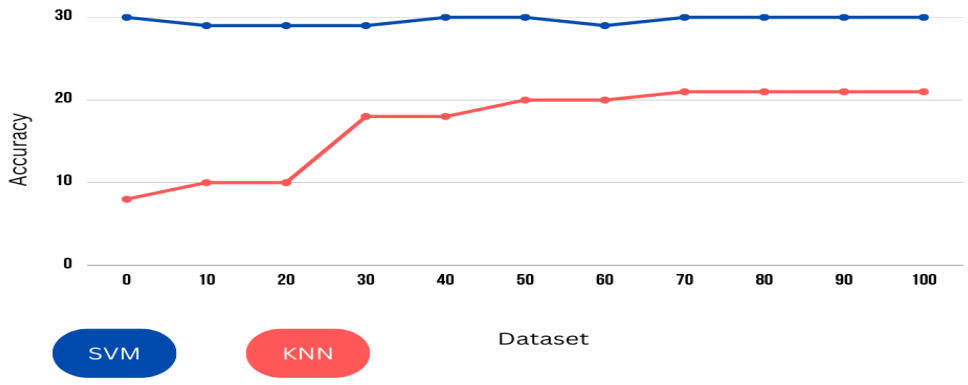
For our SVM-based classifier discussed in this paper, we utilized Matlab and the libSVM software. libSVM is a comprehensive and robust tool used for support vector classification, distribution estimation, and includes various kernel functions. During our experiments, we made modifications to libSVM to introduce randomization in sample selection.

In the case of binary data, an SVM arranges it into one class or another. To handle multi-class classification, we adopted the one-vs.-one and one-vs.-all approaches. In the one-vs.-one approach, a set of n*(n-1)/2 SVMs is trained, where each SVM is designed to differentiate between a specific pair of classes. When an unknown pattern is input, the final decision regarding its class is made by majority voting based on the outputs of all the SVMs. This approach is commonly employed for multi-class classification.

Table 2: Accuracy of Classifiers.

Classifier	Training Set	Test Set	Accuracy Rate (%)
SVM	8000	2000	98.8
KNN	8000	2000	96.47

Apparently, SVMs appear to be computationally tough, and once the data is trained, the model can be employed for predicting classes even though new unlabeled data are encountered. But in case of KNN, every time a new unlabeled data is encountered, the distance metric is computed.



Evaluation of Performance:

For carrying out performance evaluation, various performance metrics are employed such as TP, FP, TN and FN, FAR, ACC—Accuracy, DR—Detection Rate, Specificity, Sensitivity, Precision, Recall, and scores for different attributes.

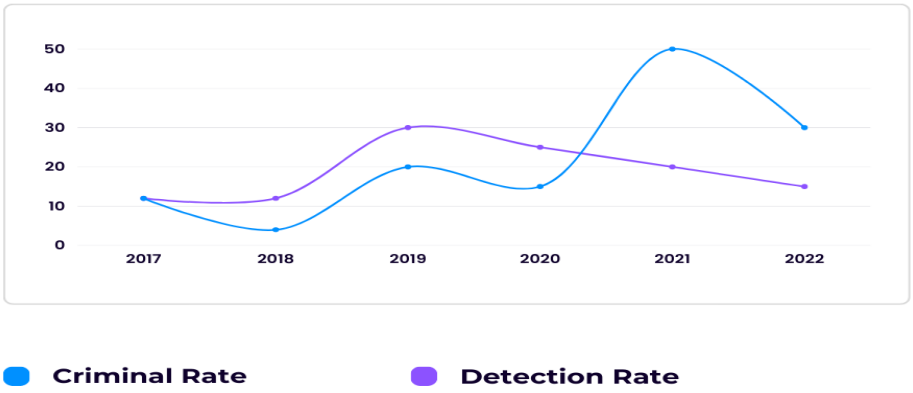
These metrics generate numeric values which are simply comparable and are described as given below.

CR: Crime Rate signifies the ratio of rightly classified instances and the total no. of instances.

$$CR = \frac{Correctly\ classified\ instances}{Total\ number\ of\ instances} = \frac{TP + TN}{TP + TN + FP + FN}$$

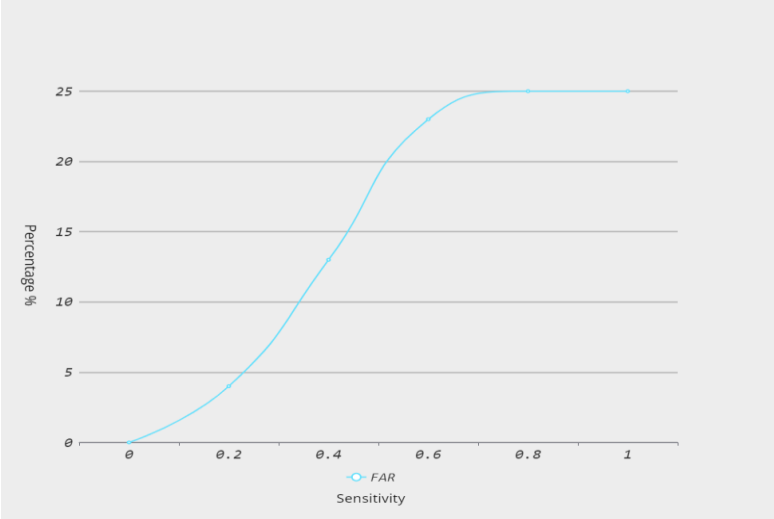
DR: Detection Rate is measured as the ratio between the no. of accurately detected attacks and the total no. of attacks.

$$DR = \frac{Correctly\ detected\ attacks}{Total\ number\ of\ attacks} = \frac{TP}{TP + FN}$$



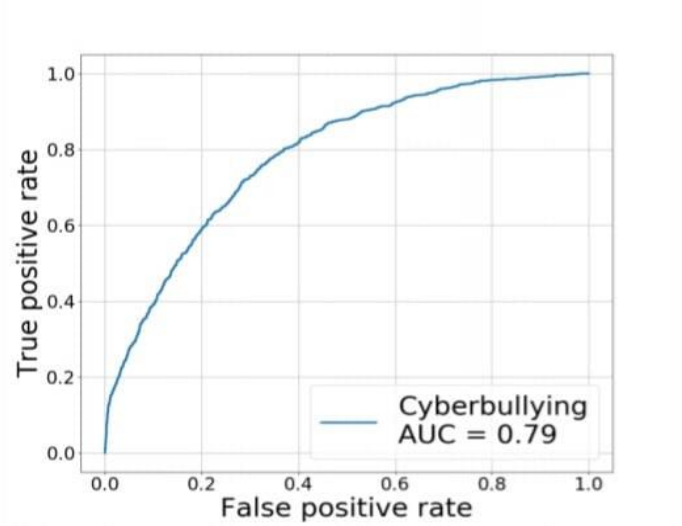
FAR: The FAR stands for "False Acceptance Rate." It's a measure of how likely a biometric security solution is to accept an unauthorized user's access attempt. The FAR of a system is usually expressed as the difference between the number of times the system failed to recognize the user and the number of times it failed to identify the user.

$$\text{FAR} = \frac{Number\ of\ false\ acceptance}{Total\ number\ of\ identification\ attempts} = \frac{FP}{FP + TN}$$



FPR: False Positive Rate, is measured as the ratio between the total no. of normal instances which are identified as an attack and the total no. of normal instances.

$$\text{FPR} = \frac{Number\ of\ normal\ instances\ detected\ attacks}{Total\ number\ of\ attacks}$$

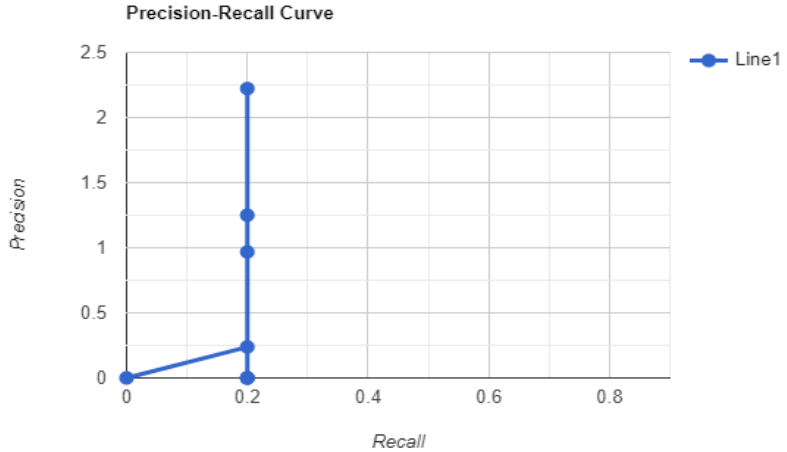


PR: Precision Rate, is the fraction of positively predicted data instances that are positive in actuality.

$$\text{PR} = \frac{TP}{TP + FP}$$

Recall: The Recall, metric computes the percentage (a missing part from the precision) from the real attack instances included by the classifier. Apparently, it is very well required that the classifier has a high recall value. The recall metric is similar to the DR metric.

$$\text{Recall} = \frac{FP + TN}{Total\ Users}$$



Performance metrics using the SVM classifier.

Attribute	TP	TN	FP	FN	FAR	DR	ACC	Precision	Recall	Specificity	Sensitivity	FMI
1	16	64	20	0	0.2381	1	0.8	0.23810	0.2	1	0.7619	0.6667
2	16	64	20	0	0.2381	1	0.8	0.23810	0.2	1	0.7619	0.6667
3	80	0	9	11	1.00	0.879	0.8	2.22222	0.2	0.8791	0.00	0.8889
4	80	0	16	4	1.00	0.952	0.8	1.25000	0.2	0.9523	0.00	0.8909
5	80	0	0	20	0.00	0.8	0.8	0.00000	0.2	0.8	0.00	0.8944

V. CONCLUSION

In our study, we put forward a methodology aimed at enhancing the accuracy of crime-related post detection in social media text messages and identifying the specifics of an intruder based on the type of attack. Our approach involved a two-step process. Firstly, we employed a keyword-based filter to narrow down the relevant messages. Subsequently, we utilized an SVM-based filter and Random Forest Classification to further refine the data by eliminating noise. It is worth noting that existing research has shown that SVM exhibits superior accuracy compared to other classifiers. Hence, we chose to utilize SVM in our study, based on its performance and reliability.

ACKNOWLEDGEMENTS

We would like to extend our heartfelt appreciation to Mrs. P. Yamuna, our Assistant Professor and Internal Guide, for her invaluable support and guidance throughout this project. We are also grateful to our project coordinators, Mrs. Soppari Kavitha and Mr. CH Vijay Kumar, for their assistance and coordination. Additionally, we express our gratitude to Dr. M.V. Vijaya Saradhi, the Head of the Department (CSE) at ACE Engineering College, for his unwavering support and the time he devoted to our project. Their contributions have been instrumental in the success of our work, and we are truly thankful for their presence and guidance.

VI. REFERENCES

[1]. S. Dixon, ACM SIGCOMM Computer Communication Review, Volume 39, Number 5, October 2009, "A Brief History of the Internet"
[2]. Cybercrime and the Victimization of Women: Laws, Rights, and Regulations, by D. Halder and K. Jaishankar. USA: IGI Global, Hershey, PA, 2011.
[3]. J. Srinivas, A. K. Das, and N. Kumar, Future Generation Computer Systems, 2019. "Government rules in cyber security: Framework, standards, and suggestions."
[4]. W. Clay. (2005), Threats and Policy Concerns for Congress Aspect Of computer Attack and Cyber Terrorism Congressional Research Service report to Congress, 2005. reached on February 3rd, 2013 at http://www.history.navy.mil/library/online/computerattack.htm#summ.
[5]. I. M. Venter, R. J. Blignaut, K. Renaud, and M. A. Venter, "The three R's and cyber security education are equally important," 2019; Heliyon
[6]. K. F. Cheung and M. G. H. Bell, Eur. J. Operat. Res., 2019. "Attacker-defender paradigm for cyber security in logistics management against attackers with quantal responses: An introductory research.
[7]. http://www.fema.gov/pdf/onp/toolkit_app_d.pdf. ComputerAttack.Accessed on 12/03/2013.

[8]. M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, J. Inf. Secur. Appl., 2020, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative analysis."
[9]. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," IEEE Access, 2019.
[10]. B. Zhu, A. Joseph and S. Sastry,"A taxonomy of cyber-attacks on SCADA systems," 4th International Conference on Cyber Physical and Social Computing and 2011 International Conference on Internet of Things, pp. 380-388, 2011.
[11]. Kemal Hajdarevic,Adna Kozic and Indira Avdgic,"Using the GNS3 Simulator to Train Network Managers in Ethical Hacking Techniques to Manage Resource Starvation Attacks,Communication and Automation Technologies (ICAT),International Conference on Information, Sarajevo, Bosnia-Herzegovina, pp. 1-6 , Oct 26- 28, 2017
[12]. B. Vijay B, G. Ajay and A. Ala, Detection of masquerade attacks on Wireless Sensor Networks, 2010. Available at http://www.ists.dartmouth.edu/library/343.pdf.Accessed on 13/03/2013.
[13]. INFOCOM,Identity-based Attack Detection in Mobile Wireless Networks. Proceedings of the IEEE held in Shanghai, April 10-15, 2011, pp. 1880-1888.Available at http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp. Accessed on 02-04-2014.