Automatic Method to Predict and Classify Cyber Hacking Breaches Using Deep Learning

[1]Ms. R Nivetha, [2]G Varun Kumar, [3]Ganesh Rokaya, [4]E Trinath, [5]E Uma Maheswara Reddy

[1]Assistant Professor, [2,3,4,5]Student
Department of Computer Science and Engineering
Bharath Institute of Higher Education And Research
Chennai, India-600073

Abstract- **Analyzing cyber incident data sets is an essential way for better understanding the development of the threat environment. We are aware of several cyber breaches and hacking incidents currently. In this project, we investigate numerous cyber-attacks and breaches, as well as the methods used to carry them out and devise alternatives. We demonstrate that instead of spreading these assaults based on their autocorrelations, we should characterize both hacking breach event inter- arrival durations and breach sizes using a stochastic process. We derive several cyber security insights, including the fact that the danger of cyber hacking is increasing in frequency.**

For our study, we will use techniques like Convolution Neural Network (CNN) and Recurrent Neural Network (RNN) to analyze our data. The findings show that Recurrent Neural Network (RNN) performs better than Convolutional Neural Network (CNN).

Keywords: **Classification, Cyberattacks, Deep Learning, Protection, Recovery.**

I. INTRODUCTION

An information breach is a security incident when sensitive, protected, or confidential data is copied, communicated, viewed, stolen, or used by an unauthorized person. An information breach is the planned or inadvertent entry of secure or private/classified data to an untrusted domain. Different expressions for this miracle involve unintended data divulgence, information leak and information spill. This may incorporate occurrences, for example, robbery or loss of advanced media, for example, PC tapes, hard drives, or smart phones such media whereupon such data is put away decoded, posting such data on the internet or on a PC generally available from the Internet without legitimate data security safeguards, exchange of such data to a framework which isn't totally open yet isn't fittingly or formally.

Authorize for security at the affirmed dimension, for example, decoded email - or exchange of such data to the data frameworks of a conceivably unfriendly office, for example, a contending organization or a remote country, where it might be presented to increasingly serious unscrambling strategies. While mechanical arrangements may harden digital framework.

This drives us to characterize the emergence of information rupture events. This not solely will depth our grasp of information breaches, however in addition provide light on new approaches for easing the damage, for example, protection. Many trust that protection will be worthwhile, but the improvement of correct cyber danger measurements to govern the work of protection rates is beyond the compass of the current knowledge of information breakdowns in this article, we make the following pledges. We believe that as opposed to circulating the ruptures we should show by stochastic technique both the hacking break incidence entomb entrance timings and rupture sizes.

II. LITERATURE REVIEW

As proposed by [1] an expert system called OPENSke that was designed by using Drool's tool. This system took system information as input and identified the vulnerability found. System information was the information of the network such as hosts present in the system, user account, assets and applications running on them. Vulnerabilities found was the output of any vulnerability scanner such as OVAL scanner or NESSUS etc. that provides the vulnerability found in the system [1]. This gave as output weaknesses found compromised software or assets and executed attack pattern. Weaknesses found in described by the CWE (Common Weakness Enumeration) that have satisfied current vulnerability in the system. It also told about the software or the assets that have been compromised were risky to use. Executed attack pattern constitute a list of CAPEC (Common Attack Pattern Enumeration and Classification) attack patterns. This system was experimented on several different systems and identified the vulnerability present. The major drawback of the system was that it had not taken into consideration the role of time in execution of attacks. In this system information about the network needed to be entered manually, so could also be automated to detect the network topology and run the system on frequent basis. Rules in this system were executed arbitrarily and this could be improved by going through the workflow process. Kerim Goztepe[2] designed fuzzy rule based expert system for cyber security named as Fuzzy Rule Based Cyber Expert System(FRBCES). Major parts in this were data collection, defining variables i.e. input and output and then implementation. Input variables used in this were cyber technique, cyber intruder’s target, cyber intruder, aim of cyber intruder. Techniques used for this were DOS attack, network International Conference on Computing, Communication and Automation (ICCCA2015) 243 attacks, worm, malware, social engineering, Trojan horse etc. Cyber intruder could be inside or outside the organization, special staff, hacker or any enemy. Aim of the intruder was identified as to gain control of the system, block web pages, capturing critical information, etc. Output provided by this was hardware, software or user i.e. what needed to be done with these components. For example in case of softwareprovide solution such as use special software or update system. This system uses Mamdani fuzzy inference engine or system. In expert system for computer security presented in this paper, data about cyber attacks, their symptoms and countermeasures is collected from the different papers of the journals and conferences online sources. Different authors classify the security and security threat each has its own view. K Ahmed et.al [3] define the security attack as active and passive attack and these attack use technique such as interruption, interception, modifications and fabrications. Attacks that use interception as technique are traffic analysis, release of message content and sniffing. Modification technique include MITM (Man-In-The Middle). Similarly other technique fabrications include attacks such as replay attack and identify spoofing. DOS attack and its variants (DDOS and DRDOS) and SQL injection comes under Interruption. D. Welch [4] defined wireless network attack taxonomy. This was designed based upon the security principles they affect and their countermeasures. For example threat violating confidentiality, integrity and availability. Threats that affect the confidentiality of the system are traffic analysis, active and passive sniffing. The threats or attack that affect integrity of the system are session hijacking, replay attack, unauthorized access to the computer or network. The solutions to remove these attacks are implement firewall to block undesirable traffic, mutual authentication and encryption. According to [5], DOS attack in networks can occur at different layers like physical layer, network layer and transport layer. Simplest of the DOS attack is to consume all the resources of the victim by sending a large number of packets. In [6] security threats in wireless sensor networks are defined and these are classified as attack on the different layers. The attacks can be broadly classified as modification, interruption and fabrication.

III. OVERVIEW

As we are living in a connected online environment, most of our routine conversations and business operations now take place over the Internet. Since cyber infrastructure is very susceptible to attacks, the dangers in cyber space move at the speed of light. With the speed of cyber activity and enormous amount of data consumed, the safety of cyber space cannot be managed by any physical equipment or by human involvement alone. It requires extensive automation to identify dangers and to make intelligent real-time judgments. It is difficult to design software using standard methods to successfully fight against the continually developing threats. It may be handled by adding bio inspired computer approaches of artificial intelligence to the program. The goal of this research is to examine the potential of artificial intelligence in tackling cybercrime challenges.

III. Deep Learning:

In the statistical context, Deep Learning is described as an application of artificial intelligence where accessible knowledge is employed via algorithms to process or help the processing of statistical data. While Deep Learning contains principles of automation, it needs human guidance. Deep Learning needs a high degree of generalization to obtain a system that works well on yet unseen data examples.

Deep learning is a relatively new topic within Computer Science that includes a set of data analysis tools. Some of these strategies are based on widely recognized statistical methods (e.g. logistic regression and principal component analysis) whereas many others are not.

Most statistical approaches follow the concept of finding a single probabilistic model that best represents observed data within a class of related models. Similarly, most Deep learning

*1.*CLASSES OF DEEP LEARNING

There are two primary kinds of Deep learning techniques:

1. Supervised Deep Learning versus unsupervised Deep Learning. An Examples of supervised learning Logistic regression (statistics) versus Support vector Deeps (Deep learning) Logistic regression, when applied for prediction purposes, is an example of supervised Deep learning. In logistic regression, the values of a binary response variable (with values 0 or 1, say) as well as several predictor variables (covariates) are observed for several observation units. These are termed training data in Deep learning language. The essential assumptions are that the response variable follows a Bernoulli distribution .

2.The logistic regression optimization problem comes from probability theory whereas that of SVM comes from geometry.

Other supervised Deep learning approaches addressed later in this briefing include decision trees, neural networks, and Bayesian networks. B. Examples of unsupervised learning Principal component analysis (statistics) versus Cluster analysis (Deep learning). The main example of an unsupervised Deep learning technique that comes from classical statistics is principal component analysis, which seeks to “summarize” a set of data points in high-dimensional space by finding orthogonal one-dimensional subspaces along which most of the variation in the data points is captured. The word “unsupervised” merely refers to the fact that there is no longer a response variable in the present context. Cluster analysis and association analysis are examples of non-statistical unsupervised Deep learning approaches. The former aims to discover intrinsic grouping structure in provided data, while the latter seeks to find co-occurrence patterns of items.

IV. ALGORITHM

1. Detecting Port Scan Attempts with Comparative Analysis of Deep Learning and Support Vector Deep Algorithms

Compared to the past, improvements in computer and communication technology have produced significant and advanced changes. The employment of new technologies provides enormous benefits to individuals, companies, and governments, but it produces some issues against them. For example, the privacy of essential information, security of stored data platforms, availability of knowledge etc. Depending on these concerns, cyber terrorism is one of the most critical issues in today’s world. Cyber terror, which caused a lot of issues to individuals and institutions, has reached a level that could jeopardize public and country security by various parties such as criminal organizations, professional persons, and cyber activists. Thus, Intrusion Detection Systems (IDS) have been designed to avoid cyber-attacks. In this work, deep learning, and support vector Deep (SVM) techniques were used to detect port scan attempts based on the new CICIDS2017 dataset and 97.80%, 69.79% accuracy rates were attained respectively.

2. Detecting cyber-attacks using a CRPS-based monitoring approach

Cyber-attacks can adversely impair the security of computers and network infrastructure. Thus, building an efficient anomaly detection technique is vital for information.

Protection and cyber security. To successfully detect TCP SYN flood attacks, two statistical techniques based on the continuous ranked probability score (CRPS) metric have been devised in this study. Specifically, by merging the CRPS measure with two conventional charts, Shewhart and the exponentially weighted moving average (EWMA) charts, innovative anomaly detection methodologies were developed: CRPS-Shewhart and CRPS-EWMA. The efficiency of the suggested approaches has been verified using the 1999 DARPA intrusion detection evaluation datasets.

3. A Taxonomy of Malicious Traffic for Intrusion Detection Systems

With the increasing number of networks threats it is vital to have a knowledge of existing and upcoming network risks to create better intrusion detection systems. In this study we propose a taxonomy for describing network attacks in a consistent fashion, allowing security researchers to focus their efforts on constructing accurate intrusion detection systems and focused datasets.

4. Parameter-Invariant Monitor Design for Cyber–Physical Systems

The intimate link between information technology and the physical environment inherent in cyber-physical systems (CPS) can challenge standard methodologies for monitoring safety and security. Data obtained for robust CPS monitoring is generally sparse and may lack sufficient training data describing crucial events/attacks. Moreover, CPS often function in various contexts that can have high inter/intra-system variability. Furthermore, CPS monitors that are not robust to data sparsity and inter/intra-system variability may result in inconsistent performance and may not be trusted for monitoring safety and security. To overcome these issues, this paper provides current work on the design of parameter-invariant (PAIN) monitors for CPS. PAIN monitors are constructed such that unknown events and system fluctuation little impair the monitor functionality. This article demonstrates how PAIN designs can achieve a constant false alarm rate (CFAR) in the presence of data sparsity and intra/inter system volatility in real-world CPS. To demonstrate the design of PAIN monitors for safety monitoring in CPS with different types of dynamics, we consider systems with networked dynamics, linear-time invariant dynamics, and hybrid dynamics that are discussed through case studies for building actuator fault detection, meal detection in type I diabetes, and detecting hypoxia caused by pulmonary shunts in infants. In all applications, the PAIN monitor is demonstrated to have (slightly) less variance in monitoring performance and (often) outperforms other competing approaches in the literature. Finally, an initial application of PAIN monitoring for CPS security is provided a long with obstacles and research possibilities for future security monitoring deployments.

3. PROBLEM STATEMENT AND METHODOLOGY

3.1 PROBLEM DEFINITION

The primary problems to produce the future generation of intelligent Systems are: -

- Slowing down systems, crashing a system, and its lack of scalability.
- System sending irregular communications it might potentially lead to the identification of the scan.
  Less security.
- Time consumption.
- Common folks can be deceived at any time.

3.2 METHODOLOGY

The system will look at ways to convert crime information into a data-mining problem, so that it will help investigators in solving crimes faster criminal analysis based on available information to extract criminal patterns. Using various data mining approaches, frequency of happening crime can be anticipated based on territorial distribution of current data Crime recognition.

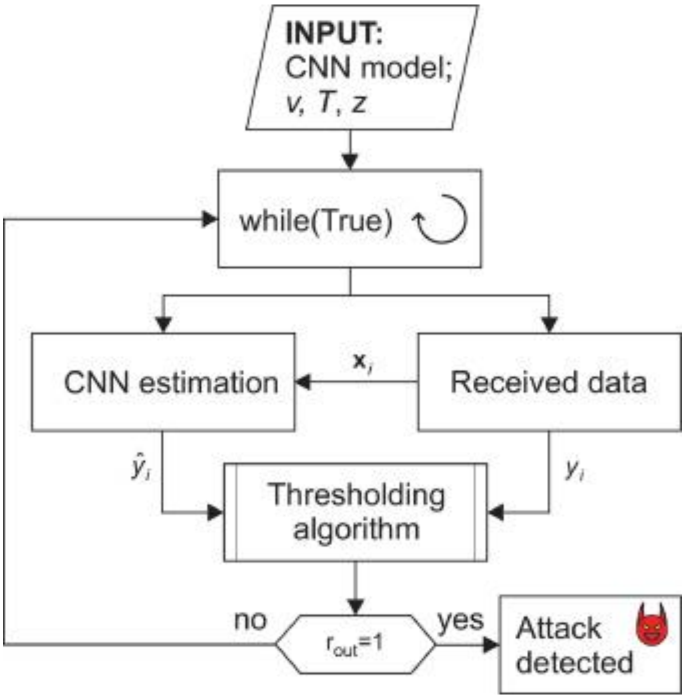
3.2.1 Existing System:

Convolution Neural Network (CNN): In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most typically employed to interpreting visual vision. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on its shared-weights architecture and translation invariance qualities. They have applications in image and video recognition, recommender systems, image classification, Image segmentation, medical image analysis, natural language processing, brain-computer interfaces, and financial time series.

[7]CNNs are regularized versions of multilayer perceptron's. Multilayer perceptron's usually denote completely connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The "fully-connectedness" of these networks renders them prone to overfitting data. Typical means of regularization include adding some form of magnitude measurement of weights to the loss function. CNNs use a different strategy towards regularization: they take advantage of the hierarchical structure in data and create more complex patterns utilizing smaller and simpler patterns. Therefore, on the scale of connection and complexity, CNNs are on the lower extreme.

Convolutional networks were motivated by biological processes in that the connection pattern between neurons matches the organization of the animal visual corte Individual cortical neurons respond to stimuli only in a narrow section of the visual field known as the receptive field. The receptive fields of various neurons partially overlap such that they cover the full visual field.

CNNs employ comparatively less pre-processing compared to other image classification techniques. This means that the network learns the filters that in older algorithms were hand-engineered. This independence from prior knowledge and human effort in feature creation is a big advantage.



3.2.2 Proposed System:

Recurrent Neural Network (RNN):
A recurrent neural network (RNN) is a class of artificial neural networks whose connections between nodes create a graph along a temporal sequence. This permits it to exhibit temporal dynamic behavior. Derived from feed forward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs. This makes them useful for tasks such as unsegmented, linked handwriting recognition or speech recognition.

The phrase “recurrent neural network” is used indiscriminately to refer to two broad kinds of networks with a similar overall structure, where one is finite impulse, and the other is infinite impulse. [8]Both sorts of networks display temporal dynamic behavior. A finite impulse recurrent network is a directed acyclic graph that can be unrolled and substituted with a strictly feed forward neural network, while an infinite impulse recurrent network is a directed cyclic graph that cannot be unrolled.
Both finite impulse and infinite impulse recurrent networks can contain additional stored states, and the storage can be under direct control by the neural network. The storage can alternatively be replaced by another network or graph, if that involves time delays or has feedback loops. Controlled states are referred to as gated state or gated memory and are part of memory networks (LSTMs) and gated recurrent units. This is also termed Feedback Neural Network (FNN).

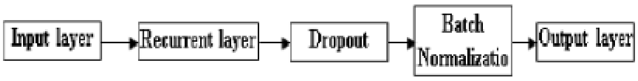


Fig: RNN and unfolded RNN

V. FLOW CHART

5.1 MODULE DESCRIPTION

- Input dataset.
- Analysis of size of data set.
- Oversampling.
- Training and Testing subset.
- Apply algorithms.
- Predicting results.

1. Input dataset: Dataset can be collected from an online source provider called UCI repository. We have collected a set of criminal datasets which we are going to evaluate. Then training the data set also for the comparison of the non-criminal datasets has also been taken.

2. Analysis of data set: Here the analysis of data set takes place. The size of data is taken into consideration for the data process.

3. Oversampling (Using SMOTE): we have produced a complete history of all offenses that been complained over a given length of time and it is sampled to fix a threshold value.

4. Training and Testing Subset: As the dataset is uneven, many classifiers display bias for majority classes. The qualities of minority class are viewed as noise and are ignored. Hence it is advised to select a sample dataset.

5. Applying algorithm: Following are the classification methods used to assess the sub-sample dataset. a. Convolution Neural Network (CNN) and b. Recurrent Neural Network (RNN)

6. Predicting results: The test subset is put to the training model. The metrices utilized is accuracy. The ROC Curve is plotted, and the desirable results are achieved.

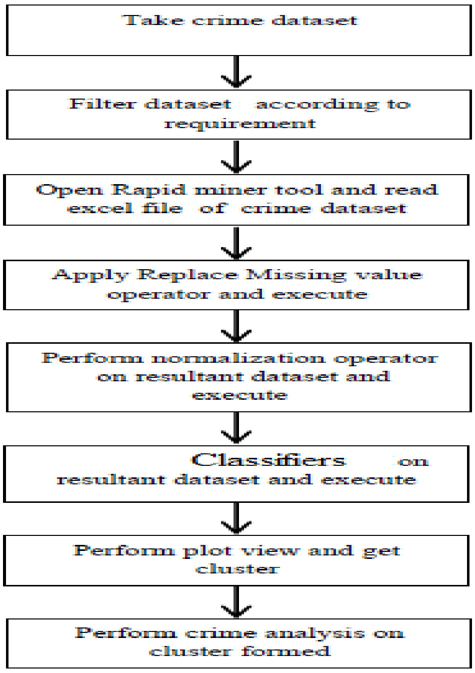
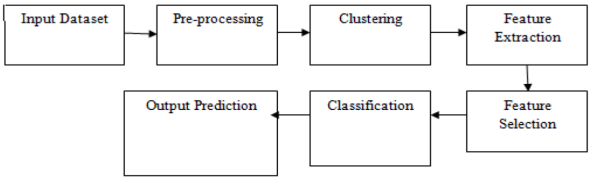
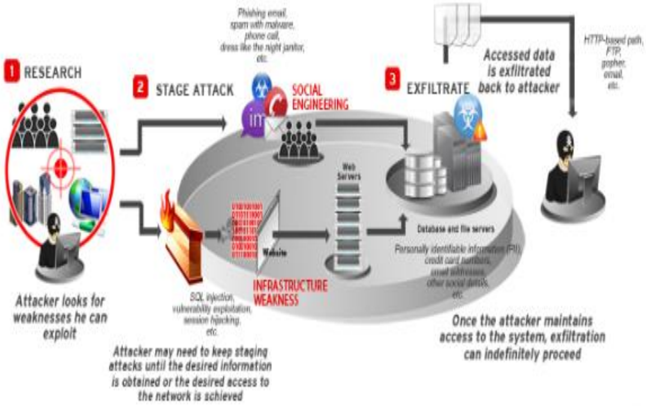

Take crime dataset
Filter dataset according to requirement
Open Rapid miner tool and read excel file of crime dataset
Apply Replace Missing value operator and execute
Perform normalization operator on resultant dataset and execute
Classifiers on resultant dataset and execute
Perform plot view and get cluster
Perform crime analysis on cluster formed

Fig 3. Flow-chart of the methodology used for Recurrent Neural Network

VI. COMPARATIVE STUDY OF EXISTING AND PROPOSED SYSTEM

[9]In our project we will be using algorithms such as Convolution Neural Network (CNN), and Recurrent Neural Network (RNN) for analyzing our outcomes. From the results is demonstrated that Recurrent Neural Network (RNN) works better than Convolution Neural Network (CNN). In our proposed system we will be employing the proposed methodology method which leads to greater accuracy compared to the existing system. From this we are receiving a decreased noise ratio and good accuracy therefore we can state that our proposed system performs better than the present system.




1 RESEARCH
Attacker looks for weaknesses he can exploit
2 STAGE ATTACK
Phishing email, spam with malware, phone call, dress like the night janitor, etc.
SOCIAL ENGINEERING
Website
Web Servers
INFRASTRUCTURE WEAKNESS
SQL injection, vulnerability exploitation, session hijacking, etc.
Attacker may need to keep staging attacks until the desired information is obtained or the desired access to the network is achieved
3 EXFILTRATE
Accessed data is exfiltrated back to attacker
HTTP-based path, FTP, gopher, email, etc.
Database and file servers
Personally identifiable information (PII), credit card numbers, email addresses, other social details, etc.
Once the attacker maintains access to the system, exfiltration can indefinitely proceed

VII. SOFTWARE DESCRIPTION

A. Python

Python is a tremendously capable dynamic, object-oriented programming language that is employed in a wide variety of application disciplines. It offers significant support for integration with other languages and technologies and comes with vast standard libraries. To be precise, the following are some unique properties of Python:

- Very straightforward, legible syntax.
- Strong introspection capabilities.
- Full modularity.
- Exception-based error management.
- High level dynamic data types.
- Supports object oriented, imperative and functional programming techniques.
- Embeddable.
- Scalable
- Mature

With so much freedom, Python encourages the user to consider problem centric rather than language focused on other circumstances. These capabilities make Python a better alternative for scientific computing.

B. Open CV

Open CV is a collection of programming functions for real time computer vision first developed by Intel and now supported by Willow garage. It is free for use under the open-source BSD license. The collection comprises more than five hundred optimized algorithms. It is utilized all over the world, with forty thousand people in the user group. Uses span from interactive art to mine inspection, to advanced robotics. The library is mostly built in C, which makes it adaptable to some

specialized systems such as Digital Signal Processor. Wrappers for languages such as C, Python, Ruby, and Java (using Java CV) have been developed to enable adoption by a wider audience. The newest releases have interfaces for C++. It focuses mostly on real-time image processing. Open CV is a cross-platform library, which may run on Linux, Mac OS, and Windows. To date, Open CV is the best open-source computer vision library that developers and researchers can conceive of.

C. Tesseract

Tesseract is a free software OCR engine that was created at HP between 1984 and 1994. HP released it to the community in 2005. Tesseract was introduced at the 1995 UNLV Annual Test OCR Accuracy and is now developed by Google distributed under the Apache License. It can currently recognize 6 languages and is fully UTF8 enabled. Developers can train Tesseract using their own typefaces and character mapping to reach ideal efficiency science and machine learning difficulties, it also offers a variety of methods that may be used for data analysis.

Python - Environment Setup

Python is available on a wide variety of platforms including Linux and Mac OS X. Let's understand how to set up our Python environment.

Local Environment Setup

Open a terminal window and type "python" to find out if it is already installed and which version is installed.

- Unix (Solaris, Linux, FreeBSD, AIX, HP/UX, SunOS, IRIX, etc.)
- Win 9x/NT/2000
- Macintosh (Intel, PPC, 68K)
- OS/2
- DOS (multiple versions)
- Palm OS
- Nokia mobile phones
- Windows CE
- Acorn/RISC OS
- BeOS
- Amiga
- VMS/OpenVMS
- QNX
- VxWorks
- Psion
- Python has also been ported to the Java and .NET virtual Deeps.

IX. CONCLUSION

The frequency of ordinary data breaches around the world indicates how real the danger of critical infrastructure assault as the hackers develop in terms of sophistication and technological expertise, and as the key information infrastructure gets bigger and more sophisticated, it is increasingly vulnerable to attack. We can treat them like an act of terrorism which requires action under the Internal Security Act. If we pursue this route, we must be prepared for the repercussions. What is more convincing is the need to strengthen the security itself. As indicated in this essay, a multi-prong action is required; one that incorporates a mixture of technology, competency of staff, prudence, and effective legal framework. At this end, it is noteworthy that there are a few issues that emerged from this initial study that can be made an agenda of future direction. Firstly, from the technological standpoint, there is a need to examine emerging approaches that endanger the security of key information infrastructure. Secondly, from the perspective of law and policy, governments need to ensure that each sector classified as essential infrastructure should be properly protected both by legal and policy instruments.

REFERENCES:

[1] M.M. Gamal, B. Hasan, and A.F. Hegazy, “A Security Analysis Framework Powered by an Expert System,” International Journal of Computer Science and Security (IJCSS), Vol. 4, no. 6, pp. 505-527, Feb. 2011.

[2] K. Goztepe, “Designing a Fuzzy Rule Based Expert System for Cyber Security,” International Journal Of Information Security Science, vol.1, no.1, 2012 [3] K. Ahmad, S. Verma, N. Kumar, and J. Shekhar, “Classification of Internet Security Attacks,” Proceedings of the 5th National Conference, March 2011.

[4] D. Welch, “Wireless Security Threat Taxonomy,” Information Assurance Workshop. IEEE Systems, Man and Cybernetics Society, pp 76-83, June 2003.

[5] G. Kulkarni, R. Shelk , K. Gaikwad, V. Solanke , S. Gujar, and P. Khatawkar, “Wireless Sensor Network Security Threats,”

[6] M. Panda “Security Threats at Each Layer of Wireless Sensor Networks,” International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 11, Nov. 2013.

[7] D. Welch, “Wireless Security Threat Taxonomy,” Information Assurance Workshop. IEEE Systems, Man and Cybernetics Society, pp 76-83, June 2003.

[8] Vidushi Sharma, Sachin Rai, Anurag Dev” A Comprehensive Study of Artificial Neural Networks” International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 10, October 2012.

[9] Shaiqua Jabeen, Shobhana D. Patil, Shubhangi V. Bhosale, Bharati M. Chaudhari, Prafulla S. Patil” A Study on Basics of Neural Network” International Journal of Innovative Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2017.

[10] Devikrishna K S, Ramakrishna B B “An Artificial Neural Network based Intrusion Detection System and Classification of Attacks”International Journal of Engineering Research and Applications (IJERA) Vol. 3, Issue 4, Jul-Aug 2013, pp. 1959- 1964.