SPCSJ

SCIENTIFIC AND PRACTICAL CYBER SECURITY JOURNAL

VOL3 No3
SEPTEMBER 2019

ISSN 2587-4667

Scientific and Practical Cyber Security Journal (SPCSJ) 3(3): 1 - 11 ISSN 2587-4667 Scientific Cyber Security Association (SCSA)

Proposed Framework for Effective Detection and Prediction of Advanced Persistent Threats Based on the Cyber Kill Chain

Faisal A. Garba
Department of Computer Science Education, Sa'adatu Rimi College of Education, Kano, Nigeria.

[1] Sahalu B. Junaidu, [2]Barroon I. Ahmad, [3]Abdoulie M. S. Tekanyi
[1,2]Department of Computer Science, Ahmadu Bello University, Zaria, Nigeria
[3]Department of Electrical & Computer Engineering, Ahmadu Bello University, Zaria, Nigeria

ABSTRACT

The cost of data breach resulting from cyber attacks is estimated to be $3.62 million dollars worldwide according to a report. Advanced Persistent Threat (APT) is a targeted cyber attack that is tailored, proceeds at a stealth and has a high objective. The state of the art security monitoring tools have failed in their attempts to detect APT. Therefore, there is a need for a solution that is fool-proof in the detection of an APT. This paper proposed the use of cyber kill chain to detect the various attack methodologies used in an APT campaign and to correlate and predict the existence of an APT attack. APT attack deploys various attack techniques which are mapped to the stages of the cyber kill chain. For each of those techniques, an attack detection methodology has been proposed in this paper. The detection result of each of these methodologies, will now be correlated in the correlation module to ascertain whether there is an ongoing APT attack and raise an alert. The result from this research work will be evaluated against a current related work. This research will therefore advance the state of the art in APT attack detection.

KEYWORDS: Advanced Persistent Threat (APT), cyber kill chain (CKC), data breach, cyber attack, APT detection.

I. Introduction

Complex, long-term set of actions aimed against specific persons, organizations or companies is referred to as Advanced Persistent Threat (APT). Adversaries often study their targets for months before launching the attack. Adversaries maintain stealth and can exfiltrate data for a long period of time (Rot and Olszewski, 2017). The targets are mostly companies, government agencies and even individuals. The APT attacker can be an individual, organized crime group or nation state actors. It could take days or years before an APT attack is detected. When the attacker discovers that he has been detected, he might become more violent, change the method of attack or resort to an alternate course of action (Baksi and Upadhyaya, 2017). The state of the art security monitoring tools have failed in their task to detect APT (NIS Platform, 2014; Oprea *et al*., 2015; ENISA, 2018). There is therefore the need for an effective APT detection framework.

II. The Cyber Kill Chain (CKC)

The CKC also called the Intrusion Kill Chain (IKC) is a model that describes the phases of intrusions proposed by Hutchins *et al.* (2011). The CKC is a seven phase model that describes the stages APT actors follow to achieve their objectives.

a. Reconnaisance

This is the planning stage of the cyber attack. During this stage the attackers conduct a research on their target. Attackers harvests email addresses, identify employees on social media, gather press releases, contract awards, conference attendee lists and search for corporate Internet facing servers (Martin, 2015). The target can be an individual or an organization. Reconnaisance can be broken down into target identification, selection and profiling. Reconnaisance can be passive or active. In passive reconnaisance, the target is unaware of the process. Active reconnaisance on the other hand involves a deeper probing of the victim's information technology infrastructure which may trigger alert of the victim's security monitoring tools.

Table 1 gives examples of reconnaissance techniques and techniques used for both passive and active attacks.

Table 1: Reconnaissance Techniques (Yadav and Mallari, 2016)

	Reconnaissance Techniques	Type of Reconnaissance	Techniques Used
1.	Target identification and selection	Passive	Domain names, WHOIS records from APNIC, RIPE and ARIN
2.	Target profiling		
	Target social profiling	Passive	Social Networks, Public Documents, Reports and Corporate Websites
	Target system profiling	Active	Pingsweeps, Fingerprinting, Port scanning and services
3.	Target validation	Active	SPAM messages, Phishing mails and social engineering.

b. Weaponization

This is an operation preparation stage. Automated tools are used in generation of malware. A weaponizer is developed by coupling malware and exploit into a deliverable payload. A decoy document is chosen to be delivered to the victim for file based exploits (Martin, 2015). It is specifically the binding of software/application exploits with a Remote Access Trojan (RAT). Weaponization involves the use of two components; RAT and exploits. RAT is the payload of the cyber weapon. RAT is a sofware that is installed on the victim's machine to give access to the attacker. RAT is made up of two parts; a client and a server. Exploits serve as a carrier for RAT and facilitates the execution of the RAT. The main reason behind the use of RAT is to avoid victim's attention while establishing a stealth backdoor access using RAT (Yadav and Mallari, 2016).

c. Delivery

At this stage the operation is launched. The malware is conveyed to the target at this stage. Some user actions may be required like downloading and executing malicious files or visiting malicious web pages on the Internet. Some attacks are performed without user interactions by exploiting network devices e.g. CVE-2014-3306, CVE-2014-9583 (Mitre, 2014). Multiple delivery methods are usually employed since no single method can guarantee 100% success (Yadav and Mallari, 2016).

d. Exploitation

This stage is where the exploits is triggered (Yadav and Mallari, 2016). The attackers exploits a vulnerability to gain access. Exploits may be triggered by an adversary for server based exploits or by a victim through opening an attachment of malicious email or clicking a malicious link (Martin, 2015). Exploits might not usually be successful unless the following conditions are matched.

1. Victim is using the operating system or software for which the exploit has been created.
2. The software/operating system not updated or upgraded to the newest version
3. End host protection mechanism should not be able to detect the exploit or payload

Table 2 gives examples of delivery mechanisms and their peculiar characteristics.

Table 2: Delivery Mechanism (Yadav and Mallari, 2016).

	Delivery Mechanism	Characteristics
1	Email attachments	Enticing email content is composed to appeal to the user.
2.	Phishing attacks	Fake websites is used to harvest user credentials
3.	Drive by downloads	Intentionally or unintentionally victim is lured into

		downloading malicious content.
4.	USB/Removable media	Malware is kept in a USB device to attack victims
5.	DNS cache poisoning	Vulnerabilities in DNS are utilized to divert internet traffic from legitimate

The payload then connects to its Command & Control (C & C) counterpart to inform about successful execution and commands to execute. Exploits are made from the Common Vulnerabilities and Exposures (CVEs) publicly made available (Yadav and Mallari, 2016). Exploits are also made available from vulnerabilities discovered through fuzzing methodology (Yadav and Mallari, 2016).

e Installation

Installation stage is when the attackers install a persistent backdoor or implant in the victim environment to maintain access for an extended length of time (Martin, 2015). Malware utilizes droppers and downloaders to maintain stronghold on the victim's machine. Dropper installs and executes the malware on the target machine. Dropper first disables the endpoint protection on the device and hides the installed malware (Yadav and Mallari, 2016). Downloaders performs a similar function as droppers with the exception that they do not contain the malicious payload. The malicious payload is downloaded later when the downloader connects to a remote repository. Malware authors now employ the following techniques to stealthy stronghold and hidden installations (Yadav and Mallari, 2016).

1. Anti debugger and anti emulation
2. Anti antivirus
3. Rootkit and bootkit installation
4. Targetted delivery
5. Host based encrypted data exfiltration

f. Command & Control

In this stage a command channel is opened by a malware to enable the attackers control the victim remotely. A two way communication channel to the Command & Control (C2) infrastructure is opened. The C & C channels are mostly over web, DNS and email protocols. The C2 might be owned by the attacker or might be another victim's network (Martin, 2015). The aim of the C & C channels is to provide a secret channel for issuing commands to infected machines. The three types of the C & C communication structures are:

1. Centralized structure
2. Decentralized structure
3. Social network based structure

Attackers employ the following techniques to achieve stealth and unidentified communication channel (Yadav and Mallari, 2016).

1. Internet Relay Chat (IRC)
2. TCP/HTTP/FTP
3. Steganography
4. The Onion Router (TOR)

Attackers also deploy the following techniques to remain stealth (Yadav and Mallari, 2016).

1. DNS Fast Flux
2. DNS as a medium
3. Domain Generation Algorithm

g. Act on Objectives

The attackers at this stage have a hands-on keyboard access to their victim and can now accomplish their mission. The objectives ranges from harvesting user credentials, escalation of privileges, information gathering, lateral movement, data exfiltration (Martin, 2015) and espionage. Sometimes attack might be physical like in the case of Stuxnet (Angle *et al*., 2017). Attack can lead to the destruction of system hard drive or device drivers. Attacker may lead to the Central Processing Unit (CPU) using its highest capability for a very long duration which leads to the damage of the CPU hardware (Yadav and Mallari, 2016).

Although there are other models that seek to describes the stages of an APT attack, for example the Mandiant (Aldridge, 2016), the Dell Secureworks models (Dell SecureWorks, 2012) and the APT attack lifecycle (Ghafir and Prenosil, 2016) the CKC model is the most widely known and more frequently cited (Herlow, 2015).

III. Comparison of Classification Algorithms Prediction Accuracy

Kotsiantis (2007) reviewed supervised machine learning classification algorithms. In the review the study ranked Support Vector Machine (SVM) as the algorithm with the highest accuracy followed by Neural Networks, Decision Trees, kNN and Rule Learners on the same position followed by Naive Bayes in the last position.

Wei-Chih and Yu (2009) performed email spam filtering using Naive Bayesian, SVM with RBF kernel, Linear kernel, SVM using Taguchi Method and SVM using grid search. SVM using grid search has the highest accuracy followed by SVM using Taguchi Method which is proposed work of Wei-Chih and Yu (2009).

Mezghani *et al*. (2010) compared the prediction accuracy of SVM kernels with three other popular learning algorithms: Naive Bayes (NB), Decision Tree C4.5 and Multi Layer Perceptron (MLP) for speaker identification. SVM trained using polynomial kernel emerged the best for speaker identification tasks and SVM was the best compared with other algorithms.

Amami *et al*. (2012) performed an empirical comparison of SVM, K-Nearest Neighbour, Naive Bayes, Quadratic Bayes Normal and Nearest Mean on TIMIT vowel data for a multi-class recognition problem. SVM using RBF kernel achieved the best performance amongst the different classifiers evaluated.

Yasin and Abuhasan (2016) utilized five classification algorithms using Random Forest, J48, Naive Bayes, SVM and Multi-Layer Perceptron (MLP) for phishing email detection. Random Forest gives the best result followed by J48.

Agarwal and Kumar (2016) performs spam filtering with SVM using different kernel functions (linear, polynomial, RBF, sigmoid) and different parameters (C-SVC, NU-SVC). The best result is achieved with linear kernels on C-SVC.

Hong *et al*.(2017) compare SVM kernel functions for landslide susceptibility mapping. The results of the study revealed that SVM-RBF is the most suitable for landslide susceptibility assessment.

According to Kotsiantis (2007) there is no a single learning algorithm that can evenly do better than other algorithms over all datasets. Whenever we are confronted with the decision of selecting the precise algorithm for a classification problem the easiest way is to approximate the preciseness of the candidate algorithms on the problem and choose the one that is more precise (Kotsiantis, 2007).

From the works reviewed, it could be clearly seen that SVM is leading in terms of accuracy followed by Random Forest. This study will therefore test the prediction accuracy of the SVM kernels (linear, polynomial, RBF, sigmoid) using C-SVC and NU-SVC parameters, SVM using Taguchi Method and SVM using grid search in the prediction module to predict the APT attack.

IV. Related Work

Sharma *et al*.(2016) proposed a distributed framework architecture for the detection of APT. The work focuses on providing intrusion detection framework especially for APT attack detection. The aim of the work is to offer a new intrusion detection system that processes the network traffic and that is intelligent enough to identify an APT attack. The recognition of APT attack depends on the relationship between the events that are generated by different

classifier methods. The study designed a new framework architecture for intrusion detection system of network traffic for APT attacks in a distributed environment. The intrusion detection process was performed in a distributed environment in the trusted platform module where it stays hidden from the attackers. Initially network traffic packets to identify all possible strategies that could be utilized as a part of an APT attack cycle are collected, processed and analyzed using four different recognition methods which are independent of each other. The outputs of these classifier methods are then submitted to the next stage which is the event correlation phase. The event correlation modules takes all events provided by the outputs of all detection classifier methods as an input and correlates all of them individually as indicated by the principles specified by the system admin to raise alert on APT attack discovery. The outputs is then submitted to the next stage which is the voting stage. In the voting stage voting service analyzes and determines final result based on the information provided by event correlation for the different methods. The rationale behind the voting techniques is to lessen the rate of false positives and enhances the accuracy of the detection. Four classification methods used for the detection are genetic programming, classification and regression trees, support vector machines and dynamic bayesian game model. The proposed methodology was evaluated with results from the individual classifiers. The study did not validate the proposed approach by comparing it against other APT detection works.

Moya *et al*.(2017) proposed the use of expert knowledge and data analysis to detect APT. The accuracy of the proposed model was measured with several samples using bayesian techniques, decision trees and artificial neural networks. Decision trees shows better fitness. Validation tests was performed over all the samples and then selected some variables to be assessed: accuracy of the model created with decision trees, improvement over the trivial model, sensitivity to harmful behaviour, resistance accuracy of the model, resistance improvement over trivial model and resistance sensitivity to harmful behaviour. To choose the best possible proportion of activity logs the study developed descriptive analysis over each sample with the values of the variable described in the study (boxplots and arithmetic mean). The sample with the highest mean points to the most adequate model. After the analysis, the final system is run with the best sample and is able to alert of log registers that might be related with APTs. The results of the analysis revealed that ID3-C4.5 decision tree provides better accuracies and errors than Naive Bayes and probabilistic neural network. This led to the selection of the decision tree to detect anomalous behaviors in the network activity (Moya *et al*., 2017). There was no evaluation of the proposed methodology to show how effective it is against other proposals.

Ghafir *et al*.(2018) proposed MLAPT for the detection of APT using machine learning correlation analysis. The proposed methodology comprises of three parts: threat detection, alert correlation and attack prediction. The threat detection uses eight methods to detect various steps used in a multi-step APT attack. This methods are disguised executable file detection (DeFD), malicious file hash detection (MFHD), malicious domain name detection (MDND), malicious IP address detection (MIPD), malicious SSL certificate detection (MSSLD), domain flux detection (DFD), scan detection (SD) and Tor connection detection (TorCD). The outputs of this phase generates events from the detection methods used. The events correlation phase correlates the events produced in the first phase with one APT attack scenario. The event correlation phase consists of three steps: alert filter (AF), alerts clustering (AC) and correlation indexing (CI). The aim of the event correlation phase is to reduce the false positive rate of the detection system. The attack prediction phase implements a machine learning based prediction module based on a historical record of the monitored network. The prediction module employs four classification algorithms thus: decision tree learning, support vector machine, k-nearest neighbours and ensemble learning. No reason was specified for choosing those classification algorithms. SVM has the highest degree of prediction accuracy and recommended to be used by the network security team to predict APT. The attack prediction module is aimed to help the network security team to predict APT attack. Other limitations of the proposal is that the attack detections modules did not adequately captured all the attack techniques used during APT. So there is a room for improvement with regards to that.

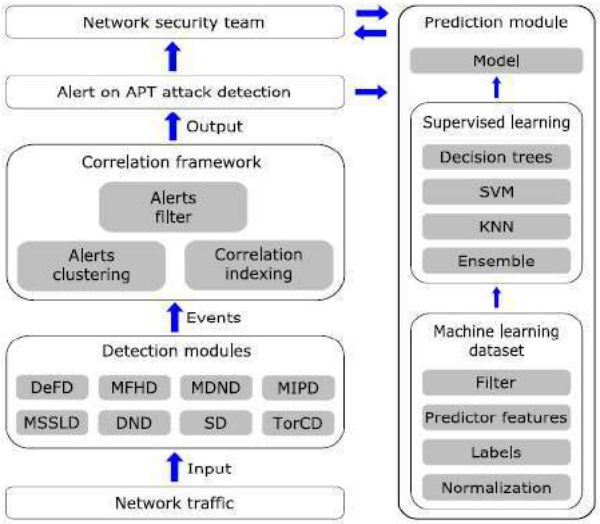

Figure 1: The Architecture of MLAPT (Ghafir *et al*., 2018)

This research work seeks to improve the work of Ghafir *et al*., (2018) by increasing the attack detection methodologies. The attack detection methodologies that will be added have been presented side by side with the proposed attack detection methodologies of Ghafir *et al*., (2018) in table 3. In the prediction module, Ghafir *et al*., (2018) utilizes four classification algorithms Decision trees, SVM, KNN and Ensemble. The classification algorithms that yields the highest detection accuracy is SVM using the linear kernel. This work seeks to improve upon the prediction accuracy by proposing to compare the prediction results of the different kernels of SVM (linear, polynomial, RBF, sigmoid), SVM using Taguchi Method and SVM using grid search. The different types of SVM have been depicted in figure 6 as SL(SVM linear), SP(SVM polynomial), SR(SVM RBF), SS(SVM sigmoid), STM(SVM using Taguchi Method) and SGS(SVM using grid search).

V. Methodology

1. Aim and Objectives

The aim of this research is to develop an effective framework for the detection and prediction of advanced persistent threat (APT) based on the cyber kill chain (CKC).

The specific objectives of this research are to:

a. design an effective APT detection and prediction framework
b. develop attacks detection modules for the attacks in the cyber kill chain stages
c. develop correlation module for the APT attacks detection
d. develop an APT prediction module
e. evaluate the effectiveness of the proposed APT detection framework with that of Ghafir *et al*., (2018)

2. System Architecture

The system architecture of the proposed APT detection framework based on the CKC is presented in figure 6.

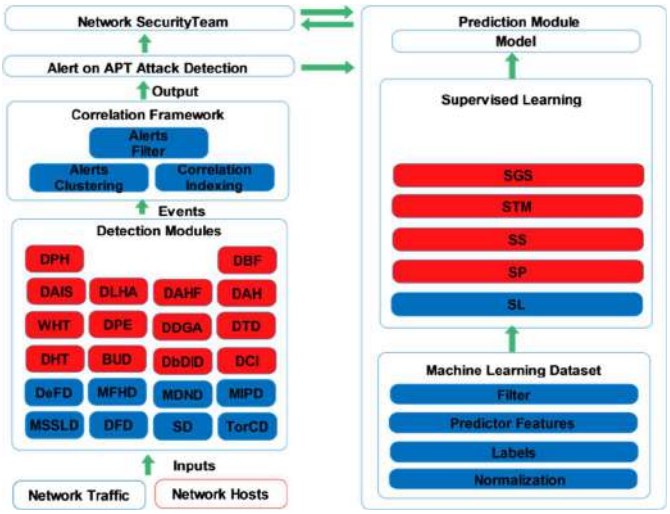

Figure 2: Proposed APT Detection System Architecture

The proposed framework is an extension of the work of Ghafir *et al*. (2018). Network traffic and network hosts will be monitored by the attack detection modules for attack. Fourteen detection methodologies have been added to the proposed work of Ghafir *et al*. (2018) to detect the several attacks employed during an APT campaign. Table 3 compares the attack detection methodologies proposed by Ghafir *et al.* (2018) which are based on the APT attack lifecycle against the attack detection methodologies proposed by this study to build on the work of Ghafir *et al*. (2018) and which are based on the CKC. These attack detection modules (methodologies) generates events which will then be fed to the correlation framework. The correlation framework correlates the events generated from the attack detection modules to one APT attack scenario. The rationale behind using the correlation framework is to lower the false positive rate of the system (Ghafir *et al*. 2018). The output from the correlation framework is an alert (event) on APT detection that will be channeled to the network security team who will utilize it to predict APT attack. The prediction module, predicts whether the event generated by the correlation framework will grow to a full APT attack scenario in the future based on the attribute of the event generated by the correlation framework. This will enable the network security team to perform more analysis on the corresponding two suspicious events (the event from the correlation framework and the full APT attack scenario) and stop the attack before it grows to a full APT (Ghafir *et al*. 2018).

Table 3: Comparison of the Proposed Cyber Kill Chain based APT Detection Methodology against Ghafir *et al*. (2018)

Cyber Kill Chain	APT Attack Lifecycle (Ghafir *et al*., 2018)	Methods of Detection Proposed by (Ghafir *et al*., 2018)	Methods of Detection Proposed by this study
Reconnaissance	Intelligence Gathering	None	Use of DNS Honey Tokens (DHT), Detection of Access to robots.txt Files, Detection of Access to Invisible Links, Detection of access to HTML Honey tokens (collectively referred to Detection of Web server Honey Tokens (WHT)) (Kollitris, 2015)

Weaponization		None	None
Delivery	Initial Compromise (Point of Entry)	Malicious Domain Name Detection (MDND), Disguised exe File Detection (DeFD), Malicious File Hash Detection (MFHD)	Malicious Domain Name Detection (MDND) (Ghafir *et al*., 2018), Disguised Exe File Detection (DeFD) (Ghafir *et al*., 2018), Malicious File Hash Detection (MFHD) (Ghafir *et al*., 2018), Bad USB Detection (BUD), Drive by Downloads/Install Detection (DbDID)
Exploitation	Initial Compromise (Point of Entry)		
Installation	Initial Compromise (Point of Entry)		Detection of Code Injection (DCI), Detection of API Hooking (DAH), Detection of Privilege Escalation (DPE)
Command & Control	Command & Control	Malicious SSL Detection (MSSLD), Malicious IP Address Detection (MIPD), Domain Flux Detection (DFD).	Detection of a Connection to a TOR (TorCD), Detection of a DGA (DDGA), Malicious SSL Certificate Detection (MSSLD) (Ghafir *et al*., 2018), Malicious IP Address Detection (MIPD) (Ghafir *et al*., 2018), Domain Flux Detection (DFD) (Ghafir *et al*., 2018), DNS Tunneling Detection (DTD).
Act on Objectives	Lateral Movement, Asset/Data Discovery and Data Exfiltration		Detection of Access to Internet Sink (DAIS) (Kollitris, 2015), Detection of Logging to Honey Account (DLHA) (Kollitris, 2015), Detection of Access to Honey Files (DAHF) (Kollitris, 2015), Tor Connection Detection (TorCD) (Ghafir *et al*., 2018), Detection of Pass the Hash (DPH), Detection of Brute Force Attack (DBF).

Figure 5 presents the system architecture for the proposed APT detection framework. The detection results of each of the method in the detection module will generate an event which will serve as an input to the correlation framework. The correlation framework aim is to find events that are related and belonging to one APT attack situation (Ghafir *et al*., 2018). To find out the probability of the early alerts leading to a complete APT attack, a machine learning based prediction module will be used in the final stage (Ghafir *et al*., 2018).

VI. Conclusion

The result of this research work will be a framework that will effectively detect APT. The attack detection modules proposed in the study will be developed and evaluated against recent study. The events generated from the attack detection modules will be fed to the correlation framework and subsequently the various SVM kernels will be used to develop a model to predict APT attack. The model that supersedes in accuracy will be recommended for use by the network defense team. The APT prediction accuracy of the proposed framework will be evaluated against the work of Ghafir *et al*. (2018).

An effective framework capable of effectively detecting APT based on cyber kill chain has been proposed. The proposed study builds upon the work of Ghafir *et al*. (2018). This will be achieved by increasing the number of attack detection modules in the proposed framework and the use of several SVM kernels have also been proposed to predict APT attack.

REFERENCES

Agarwal, D. K., & Kumar, R. (2016). Spam Filtering using SVM with different Kernel Functions. International Journal of Computer Applications, 136(5), 16-23. Retrieved from https://www.ijcaonline.org/research/volume136/number5/agarwal-2016-ijca-908395.pdf

Aldridge, J. (2016). Remediating Targeted-threat Intrusions. Fire Eye. Retrieved from https://www2.fireeye.com/rs/848-DID-242/images/WP-Remediating-Intrusions.pdf?mkt_tok=eyJpIjoiT1RNMk1HWmxNalF3WkRBNSIsInQiOiJ6dVwvVXR0cGFZS2UzaFF1UlBsdUZ3Sjl0b2NUbVJWTWpIK3dLS04yazUxcFowN0dJQU9rUlM4ZnF2cGRsMStDb2paU3o5RzFyXC9LdnZyQVpWS29EbUdNaE1ia0p2QXFmQn

Amami, R., Ayed, D. B., & Ellouze, N. (2012). An Empirical Comparison of SVM and Some Supervised Learning Algorithms for Vowel Recognition. International Journal of Intelligent Information Processing (IJIIP), 3(1.6), CoRR. doi:doi: 10.4156/IJIIP

Angle, M. G., Madnick, S., & Kirtley, J. (2017). Identifying and Mitigating Cyber Attacks that Could Cause Physical Damage to Industrial Control Systems . IEEE Power and Energy Technology Systems Journal, 1-10 .

Baksi, R. P., & Upadhyaya, S. J. (2017). Kidemonas: The Silent Guardian. SKM'17, (pp. 1-6). Tampa, FL, USA

Dell SecureWorks. (2012). Lifecycle of an Advanced Persistent Threat. Dell. Retrieved from http://www.redteamusa.com/PDF/Lifecycle%20of%20an%20Advanced%20Persistent%20Threat.pdf

ENISA. (2018). ENISA Threat Landscape Report 2017: 15 Top Cyber-Threats and Trends. Heraklion, Greece: ENISA. doi:DOI 10.2824/967192

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification. Informatica, 249-268.

Ghafir, I., & Prenosil, V. (2016). Proposed Approach for Targeted Attacks Detection. In H. Sulaiman, M. Othman, M. Othman, Y. Rahim, & N. Pee, Lecture Notes in Electrical Engineering, (Vol. 362, pp. 73-80). Springer, Cham.

Ghafir, I., Hammoudeh, M., Prenosil, V., Han, L., Hegarty, R., Rabie, K., & Aparicio-Navarro, F. J. (2018). Detection of Advanced Persistent Threat using Machine-Learning Correlation Analysis. Future Generation Computer Systems, 89, 349-359. doi:https://doi.org/10.1016/j.future.2018.06.055

Herløw, L. (2015). Detection and Prevention of Advanced Persistent Threats: Evaluating and Testing APT Lifecycle Models Using Real World Examples and Preventing Attacks through the Use of Mitigation Strategies and Current Best Practices. Denmark: DTU Compute: Department of Applied Mathematics and Computer Science.

Hong, H., Pradhan, B., Bui, D. T., Xu, C., Yousseff, A. M., & Chen, W. (2017). Comparison of Four Kernel Functions used in Support Vector Machines for Landslide Susceptibility Mapping: A Case Study at Suichuan Area (China). Geomatic, Natural Hazards and Risk, 8(2), 544–569. doi:http://dx.doi.org/10.1080/19475705.2016.1250112

Hutchins, E. M., Cloppert, M. J., & Amin, R. M. (2011). Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains. 6th Annual International Conference on Information Warfare and Security (pp. 1 - 14). Washington DC: Academic Conferences and Publishing International.

Kollitris, N. V. (2015). Detecting Advanced Persistent Threats through Deception Techniques. Greece: Information Security and Critical Infrastructure Protection (INFOSEC) Laboratory.

Mandiant. (2004). APT1: Exposing One of China's Cyber Espionage Units. Mandiant.

Martin, L. (2015). Gaining the Advantage: Applying Cyber Kill Chain Methodology to Network Defense. Lockheed Martin Corporation.

Mezghani, B. A., Boujelbene, Z., & Ellouze, N. (2010). Evaluation of SVM Kernels and Conventional Machine Learning. International Journal of Hybrid Information Technology, 3(3), 23-34. Retrieved from http://www.sersc.org/journals/IJHIT/vol3_no3_2010/3.pdf

Mitre. (2014). Search Results. Retrieved May 5, 2018, from Common Vulnerabilities and Exposures: https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=CVE+-2014-3306

Moya, J. R., García, N. D., Díaz, R. Á., & Tamargo, J. L. (2017). Expert Knowledge and Data Analysis for Detecting Advanced Persistent Threats. *Open Mathematics, 15*(1), 1108-1122. doi:https://doi.org/10.1515/math-2017-0094

NIS Platform. (2014). State of the Art of Secure ICT Landscape. NIS. Retrieved from https://resilience.enisa.europa.eu/nis-platform/shared-documents/wg3-documents/state-of-the-art-of-the-secure-ict-landscape/at_download/file

Oprea, A., Li, Z., Yen, T.-F., Chin, S., & Alrwais, S. (2015). Detection of Early-Stage Enterprise Infection by Mining Large-Scale Log Data. 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (pp. 45-56). Rio de Janeiro, Brazil: IEEE. doi: 10.1109/DSN.2015.14

Rot, A., & Olszewski, B. (2017). Advanced Persistent Threats Attacks in Cyberspace Threats, Vulnerabilities, Methods of Protection. Federated Conference on Computer Science and Information Systems. 12, pp. 113–117. Prague, Czech Republic: ACSIS. doi:DOI: 10.15439/2017F488

Sharma, P. K., Moon, S. Y., Moon, D., & Park, J. H. (2017). DFA-AD: A Distributed Framework Architecture for the Detection of Advanced Persistent Threats. *Cluster Computing, 20*(1), 597–609. doi:https://doi.org/10.1007/s10586-016-0716-0

Wei-Chih, H., & Yu, T.-Y. (2009). E-mail Spam Filtering Using Support Vector Machines with Selection of Kernel Function Parameters. 2009 Fourth International Conference on Innovative Computing, Information and Control (pp. 764-767). Kaohsiung, Taiwan: IEEE. doi:DOI: 10.1109/ICICIC.2009.184

Yadav, T., & Mallari, R. A. (2016). Technical Aspects of Cyber Kill Chain. 1 - 7.

Yasin, A., & Abuhasan, A. (2016). An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security & Its Applications (IJNSA), 8(4), 55-72. doi:DOI: 10.5121/ijnsa.2016.8405

Scientific and Practical Cyber Security Journal (SPCSJ) 3(3): 12 - 15 ISSN 2587-4667 Scientific Cyber Security Association (SCSA)

TextFort: An Efficient Hybrid Short Message Service Encryption Scheme for Mobile Devices

Faisal A. Garba
Department of Computer Science, Ahmadu Bello University, Zaria, Nigeria.
Department of Computer Science Education, Sa'adatu Rimi College of Education, Kano, Nigeria.
Cyberforce Pentest Ltd, Kano, Nigeria.

[1]Prof. Afolayan A. Obiniyi and [1,2]Prof. Saleh E. Abdullahi
[1]Department of Computer Science, Ahmadu Bello University, Zaria, Nigeria
[2]Department of Computer Science, Nigeria Turkish Nile University, Abuja, Nigeria

ABSTRACT

Mobile device users prefer to preserve the privacy of their SMS communication from mass government surveillance and other adversaries using mobile device SMS encryption solutions. The mobile devices in use however, are highly constrained in terms of memory, power and computing capability to utilize the current SMS encryption solutions. There is a room for improvement in term of the speed efficiency of the SMS encryption schemes proposed for use on mobile devices. This paper propose an end-to-end SMS encryption scheme ideal for use on mobile devices using a hybrid combination of cryptographic algorithms: Blowfish symmetric encryption algorithm, Elliptic Curve Diffie Hellman (ECDH) and Elliptic Curve Digital Signature Algorithm (ECDSA). The proposed scheme will be implemented using Java programming language to develop SMS encrypting Android application. The time taken for the proposed SMS cryptographic operations will be measured on five different Android mobile devices with varying processor speed and will be compared with a related work to evaluate the proposed scheme's speed. The cryptographic operations to be measured are the time taken for encryption and decryption and key generation.

Keywords: encryption, SMS, Blowfish, ECDH-ECDSA, cryptography, security, privacy.

Introduction

According to Susanto and Godwin (2010), using SMS over voice calls is the choice of majority of mobile users since it is cheap and trivial. Banks use SMS to send one time password (OTP), bank account details, exchange of security codes. However, this sensitive data could easily be hacked on their way to the intended recipient or send to the wrong recipient. Cryptography could be used to secure the transmission of SMS. Cryptography comes in three forms: symmetric key cryptography (secret key cryptography), asymmetric key cryptography (public key cryptography) and cryptographic hash functions. To ensure privacy of data symmetric key cryptography and asymmetric cryptography are used while cryptographic hash functions are used to preserve integrity. Each of these

forms of encryption has it's weakness as well as strength. To eliminate the weakness and gain the strength, the three forms of encryption are joined together to form a hybrid encryption scheme (Kuppuswamy and Al-Khalidi, 2014). These strengths are speed, security and the elimination of the key distribution problem.

Methodology

An efficient end-to-end SMS hybrid encryption scheme using a combination of cryptographic algorithms: Elliptic Curve Diffie Hellman (ECDH) which is a key negotiation algorithm, and also an asymmetric encryption algorithm with Elliptic Curve Digital Signature Algorithm (ECDSA) and Blowfish encryption algorithms which is a symmetric encryption algorithm. The SMS encryption scheme will be implement using Java programming language to develop SMS encrypting Android app. The target Android version is Android 4.0 (Ice Cream Sandwich). The work of Azaim *et al*. (2016) will also be implemented using Java programming language to develop SMS encrypting Android app. The target Android version is also Android 4.0 (Ice Cream Sandwich). The encryption and decryption rate of the proposed scheme will be compared with the work of Azaim *et al*. (2016) against the CPU clock rate of 5 android mobile devices. Figure 1 is the proposed efficient SMS hybrid encryption scheme. In the system architecture we have two entities Aisha and Buhari trying to exchange SMS. Any of the entities can initiate the communication process. ECDH-ECDSA is being used to generate a shared secret which serve as a temporary key. To encrypt and exchange the permanent Blowfish key, the temporary key is used alongside Blowfish encryption algorithm. The permanent Blowfish key, can now be used with the Blowfish encryption algorithm to exchange SMS. Other entities in the architecture are the database which is used in storing the keys as well as the SMS messages and the mobile network operator.

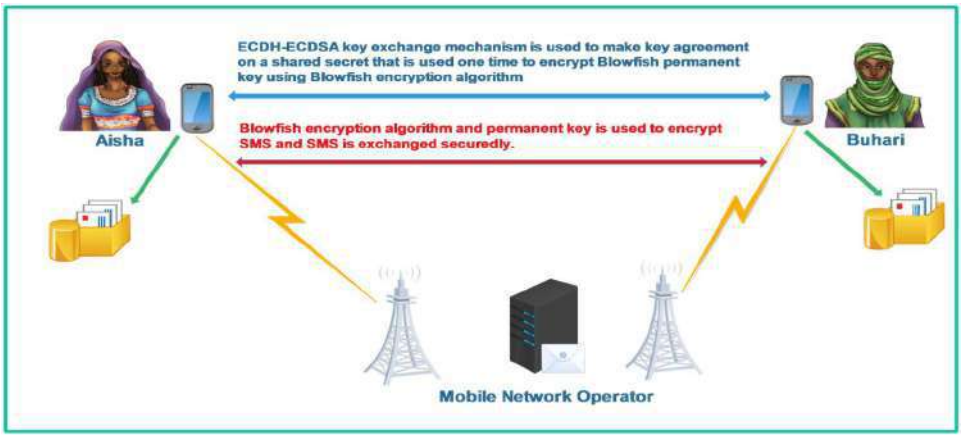

Figure 1: Proposed Efficient SMS Hybrid Encryption Scheme Architecture

Proposed Scheme's Pseudocode

Step 1: Aisha selects an integer X_A to serve as her private key and go on to generate $Y_A = X_A \times G$ to serve as her public key.

Step 2: Aisha sends the public key Y_A to Buhari signed with her ECDSA private key.

Step 3: Buhari verifies that the public key Y_A is from Aisha by using Aisha's ECDSA public key and then picks an integer X_B to be his private key and calculate his public key thus, $Y_B = X_B \times G$.

Step 4: Buhari sends the public key Y_B to Aisha signed with his ECDSA private key.

Step 5: Aisha verifies that the public key Y_B is from Buhari using Buhari's ECDSA public key, Aisha computes her secret shared session key thus $K = X_A \times Y_B$.

Step 6: Buhari also calculates his shared session key thus $K = X_B \times Y_A$.

Step 7: Aisha uses Blowfish encryption algorithm and K to encypt permanent Blowfish key K' and send it to Buhari.

Step 8: Buhari accept the encrypted message and decrypt it with his shared secret key generated in step 1 to recover the permanent Blowfish key.

Step 9: Aisha and Buhari can now exchange SMS encrypted with Blowfish encryption algorithm

Results & Discussion

The result of this research work will be compared with the work of Azaim *et al*. (2016) to evaluate the speed of the proposed efficient end to end SMS encryption scheme.

Conclusion

The result of this research work will be an efficient end-to-end SMS encryption scheme ideal for use on mobile devices which shall advance the state of the art in SMS encryption techniques on mobile devices.

REFERENCES

Azaim, M. H., Sudiharto, D. W., & Jadied, E. M. (2016). Design and Implementation of Encrypted SMS on Android Smartphone Combining ECDSA - ECDH and AES. *The 2016 Asia Pacific Conference on Multimedia and Broadcasting (APMediaCast),* 18-23.

Kuppuswamy, P., & Al-Khalidi, S. Q. (2014). Hybrid Encryption/Decryption Technique Using New Public Key

and Symmetric Key Algorithm. *International Journal of Information and Computer Security, 6*(4), 372-382.

Susanto, T. D., & Goodwin, R. (2010). Factors Influencing Citizen Adoption of SMS-Based e-Government Services. *Electronic Journal of e-Government, 8*(1), 55 - 71.

Scientific and Practical Cyber Security Journal (SPCSJ) 3(3): 16 - 17 ISSN 2587-4667 Scientific Cyber Security Association (SCSA)

Beyond The Vault: Evaluating authentication controls within secure storage mobile applications.

Gionathan Armando Reale
Honorary Security Team Member of Stratus5, USA.

ABSTRACT

With the dramatic increase in smartphone usage within the last decade and an increase in privacy demands in modern post-Snowden world, many users strive for a safe and convenient way to store personal data and media. In order to fill these demands, software developers have filled the market with mobile applications designed to safely store sensitive content. Within this article I evaluate the efficiency of authentication controls from a sample of secure storage mobile applications.

KEYWORDS: Infosec, Mobile Security, Vulnerabilty, Pentesting, Authentication, Bruteforce

Introduction

It is reported that 49% of smartphone users have sent or received intimate content [1], It is no wonder that many people opt to install mobile applications in the hopes of controlling and securing personal content which, if left insecure, could cause them and others significant harm or embarrassment. The potential problem arises when the mobile applications which are trusted to secure data and media are not regularly tested and are poorly built.

Method

The most popular approach to preventing unauthorised access within secure storage mobile applications is to use a form of password/pattern/fingerprint or PIN based authentication. I set out to test, with a small sample of secure storage mobile applications, how well these controls were implemented.

My sample consisted of ten secure storage mobile applications. All applications were tested on an Android 8.0.0 smartphone device and sourced from the Google Play Store[2]. The testing consisted of reviewing the options available to users to protect against unauthorised access, attempting to trigger and detect anti-bruteforce mechanisms, as well as evaluating risk, based on the outcome of the two previous factors.

When intending to trigger and detect anti-bruteforce mechanisms I manually submitted failed login attempts over a period of ten to twenty minutes. I used the documentation and settings within the applications to review their authentication options, and when evaluating risk I took into account the protective mechanisms (or lack thereof) I had detected and the potential ease a motivated attacker would have to bypass authentication given the settings and controls in place.

Results

Upon testing, I discovered that only two out of the ten mobile applications in my sample had anti-bruteforce protection and adequate controls for strong authentication. The remaining mobile applications (8/10) in my sample did not have anti-bruteforce protection. Three mobile applications offered users the option to set adequate authentication controls, such as the ability to set a long complicated passphrase or ability to enable multi-factor authentication. The majority (7/10) of mobile applications tested did not offer secure authentication options to users. Instead they opted for a less secure four digit PIN, which would, given the correct circumstances, allow a motivated attacker to gain unauthorised access.

Limitations

The project as a whole has significant limitations, the first being that the sample size I used for testing was rather small. There are numerous mobile applications offering secure storage solutions. It would be wrong to assume any firm conclusions could be made based on this research alone.

The next limitation was that risk suggested by my data may not reflect the actual risk users face by using a particular application mentioned in this article. The actual risk would depend on other factors and the threat model of the user, for example: if a user leaves their phone unattended in public and unlocked, they would be at more risk than a user who kept their phone encrypted and within sight at all times.

Another limitation is that all the mobile applications were tested within an Android environment. It may be possible that other versions of the application for other platforms may have been more secure.

The final limitation was that I only tested the product within the «FREE» version, it is possible that upon payment some of the mobile applications tested may have been a lot more secure, allowing users to pay more if their threat model required it. Other limitations may exist.

Discussion

Privacy is an important value within today's society, given the percentage of smartphone users worldwide[1] and the use of smartphones to store, send and receive sensitive content. It is important that solutions offered as secure are reviewed and evaluated on a consistent basis. My findings suggest that some of the secure storage mobile applications simply may not have sufficient security that could stop a motivated and persistent attacker. The limited size of the sample group limits the overall significance of the results, but this data implies that poor authentication controls within this type of mobile application could be a widespread issue.

REFERENCES

[1] Mcafee. Feb 2014. [Online]
https://securingtomorrow.mcafee.com/consumer/identity-protection/love-and-tech/?culture=en-us&affid=0&cid=140623
[2] Google Play Store [Online]
play.google.com

Scientific and Practical Cyber Security Journal (SPCSJ) 3(3): 18 - 24 ISSN 2587-4667 Scientific Cyber Security Association (SCSA)

Evaluation of the Level of Cyber Security of Information

Khoroshko Vladimir, National Aviation University of Kiev, Doctor in Technical Sciences, Professor Kiev, Ukraine, professor

Mykola Brailovskyi, Taras Shevchenko National University of Kyiv, PhD in Engineering Science, Associate Professor Kiev, Ukraine

Khokhlachova Yulia, National Aviation University of Kiev, PhD in Technical Sciences, Associate Professor Kiev, Ukraine

Ayasrah Ahmad Rasmi Ali, graduate student of the National Aviation University, Kiev, Ukraine

ABSTRACT

A comparative analysis of the concepts of "cyber threat" and "cybersecurity" is given. However, opposite these concepts may be, they are interdependent and have much in common. It is proved that the level of cyber-threat of information simultaneously characterizes the level of cybersecurity, and the quantitative indicator of this could be - the cybersecurity index. The method of calculation of cybersecurity index of is presented in the work. Mathematical modeling of the cybersecurity index of information has not only practical but also predictive value. By employing the values of variables that are included in mathematical dependencies to calculate a cybersecurity index, one can evaluate the effectiveness of implementing certain measures aimed at its dynamics. Therefore, the functional relationship between the cybersecurity index and the value of the information indicators around the information can be an instrument for an in-depth study of the cybersecurity problem of information.

Keywords: cyber security index, security rating, cyber threat

The gradual transition in the development of the human formation of the "information society" to "high-tech society" causes the evolution of approaches to security in the new conditions at different levels. The gradual transformation of citizen, society and state information security concept requires supplementing it with a new concept of cybernetic security. At the same time, there is a process of distinguishing different types of security at the geopolitical levels and understanding the role of cybernetic security at each of them. The need for awareness of the role of cybernetic security is primarily due to the intensification of international, terrorist, extremist organizations and criminal gangs, individual states that exercise cybernetic effects on citizens, society and the state in order to reach their goals.

Cybersecurity is becoming increasingly important in ensuring the national security of the developed states. The cybernetic affects are probably the most effective in achieving the objective of controlling various objects (i.e. individuals, organizations, regions, states, etc.) in the modern world. In fact, a new phenomenon has emerged in international politics: the possibility of achieving political goals, changing legitimate governments and even political, economic and spiritual subjugation of civilians without any military force. Cyberattacks and cyber impacts are evolving rapidly.

With regard to it, one can make the following conclusions concerning the changes in cyberspace and national security of states in the context of transformation of the existing and new methods of cyber-threats, cyber-impacts and their impact on cybersecurity systems, including national, regional and international.

Recently, there has been a surge in research aimed at shaping the cybersecurity of the state. This is necessary due to the need to provide an opportunity to solve the main tasks of cybersecurity in various areas of the state's activity from the common methodological positions.

On the basis of the definition and essence of cybernetics as a science of the general laws describing processes of management and information transmission in society and information systems, and security as the protection of certain objects from threats, cybersecurity can be defined as the state of security management in all spheres (social, technical, sociotechnical), which ensures its effective implementation [1,2].

In order to implement cyber security, a priority task is to ensure the counteraction towards destructive influences in this area, and this requires relevant information. That is why a powerful counteraction is required.

Cybersecurity is an integral part of information security and of each area of national security. Therefore, the state policy of ensuring high-tech cybernetic security becomes one of the most important components of national security policy, which is becoming increasingly independent.

Accordingly, implementing cybersecurity at the international, national and regional levels is one of the most important components of the national security system for any state.

In addition, a comparison of the concepts of "cybersecurity" and "protection" should be made. The protection of information by its main task means the rejection or reflection of a cyberattack on information or unauthorized access to it. Cyber security, compared to protection, is more complex and multifaceted. The information security is achieved not only through the organization of cyber defense, but also through a variety of activities in the political, economic and through other spheres of public life.

When discussing the concepts of "cyber threats" and "cybersecurity" of information, a range of questions may arise: which of these two phenomena is primary; what are their mutual relationship and influence; what are the criteria for their assessment?

It is obvious that the primary concept of "cyber security" lies in information. It is directed against the information dangers through employing the ways, methods and means of cyber security.

However, despite the semantic prominence of the concepts of cyber-threats of information, much is to be found between them.

First, both concepts of the phenomenon are purposefully arisen in the same spheres of human activity.

Secondly, cyber-threats of information and cyber security of information are created by the same subjects.

Thirdly, both cyber threats and cyber security information can be created using the same methods and tools.

Regarding the differences between cyber-information and cybersecurity, they lie in different contexts.

First and foremost, this is the difference between the cyber threats of information and cyber security in relation to the objects of activity: the object of information hazard - mastering, receiving, while the object of cybersecurity - protection, preservation, provision of conditions for the direct existence, information storage and use.

Another fundamental difference between cyber threats and cybersecurity of information is in their relationship with the objects of activity. Cybersecurity information is for its objects an external hostile factor. Cybersecurity is united with its objects by the commonality of the personal unity of goals and interests, especially in experimental situations. In spatially presented objects of cybersecurity information seems to be surrounded by a protective shell, and the cyber threat of information is aimed

at unauthorized access to it and the destruction of both this protection itself and the information itself (object).

Finally, cyber-threats of information and cyber security of information are also distinguished by the arsenal of means by which these phenomena are created in the sphere of life-type information. If the information cyber threat is, first, the means of attack and the impact on it, the cybersecurity of information, which also relies on active counteraction, should be achieved first, ways to prevent unauthorized actions and attacks on information.

The interdependence of cybersecurity information and cyber threats is unequivocal. It has several important features that greatly affect the situation around information.

First, it is a deterrent effect of cybersecurity on cyber-threat information. The cybersecurity information-suppressive information, if it is carried out mainly by one type of protection, is often temporary if it does not eliminate the root causes of a conflict or does not use integrated security systems.

Secondly, there is a stimulating effect of cyber threats on information. Any increase in the cyber-threat of information causes a certain reaction in society, which, of course, reflects the growth of efforts and strengthening of the integrated information security system.

At the same time, the questions of methodical bases for assessing the level of cybersecurity of information are very relevant. Logical methods for analyzing cybersecurity information problems are quite effective, however, they do not allow the establishment of clear functional relationships between the actions of individual factors and their combined outcomes. Therefore, the initial need is to develop a method for quantitative and qualitative analysis and to objectively determine the level of cyber security information.

When considering the concept of cyber-threat information, one can conclude that the cyber-threat of information can be estimated using an integral indicator (the level of cyber-threat information), related in a way to the degree of application of the situation and the expected scale of potential attack or influence. Turning to the question of assessing the level of cybersecurity of information, it is necessary first, to find out the essence of this assessment.

Let's make a few questions to answer.

First, can we talk about cybersecurity of information in the absence of cyber threats? Obviously yes, because the cyber security of information, in fact, is the lack of cyber threats of information. Thus, the complete lack of cyber-threat information means full cybersecurity of information.

Secondly, can we talk about cybersecurity of information in the presence of cyber-threats of information? At the same time, it is possible with a certain caveat: the higher the level of cyber security information. It could seem to be a paradox at first glance, however, the conclusion can be quite simply proved.

As mentioned before, [3,4], cybersecurity of information is achieved in two main ways:

- Prevention of the attack (threat) associated with the use of passive or active actions against the attacker, i.e. the use of various leverages to influence it in order to prevent attempts to resolve the conflict;

- Counterattack, that is, deterrence (or reflection) of attack, using certain methods and means.

The focus of attention is a potential attack on information as an event that may or may not occur, depending on the degree of attacker’s commitment in it and the effectiveness of the cybersecurity implementation system of the object of potential attack. If the attack is to be started, the ability to provide cybersecurity, however, is preserved through the ability to successfully counteract the attack. In this case, it is the event that could acquire a local, regional, nationwide or global scale, depending on the level of attack.

It is possible to build a corresponding scheme of events related to the implementation of cyber threats of information and the provision of cybersecurity information through identifying the cyber threat of information with a potential attack and its consequences, and the cyber security of information - with the successful protection (in any way) of information and the preservation of its value (figure.1).

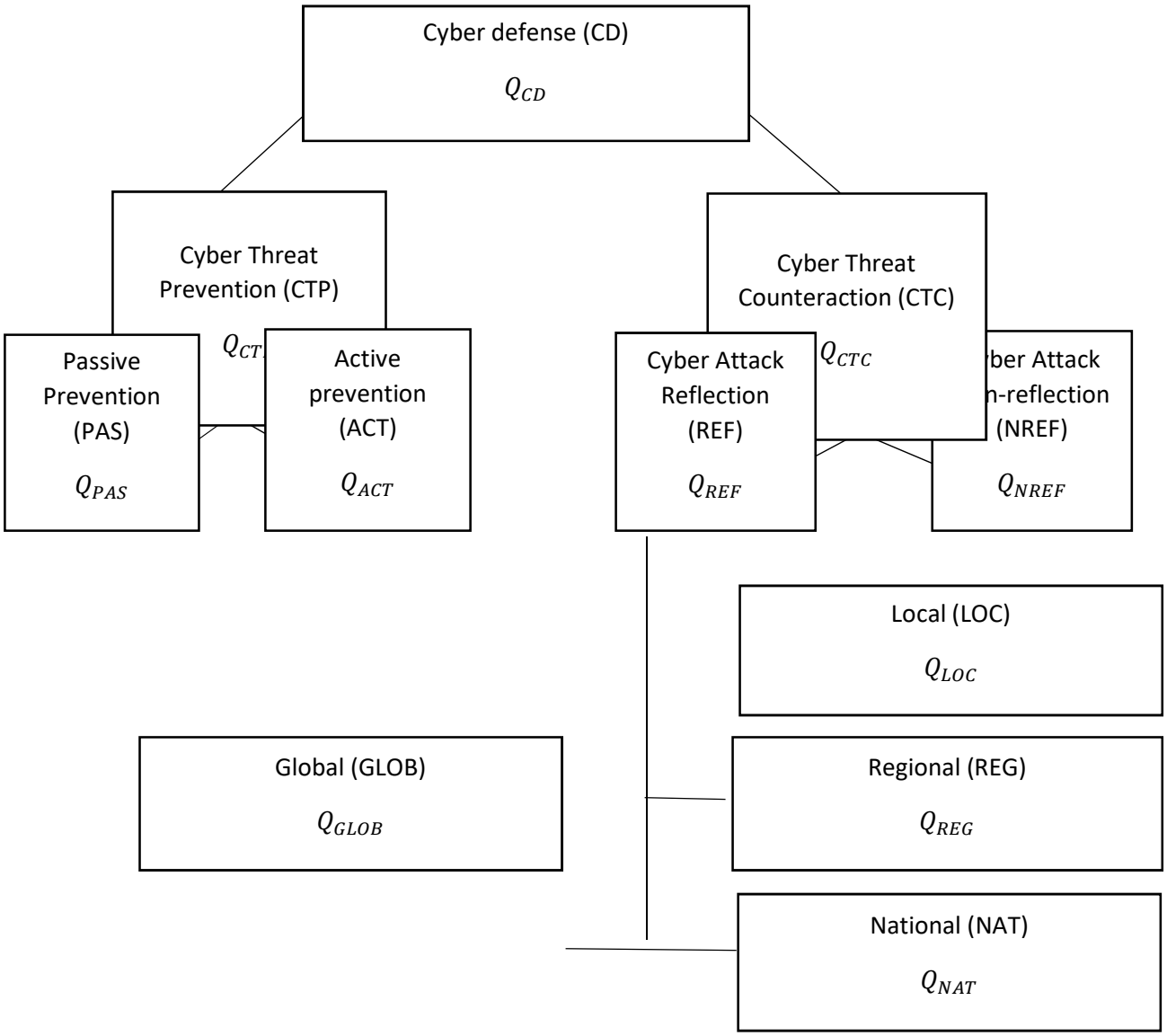

Figure 1. Scheme of events related to cyber s~~ecurity.~~

The key is the following pair of opposite events:
- Dealing with cyberattacks and countering them;
- Passive and active prevention of cyberattacks;
- Repulsing of the cyberattack and its success.

To analyze these events, a mathematical apparatus of probability theory can be used. However, an appeal to the theory of probabilities in this case requires a certain justification.

The fact is that the theory of probabilities operates, as a rule, events and phenomena that have such a property as statistical stability. The story gives thousands of examples of various conflicts, the

conditions for their occurrence, development and completion are so diverse that it is very difficult to distinguish stable statistical features. However, there are many arguments in favor of the probable approach to use in the field of cyber security.

The probability theory has many ways to determine the probability of events indirectly, because of the likelihood of other events associated with the first [5].

Significant help in solving this problem can give rise to the well-known Laplace uncertainty principle [4], the essence of which is that, in the presence of several hypotheses, none of which cannot be defeated, the probability of occurrence of the corresponding events should be considered the same. Since in this case we consider pairs of opposite events, then the starting point can serve as an axiom that for such events the sum of probabilities of their onset is equal to one.

These are the fundamental foundations for the application of the probability theory in the interest of investigating mechanisms for the emergence and termination of conflicts and attacks in the information (cyber) space.

By the way, the probabilistic approach in the analysis of conflict situations is also used in foreign studies [4].

Returning to figure 1, we note that here is the event, which is to provide cybersecurity, is marked as CD, and its probability is marked as Q_{CD}. This event can occur simultaneously with one of two other inconsistent events: with a cyberattack (CTP) with a probability of as Q_{CTP}. or with a cyberattack (CTC) with probability Q_{CTC}. At the same time, since events CTP and CTC form a complete group, then

$$Q_{CTC} = 1 - Q_{CTP} \quad (1)$$

Considering happening of the events CTP and CTC under the conditions of a specific level of cyber-threats, as only two possible hypotheses, in connection with which, with the probability Q_{CD}, according to the expression of complete probability [3] can be written

$$Q_{CD} = Q_{CTP} * Q\left(\frac{CD}{CTP}\right) + Q_{CTC} * Q(\frac{CD}{CTC}) \quad (2)$$

Or considering (1) we write

$$Q_{CD} = Q_{CTP} * Q\left(\frac{CD}{CTP}\right) + (1 - Q_{CTP}) * Q(\frac{CD}{CTC}) \quad (3)$$

where Q (CD / CTP) - conditional probability of occurrence of the event of a short-term damage in a series of offensive HQ; Q (CL / CPZ) - conditional probability of occurrence of a CF event in case of occurrence of the PCZ event.

Note that the onset of the CTP event means that the cyberattack is reflected, the cyber-threat is neutralized. In this case, the event CD is true, i.e :

$$Q\left(\frac{CD}{CTP}\right) = 1 \quad (4)$$

If a CTC event occurs, then the probability of a CD event is determined by the probability of a successful reflection of the cyberattack on the information (1), that is,

$$Q_{CD} = Q_{CD} \quad (5)$$

Then, considering (4) and (5), it is possible to write down

$$Q\left(\frac{CD}{CTC}\right) = Q_{CTP} + (1 - Q_{CTP}) * Q_{CTP} \quad (6)$$

It is important to determine the physical meaning of the value Q_{CTP}. If we denote the maximum damage to national interests and organizations as a result of external cyberattacks on information as G_{max}, then we will assume that with some probability of cyberattack reflection Q_{CTP} loss will be equal to $G_{max}(1 - Q_{CTP})$ and if $Q_{CTP} = 1$ (hypothetical case) the loss will be around zero.

Further consideration of the relationship of events is shown in Fig. 1 can be carried out according to a similar scheme. The probability of averting a cyberattack by passive action (PAS) or active containment of a cyberattack (ACT) is determined as follows:

$$Q_{CTP} = Q_{PAS} + (1 - Q_{PAS}) * Q_{ACT} \quad (7)$$

Note that the probability of deterrence or distraction of a cyberattack can be, to a certain extent, an assumption comparable to the probability of its successful reflection, since a potential attacker, when deciding on unauthorized access to information, derives, above all, from the capabilities of the party protecting the information. Thus, you can write (7) as

$$Q_{CTP} = Q_{PAS}+(1-Q_{PAS}) * Q_{REF} \quad (8)$$

As regards the reflection of a cyberattack on the information we protect, it can occur in the conditions of its local, regional, global or national nature, with the probability that the corresponding hypotheses form a complete group

$$Q_{LOC} + Q_{REF} + Q_{NAT} + Q_{GLOB} = 1 \quad (9)$$

Then

$$Q_{REF} = Q_{LOC} * Q_{REF(LOC)} + Q_{REG} * Q_{REF(REG)} + Q_{NAT} * Q_{REF(NAT)} + Q_{GLOB} * Q_{REF(GLOB)} \quad (10)$$

where $Q_{REF(LOC)}$, $Q_{REF(REG)}$, $Q_{REF(NAT)}$, $Q_{REF(GLOB)}$- is the probability of a reflection of a cyberattack of a corresponding type.

Considering (8) and (10) the level (6) is a mathematical model that reflects the degree of development of cyber threats or cyberattacks and the ability to address them by preventing or countering cyberattacks.

The conducted researches make it possible to draw the following conclusions:

1. As the main quantitative indicator of the level of cyber security, the probability of successful protection of information, preservation of its integrity in the conditions of the projected cyber-threat information may be accepted. This indicator can be determined by the cybersecurity index of information, the quantitative meaning of which makes it possible to draw certain conclusions about the level of cyber security information.

2. The methodology for calculating the cybersecurity information index should be based on the results of the assessment of the cyber-threat of information, since the schemes of events related to the provision of cybersecurity information and the implementation of cyber-threats of information are similar and are characterized by the probabilities of the same events. In addition, the basic output data for calculating the cybersecurity index can be attributed to indicators that characterize the cyber-threat of information, and their quantitative values can be determined when evaluating the latter.

Thus, the assumption that the level of information cyber threats simultaneously characterizes the level of cyber security can be considered proven.

3. Based on the interdependence of cyber-threats and cybersecurity as the main quantitative indicator of cyber-threat information, the probability of causing significant damage to the integrity and value of information as a result of cyberattacks from the outside can be accepted. This indicator should be called the index of cyber-threat information, which, in comparison with the scale of cyber-threat information, allows, in the presence of a certain criterion, to determine the level of cyber-threat information.

4. The indexes of cyber threats of information and cyber security of information are the probabilities of opposite events, which are incompatible and form a complete group, that is,

$$Q_{NCD} = 1 - Q_{CD} \quad (11)$$

Expression (2) makes it possible to argue about the possibility of applying a unified methodological approach to the evaluation of cybersecurity information and cyber security indices.

5. The quantitative assessment of the cybersecurity information index, due to the inevitable errors and the inaccuracy of the initial data, may not be of a predominant importance. More important is another: mathematical modeling of the cyber security index has not only practical but also predictive value. Operating the values of the variables included in the mathematical dependencies for calculating

the index of cybersecurity, one can evaluate the effectiveness of the implementation of certain measures aimed at its dynamics. Therefore, the functional dependence between the cybersecurity index and the value of partial information about the information environment can be an instrument for in-depth study of the cyber security problem.

References:

1. Grishchuk R.V. Fundamentals of cybernetic security / R.V. Gryshchuk, Yu.G. Danik - Zhitomir: ZNAEU; 2016 - 636 pp.

2. Danik Yu.G. National security: prevention of critical situations / Y.G. Danik, Y.I. Katkov, M.F .Pichugin - K: Ministry of Defense of Ukraine; Zhytomyr: Ruta, 2006 – 388pp.

3. Khoroshko V.O. Methodological Approach to Assessing the Level of Information Security / V.O.Khoroshko, V.S. Cherunichenko // Coll. Sci. Works of the Kyiv Taras Shevchenko National University, Vip. 14, 2008. - P. 176-181

4. Saati T.L. Mathematical methods of conflict situations / T.L. Saati - M: Sov.radio, 1997. - 304 pp.

5. Ventsel V.C. Theory of probabilities / VS Ventzel - M: Gos. issuance physical math. Lit., 1962 - 560 pp.

Scientific and Practical Cyber Security Journal (SPCSJ) 3(3): 25 - 39 ISSN 2587-4667 Scientific Cyber Security Association (SCSA)

Vehicle Traffic Flow Forecasting on Caltrans PeMS Dataset Using Machine Learning Algorithms and LSTM Networks

Jiarui Chang, Rice University (Houston, USA)
Jingwen Du, Cornell University (Ithaca, USA)
Hojin Chung, Gyeonggi Suwon International School (Suwon, Korea)

ABSTRACT:

In Intelligent transportation systems, accurate traffic flow prediction is fundamental in transportation modeling and management.Previous studies have classified prediction approaches into three categories including a time series approach with ARIMA model for finding traffic flow patterns and using those patterns for prediction, a probabilistic approach for modeling and forecasting from a probabilistic perspective, and nonparametric approaches that can perform better by handling un-deterministic and complex time series traffic datasets.This paper analyzes historical timeseries traffic data from sensors using machine learning algorithms as baseline models and designs a deep learning LSTM model to train using the historical dataset to forecast traffic flow using the trained model. The paper also compares the performance of machine learning algorithms and the deep learning model. The results show the deep learning LSTM model to outperform machine learning models.

KEYWORDS: Deep learning model, Traffic flow prediction, Caltrans PeMs dataset, LSTM model

I. Introduction:

Traffic flow prediction is a major issue in intelligent traffic systems and for public as well as private sector. It helps road users make better travel decisions and enable reduced carbon emissions and improved traffic conditions. Accurate traffic flow prediction on real-time basis providesroad users with information to optimize travel decisions and reduce travel costs, also helping traffic authorities to better mitigate congestion.

However, accurate traffic prediction is a challenging problem. Traditional traffic prediction methods include models such as autoregressive integrated moving average(ARIMA),multi-variable linear regression, and support vector regression. However, these linear models do not consider the whole range of features in traffic flow and thus do not perform optimally [8].In addition, because of stochastic and nonlinear featuresof traffic flow, parametric approaches with linearity cannot provide high traffic flow prediction performance, motivating greater attention to nonparametric approaches.[5]

Without accurate traffic flow prediction, no intelligent transportation systems could optimally perform. Previous studies have addressed this problem and classified prediction approaches into three categories including time series approach with ARIMA model for finding patterns of traffic flow and using those patterns for prediction, probabilistic approach

for modeling and forecasting from probabilistic perspective, and nonparametric approaches that performed better due to their ability to handle undeterministic and complex time series traffic datasets.

Deep learning[8] is a nonparametric approach and a type of machine learning based on neural networks. Through dependency in high-dimensional sets of variables, clear discontinuities in traffic flow emerging in large-scale networks can be captured. Deeplearning is increasingly seen as an essential tool for artificial intelligence research in areas such as traffic flow prediction. Recent years have seen the use of deep learning in traffic prediction, which this paper focuses on.

The rest of this paper is organized as follows: Section II provides a literature review. Section III discusses the materials and methods, and Section IV explains the experiments. Section V addresses the results and discussion, and Section VI concludes.

II. Literature Review:

Dai et al. [1] consider temporal patterns in traffic flow and propose a deep learningmodel for traffic flow prediction by considering DeepTrend, a deep hierarchicalneural network for predicting traffic flow based on time-variant trends. They find that DeepTrend can improve prediction performance for some popular prediction models.

Chen et al. [2] propose a novel fuzzy deep-learning approach called FDCN to better predict city traffic flow based on fuzzy theory and a deep residual network model, introducing fuzzy representation to reduce data uncertainty and proposing pretraining and fine-tuning strategies for more efficient learning of FDCN parameters. They find the proposed approach to outperform existing approaches.

Manoranjitham et al. [3] demonstrate potential benefits of deeplearning in short-term traffic flow prediction by using traffic flow data to train adeep neural network to recognize traffic patterns and provide short-term forecasts. They highlight the potential of existing GPS-based systems in improving traffic prediction accuracy and efficiency.

Jia et al. [4] introduce the deep belief network (DBN) and long short-term memory (LSTM) to better predict urban traffic flow in rainfall conditions, finding the capability of rainfall-integrated DBN and LSTMto learn traffic features and showing deep learning predictors to have better accuracy than existing predictors.

Yang et al. [5] connect long time step sequences to currenttime steps by including high-impact traffic flow values using the attentionmechanism and smoothening data beyond normal rangesfor better prediction, demonstrating the proposed predictionmodel to be better for short-term traffic flow prediction.

Kenworthy-Groen [6] reviews traffic data from three metropolitan arterialroads in Perth and compares the traditional compound traffic growth rate model to the linear traffic growth rate model, highlighting sound long-term traffic data to ensure appropriate traffic growth rate models for optimal sustainable pavement projects.

Polson and Sokolov [7] develop a deep learning model for traffic flow prediction by combining a linear model fitted using 1 regularizationand a sequence of tanh layers, demonstrating deep learning architecture to capture nonlinear spatiotemporal effects and provide accurate short-term traffic flow predictions.

Du et al. [8] propose a hybrid multimodal deep learning method for short-term traffic flow forecasting, namely one-dimensional Convolutional Neural Networks (1D CNN) and Gated Recurrent Units (GRU) with attention mechanism. They incorporate representation features of modality traffic data and find the proposed model to accurately predict complex nonlinear urban traffic flow.

Lv et al. [9] proposea novel model called LC-RNN for traffic speed prediction by considering RNN andCNN models. They also propose a network-embedded convolution structure to better incorporate topology-aware features and consider periodicity and other context factorsfor better prediction accuracy, finding the proposed LC-RNNto outperform some popular methods.

Wang et al. [10] evaluate a path-based deep learning framework for traffic speedprediction by dividing a road network into core paths and modeling each path based on the bidirectional long short-term memory neural network (Bi-LSTM NN). They find the proposed model to outperform various benchmark methods.

Zhang et al. [11] propose a method based on the cascaded artificial neuralnetwork (CANN) to predict traffic flow by incorporating actualroad network distance into the model and using real-world data from video surveillance cameras in Xiamen, China.

Xiao and Yin [12] propose a hybrid Long Short–Term Memory (LSTM) neural networkbased on the LSTM model and optimize it for various traffic environments. The prediction error of the hybrid LSTM model is lower than others but requires a slightly longer running time.

Tian et al. [13] consider a novel approach based on Long Short-TermMemory (LSTM) and multiscale temporal smoothing, demonstrating its higher accuracy in traffic flow prediction.

III. Materials & Methods:

a) Dataset:

This study uses the California department of transportation (Caltrans) dataset.In Caltrans Performance Measurement System (PeMS) dataset, data are collected on a real-time basis from individualsensors along the freeway system across all major metropolitan areas of California.

PeMS provides real-time data from over 39,000 sensors and is an Archived Data User Service (ADUS), which provides more than a decade of historical data.The dataset had 7776 instances for training.

The study focuses on analyzing historical traffic data in PeMs dataset and considers the traffic prediction model at various time interval using Long short-term memory (LSTM) networks to predict traffic flow during peak and nonpeak hours for a given city. The study also compares the LSTM model with various other baseline models.

b) Metrics:

Two popular performance indices used as metrics include the mean absolute percentage error(MAPE) and the RMS error (RMSE). MAPE measures prediction accuracy of a forecasting method typically shown in as a percentage, and RMSE is the standard deviation of residuals (prediction errors). This study's LSTM model uses Keras deep learning library, which calculates and provides a suite of standard metrics in training deep learning models. In addition, Keras defines and gives custom metrics in training deep learning models, which is useful when tracking performance measures. The Results and Discussion section provides a comparison of various algorithms with respect to metrics for a period of 5 minutes.

c) Methodology:

For experiments, a process flow is followed where the initial phase gathers required datasets from repositories (Fig.1). After relevant datasets are obtained, the data are cleaned in

preprocessing stage, followed by feature selection to identify important features that are correlated. After feature selection, training is conducted using training data, and the model will be saved if performance metrics are satisfactory. Otherwise, training is repeated using additional training data.

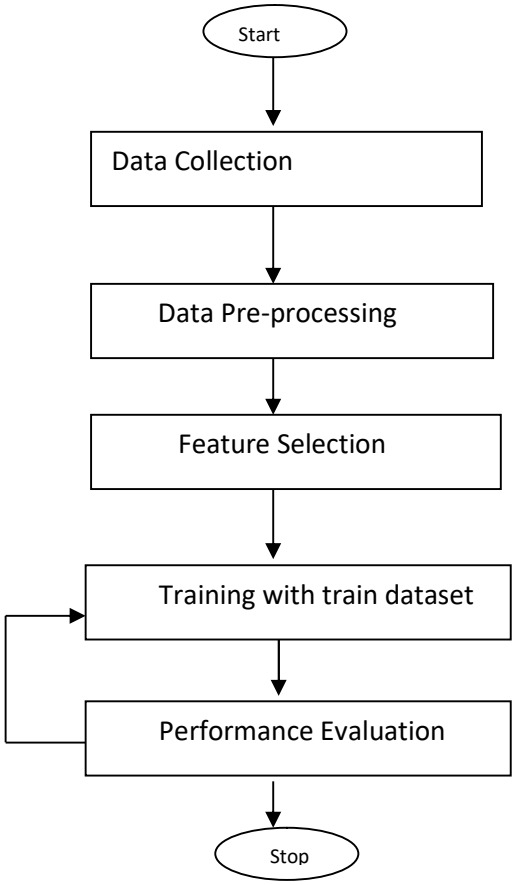

Fig.1. Methodology

IV. Experiments

The experiments are conducted in both machine learning and deep learning models. For machine learning models, WEKA tools are used to train the model with the training dataset, and performance metrics are measured using test data.Weka 3.8 is used fordata preprocessing,and regression models are trained using popular models such as Linear Regression, Multi-Layer Perceptron (MLP), RBF Network, RBF Regressor, and SMO Reg algorithms.

The experimentsuse a dataset with 7776 instances for training obtained from PeMS. As discussed in Section 4.1, the experiments are conducted using machine learning models. Section 4.2 discusses experiments using the deep learning model. For LSTM models, 5

different cases with increasing numbers of epochs are considered with reduced errors and increased performance.

a) Machine Learning Models

(i) Linear Regression

Linear regression is used to determine the linear relationship between the target and one or more predictors. Simple linear regression is useful for determining relationships between two continuous variables. One is the predictor or independent variable, and the other is the response or dependent variable. Here statistical, not deterministic, relationships are focused on. Relationships between two variables are considered deterministic if one variable is accurately expressed by the other. A statistical relationship is not accurate in determining the relationship between two variables. Here the core idea is to obtain the best-filling line, that is, the one for which the total prediction error (all data points) is the smallest. Error is the distance between a point to the regression line

(ii) Multi-Layer Perceptron

The multilayer perceptron (MLP) is a class of feedforward artificial neural networks. The MLP consists of at least three layers of nodes: input, hidden, and output. Except for input nodes, each node is a neuron using a nonlinear activation function. The MLP uses a supervised learning technique called back propagation for training, and its multiple layers and nonlinear activation distinguish the MLP from the linear perceptron. It can distinguish data that are not linearly separable.

(iii) RBF Network

The RBF network is an artificial neural network with input, hidden, and output layers. The hidden layer includes hidden neurons whose activation functionis a Gaussian function. The hidden layer generates a signal corresponding to an input vector in the input layer, and the network generates a response corresponding to the signal.

(iv) RBF Regressor

The RBF network trains hidden layers in an unsupervised manner, and RBFRegressor and RBFClassifier are fully supervised. RBFNetwork implements a normalized Gaussian radial basis function network and uses the k-means clustering algorithm for basis functions, also learning either logistic regression (discrete class problems) or linear regression (numeric class problems). Symmetric multivariate Gaussians are fit to data from each cluster, and if the

class is nominal, it uses a given number of clusters per class. RBFRegressor implements Gaussian radial basis function networks for regression, trained fully supervised using WEKA Optimization class by minimizing squared error using BFGS. Here it is possible to use conjugate gradient descent instead of BFGS updates, which is faster with many parameters, and normalized basis functions instead of unnormalized ones. RBFClassifier is the equivalent of RBFRegressor in classification problems.

(v) SMO Regressor

The sequential minimal optimization algorithm (SMO) is effective in training support vector machines (SVMs) for classification defined on sparse datasets. SMO differs from many other SVM algorithms in that it requires no quadratic programming solver. SMOreg implements the support vector machine for regression, and parameters can be learned using various algorithms. The algorithm is selected by setting the RegOptimizer.

b) Deep Learning Model

LSTM Networks

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) in deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections and can not only process single data points but also whole data sequences. LSTM networks are optimal for data collected over a time period at regular intervals, namely timeseries data.

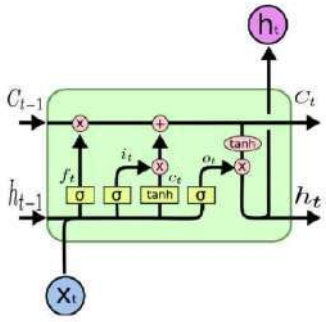

Fig.2. LSTM cell

A common LSTM unit is composed of a cell and input, output, and forget gates. The cell remembers values over arbitrary time intervals, and these three gates regulate information flow across the cell.Fig. 2 shows a simple LSTM cell.

LSTM networks are optimal for classifying, processing and making predictions using time series data since there may be lags of unknown duration between important events in time series. LSTMs can address exploding and vanishing gradient problems that can be encountered when training traditional RNNs. Relative lack of sensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov, and other sequence learning methods.

V. Results & Discussion

a) Linear Regression

Metric	1-step-ahead	2-step-ahead	3-step-ahead	4-step-ahead	5-step-ahead	6-step-ahead	7-step-ahead	8-step-ahead	9-step-ahead	10-step-ahead
MAE	7.6093	8.5389	9.6568	10.7303	11.7833	12.8264	13.7602	14.7962	15.8355	16.7475
RAE	90.444	101.481	104.5887	105.7303	105.3203	105.5449	104.8005	106.0831	107.0088	105.768
RMSE	10.3206	11.6506	13.1329	14.5647	15.9731	17.3717	18.6284	19.9162	21.1754	22.3
MSE	106.5142	135.7361	172.4725	212.1391	255.1391	301.7745	347.0184	396.6546	448.3987	497.2889

Table 1. Linear Regression metrics

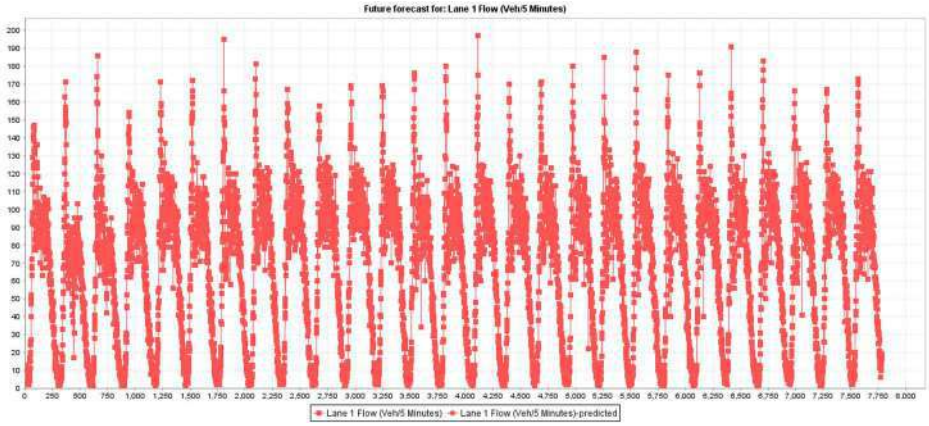

Fig.3. Future forecast for target using linear regression

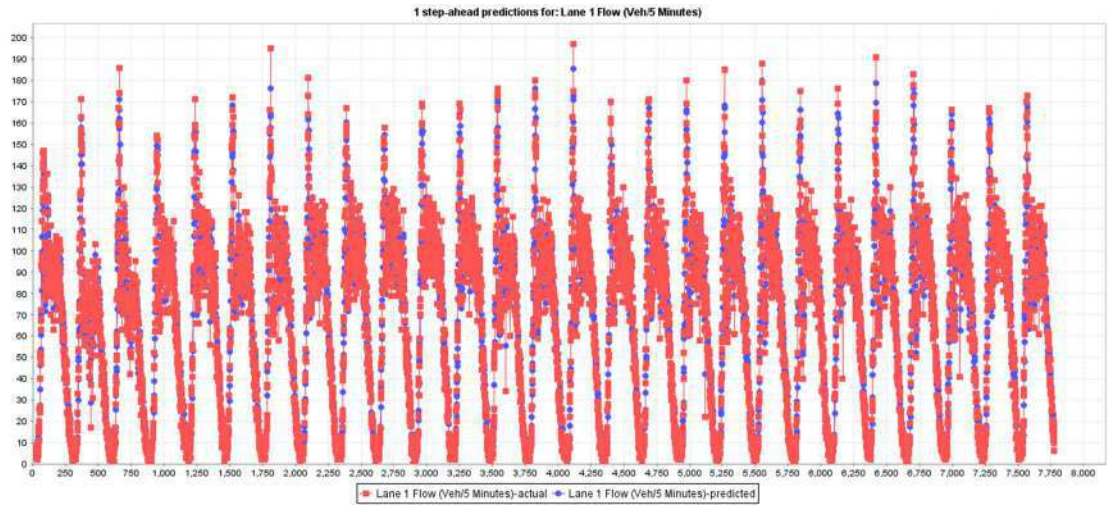

Fig.4. 1-step ahead predictions for target using linear regression

b) Multi-Layer Perceptron

Metric	1-step-ahead	2-step-ahead	3-step-ahead	4-step-ahead	5-step-ahead	6-step-ahead	7-step-ahead	8-step-ahead	9-step-ahead	10-step-ahead
MAE	11.2635	13.7831	16.3976	18.8599	21.0923	23.4731	25.8266	28.3938	31.0433	33.5108
RAE	133.863	163.8068	177.5945	184.5593	108.5255	193.1534	196.7012	203.5733	209.7759	211.6359
RMSE	13.6397	16.3185	19.0884	21.7114	24.1751	26.8254	29.4279	32.2339	35.1494	37.9459
MSE	186.0425	266.2947	364.3683	471.3865	584.4355	719.6047	866.0039	1039.0264	1235.4822	1439.8896

Table 2. Multi-Layer Perceptron Metrics

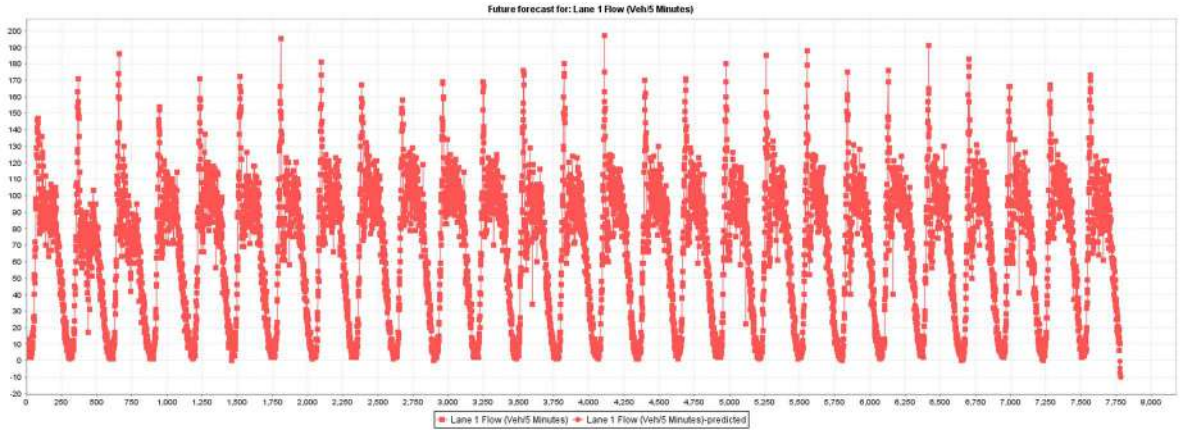

Fig. 5. Future forecast for target using multi-mayer perceptron

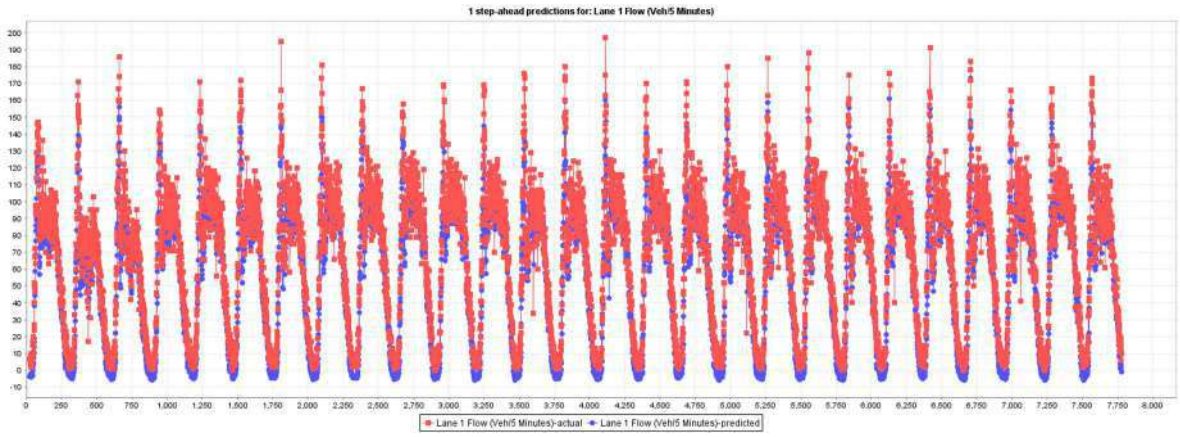

Fig. 6.Step-ahead predictions for target using multi-layer perceptron

c) RBF Network

Metric	1-step-ahead	2-step-ahead	3-step-ahead	4-step-ahead	5-step-ahead	6-step-ahead	7-step-ahead	8-step-ahead	9-step-ahead	10-step-ahead
MAE	20.5881	20.858	21.1335	21.4303	21.7275	22.0565	22.3605	22.6948	23.0372	23.3828
RAE	244.6782	247.8889	228.8866	209.7128	194.2027	101.4967	170.3026	162.7135	144.6749	147.673
RMSE	27.1056	27.5907	28.1361	28.7435	29.3172	29.9695	30.5983	31.2313	31.9144	32.5965
MSE	734.7131	761.2447	791.6411	826.1909	859.4997	858.1735	936.2555	975.2555	1018.5315	1062.5295

Table 3. RBF metrics

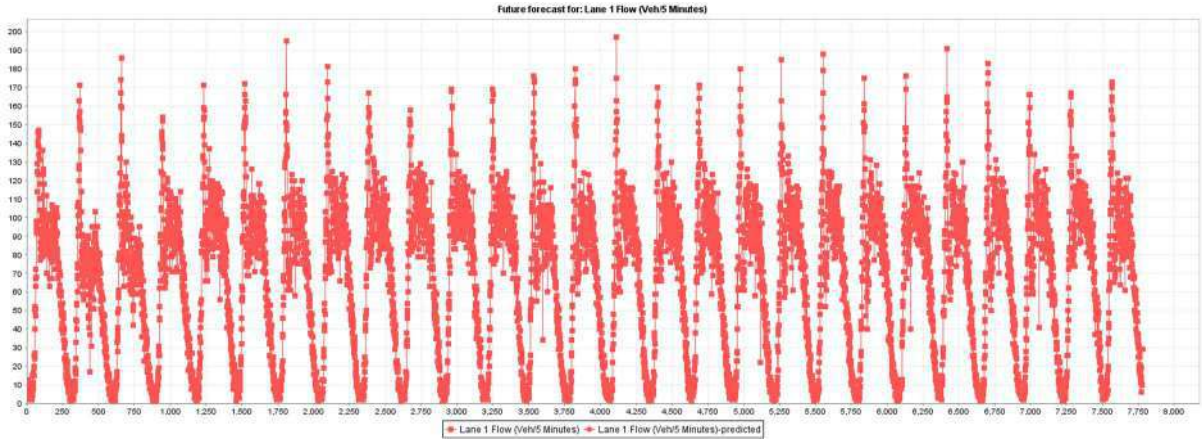

Fig. 7. Future forecast for target using RBF network

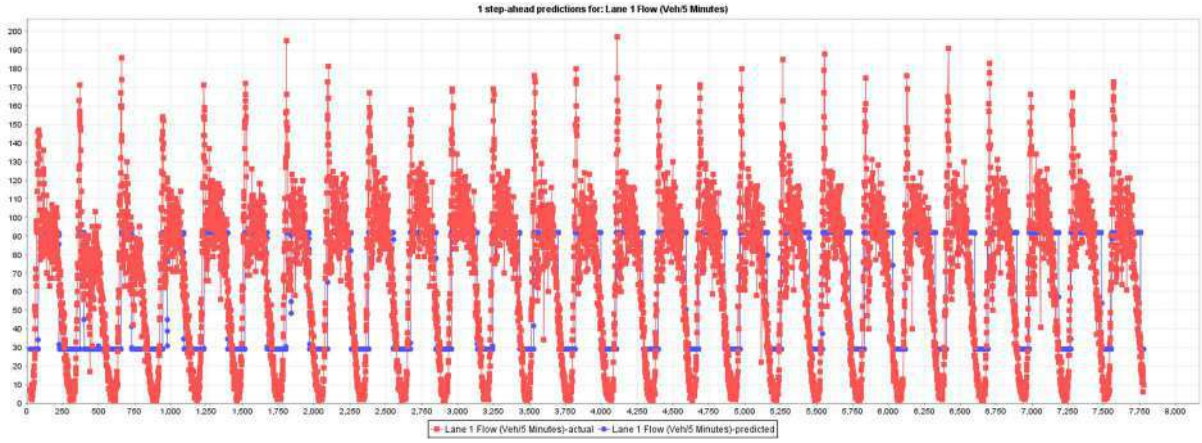

Fig.8. 1-step-ahead predictions for target using RBF network

d) RBF Regressor

Metric	1-step-ahead	2-step-ahead	3-step-ahead	4-step-ahead	5-step-ahead	6-step-ahead	7-step-ahead	8-step-ahead	9-step-ahead	10-step-ahead
MAE	7.2196	7.8062	8.4618	9.0304	10.0404	10.0404	10.4784	10.9455	11.4588	11.9374
RAE	85.8086	92.7736	91.6459	88.3695	82.2457	82.6199	79.8059	78.4751	77.4334	75.3904
RMSE	9.7249	10.5025	11.3025	11.9471	12.5349	13.0998	13.5589	14.6372	14.6372	15.2118
MSE	94.5731	110.3029	127.7474	142.7326	157.1247	171.6049	183.8441	197.988	214.248	231.1992

Table 4. RBF regressor metrics

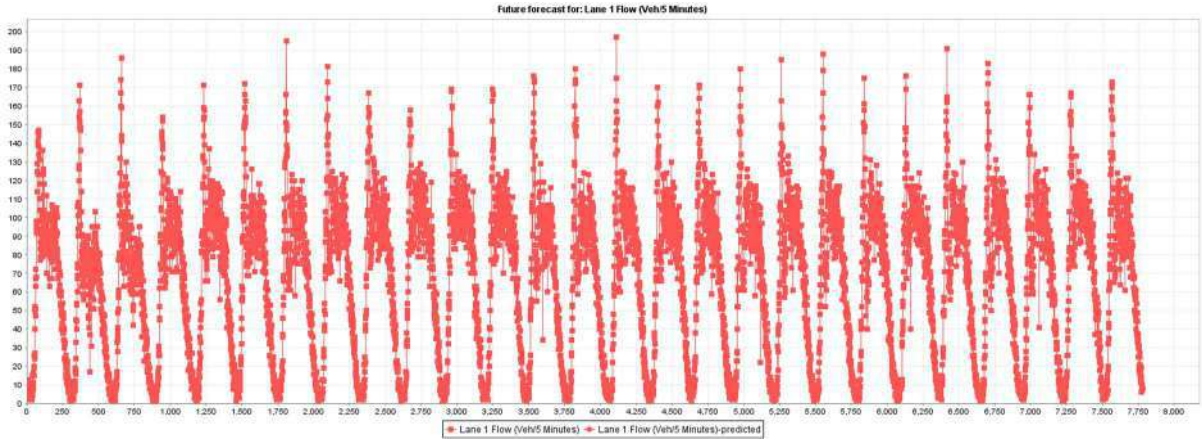

Fig.9. Future forecast for target using RBF regressor

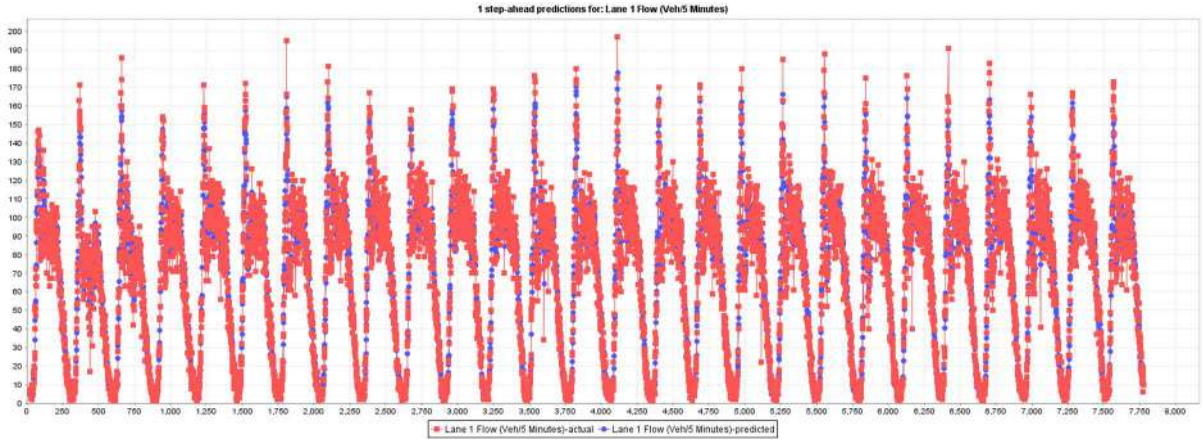

Fig.10. 1-step-ahead predictions for target using RBF regressor

e) SMO Reg

Metric	1-step-ahead	2-step-ahead	3-step-ahead	4-step-ahead	5-step-ahead	6-step-ahead	7-step-ahead	8-step-ahead	9-step-ahead	10-step-ahead
MAE	7.5506	8.4396	9.472	10.4654	11.4144	12.3323	13.1224	14.0219	14.9454	15.7413
RAE	89.7457	100.3014	102.5869	102.4128	102.0229	101.4788	99.9429	100.9941	100.9941	99.4133
RMSE	10.3603	11.7133	13.2317	14.6998	16.155	17.6047	18.9205	21.6009	21.6009	22.7999
MSE	107.3354	137.2019	174.079	216.0831	260.9834	309.9269	357.9841	466.9046	466.5988	519.8339

Table 5. SMO reg metrics

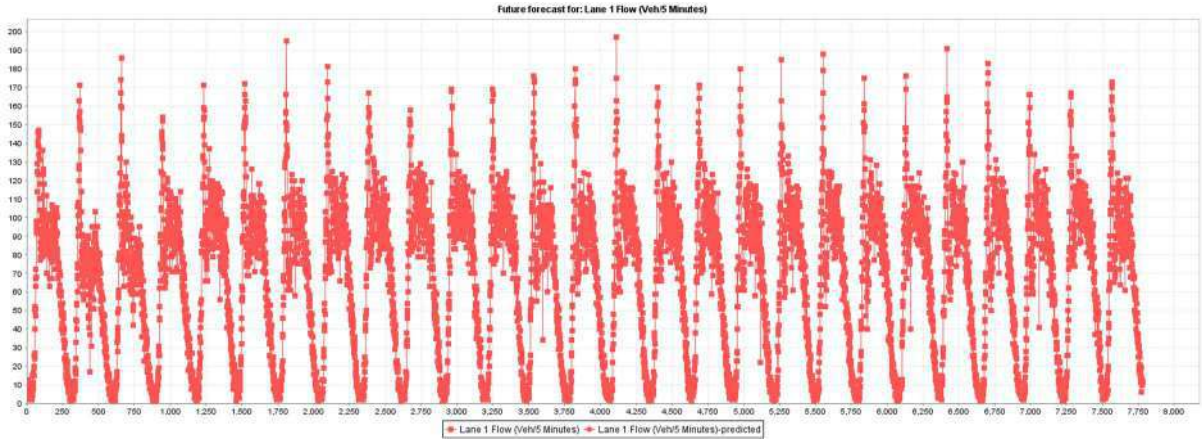

Fig.11. Future forecast for target using SMO reg

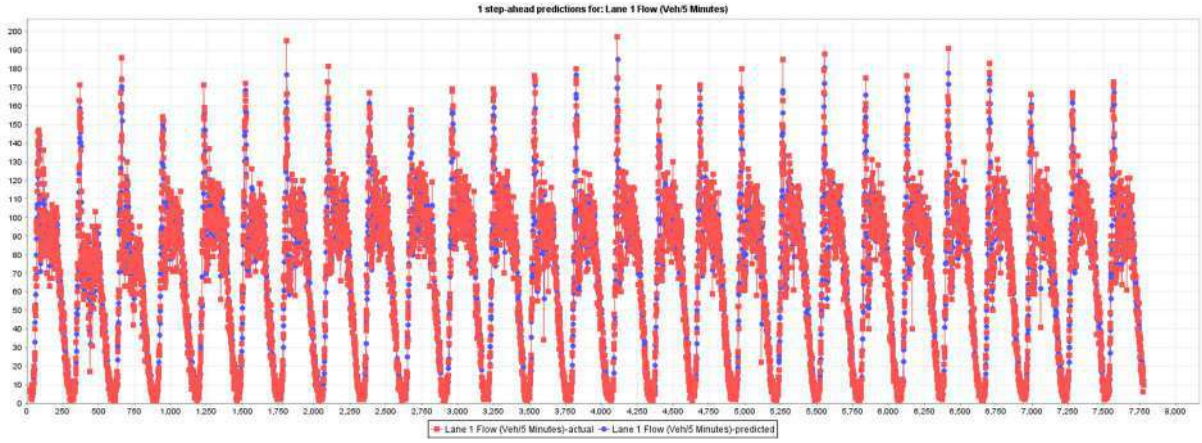

Fig.12. 1-step-ahead predictions for target using SMO reg

f) Deep Learning Model (LSTM)

Case 1: 5 Epochs

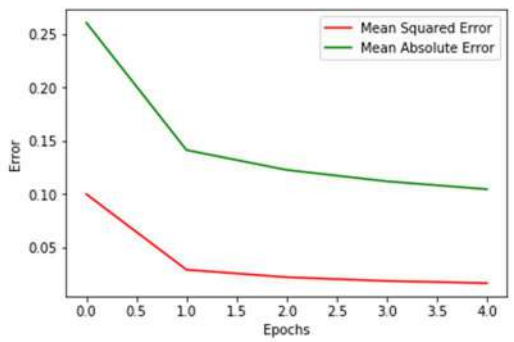

Fig.13. MSE vs MAE with 5 Epochs

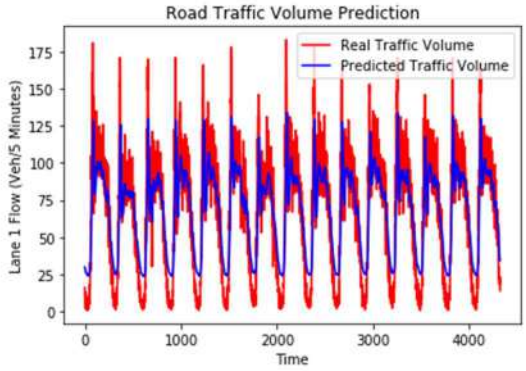

Fig.14. Real Traffic Vs Predicted Traffic with 5 Epochs

Case 2: 10 Epochs

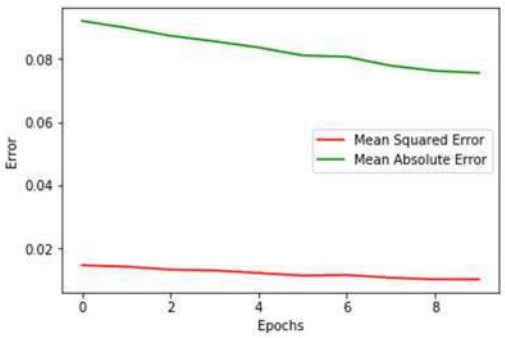

Fig.15. MSE vs MAE with 10 Epochs

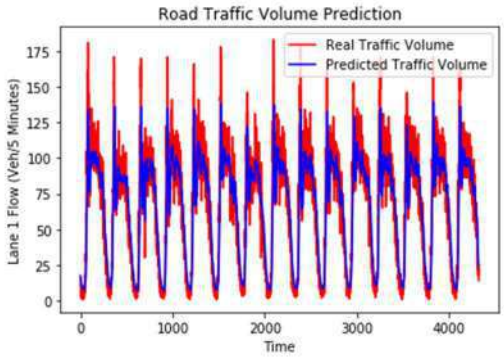

Fig.16. Real Traffic Vs Predicted Traffic with 10 Epochs

Case 3: 25 Epochs

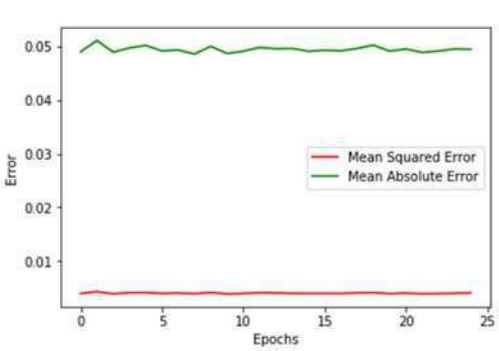

Fig. 17. MSE vs MAE with 25 Epochs

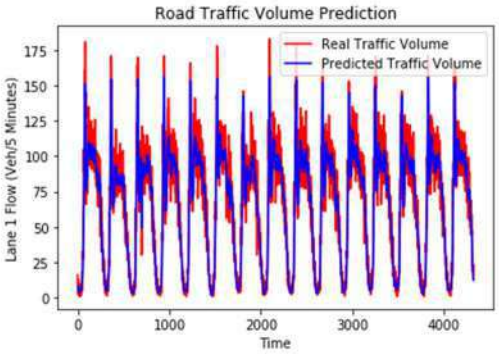

Fig.18. Real Traffic Vs Predicted Traffic with 25 Epochs

Case 4: 50 Epochs

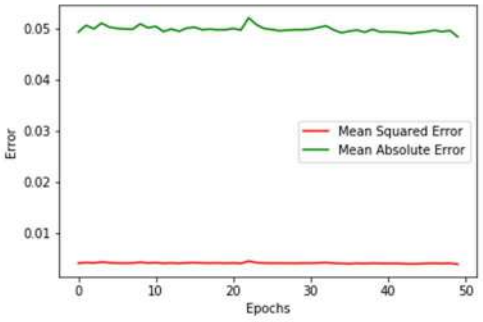

Fig. 19. MSE vs MAE with 50 Epochs

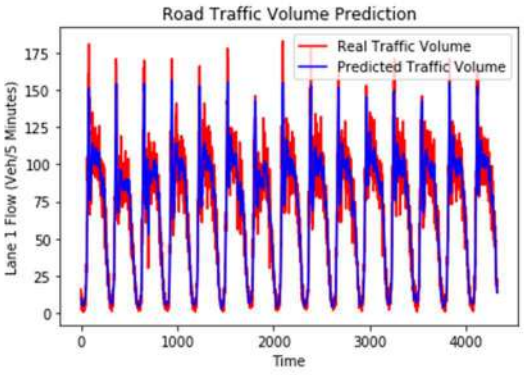

Fig. 20. Real Traffic Vs Predicted Traffic with 50 Epochs

Case 5: 100 Epochs

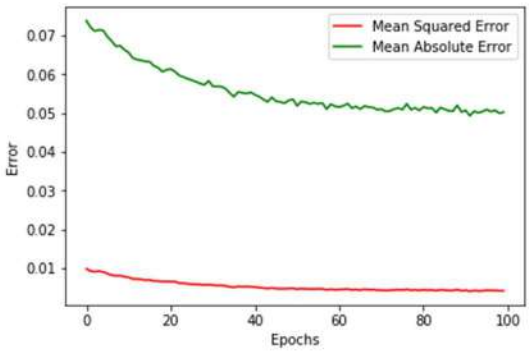

Fig. 21. MSE vs MAE with 100 Epochs

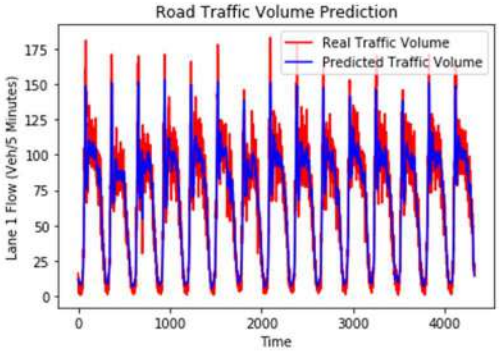

Fig. 22. Real Traffic Vs Predicted Traffic with 100 Epochs

VI. Conclusions

This study conducts experiments in WEKA using a dataset from the Caltrans Performance Measurement System (PeMS) containing vehicle traffic volume across all major metropolitan areas of California to analyze historical traffic volume and forecast traffic volume for a given day. This is a regression problem, so experiments are conducted using 5 popular regression algorithms including Linear Regression, Multi-Layer Perceptron, RBF Network, RBF Regressor, and SMO Reg algorithms. For these 5 algorithms, regression metrics are tabulated in Tables 1 to 5. Figs. 3 to 12 forecast the target variable and the future forecast of 1-step ahead data for all 5 regression algorithms. The results suggest that RBF Regressor provides the best results over linear regression, followed by SMO Regression, Linear Regression, RBF Network and MLP Network. The MLP network is not suitable. Experiments with LSTM deep learning model are conducted using the same dataset to compare LSTM with RBF Regressor. In LSTM model, results for 5 different cases are obtained for increasing numbers of epochs. Figs. 13 to 22 plotmetrics and predict real versus predicted traffic for various epochs.The results show that performance metrics improve with increasesin the number of epochs. MSE and MAE for the LSTM model is far superior to regression models.

Acknowledgements

The authors acknowledge Weiqi Wang of St. Francis Prep and Yuchen Ji of Bodwell High Schoolfor participating in this research.

REFERENCES

[1] Dai, X., Fu, R., Lin, Y., Li, L., & Wang, F. Y. (2017). DeepTrend: A deep hierarchical neural network for traffic flow prediction. *arXiv preprint arXiv:1707.03213*.

[2] Chen, W., An, J., Li, R., Fu, L., Xie, G., Bhuiyan, M. Z. A., & Li, K. (2018). A novel fuzzy deep-learning approach to traffic flow prediction with uncertain spatial–temporal data features. *Future Generation Computer Systems*, *89*, 78-88.

[3] Manornjitham, Raj. P, Lal, H.K (2018). A Survey of Road Traffic Prediction with deep learning. *International Journal of Pure and Applied Mathematics,* 2065-2073.

[4] Jia, Y., Wu, J., & Xu, M. (2017). Traffic flow prediction with rainfall impact using a deep learning method. *Journal of advanced transportation*, *2017*.

[5] Yang, B., Sun, S., Li, J., Lin, X., & Tian, Y. (2019). Traffic flow prediction using LSTM with feature enhancement. *Neurocomputing*, *332*, 320-327.

[6] S.K.Groen. (2012), A Review of Traffic growth rate calculations. *Shaping the future: Linking policy, research and outcomes, 25th ARRB Conference, Perth, Australia,* 1-15.

[7] Polson, N. G., & Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, *79*, 1-17.

[8] Du, S., Li, T., Gong, X., Yu, Z., Huang, Y., &Horng, S. J. (2018). A hybrid method for traffic flow forecasting using multimodal deep learning. *arXiv preprint arXiv:1803.02099*.

[9]Lv, Z., Xu, J., Zheng, K., Yin, H., Zhao, P., & Zhou, X. (2018, January). LC-RNN: A Deep Learning Model for Traffic Speed Prediction. In *IJCAI* (pp. 3470-3476).

[10] Wang, J., Chen, R., & He, Z. (2019). Traffic speed prediction for urban transportation network: A path based deep learning approach. *Transportation Research Part C: Emerging Technologies*, *100*, 372-385.

[11] Zhang, S., Kang, Z., Hong, Z., Zhang, Z., Wang, C., & Li, J. (2018, July). Traffic flow prediction based on cascaded artificial neural network. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*(pp. 7232-7235). IEEE.

[12] Xiao, Y., & Yin, Y. (2019). Hybrid LSTM Neural Network for Short-Term Traffic Flow Prediction. *Information*, *10*(3), 105.

[13] Tian, Y., Zhang, K., Li, J., Lin, X., & Yang, B. (2018). LSTM-based traffic flow prediction with missing data. *Neurocomputing*, *318*, 297-305.

Scientific and Practical Cyber Security Journal (SPCSJ) 3(3): 40 - 44 ISSN 2587-4667 Scientific Cyber Security Association (SCSA)

Improvement of Implementation of Merkle Crypto System

Улучшение Реализации Крипто Системы Меркле

Артуро Аракельян[1], Олег Полихенко[2]
Грузинский университет, Тбилиси, Грузия[1]
Национальный авиационный университет, Киев, Украина[2]

ABSTRACT:

This article describes hash-based digital signature systems. These systems are safe against quantum computer attacks. These systems have performance problems. We implemented the Merkle digital signature algorithm using recursion. A performance analysis was conducted. To improve the efficiency in the implementation of this algorithm, we replaced the recursion with loops. An analysis of the resulting implementation was carried out. Changing the implementation gave us very good results.

Резюме.

В данной статье описаны системы электронной подписи, основанные на хешировании. Данные системы защищены от атак квантового компьютера. Данные системы имеют проблемы эффективности. Мы реализовали алгоритм электронной подписи Меркле, с помощью рекурсии. Проведен был анализ эффективности. Для улучшения эффективности в реализации данного алгоритма рекурсию мы заменили циклами. Был проведен анализ полученной реализации. Изменение реализации дало довольно хорошие результаты.

KEYWORDS: Merkle, crypto system, improvement, cryptography

Идет активная работа над разработкой и усовершенствованием квантовых компьютеров. Криптосистемы, которые используются в практике уязвимы к атакам квантовых компьютеров. Безопасность этих систем основана на проблеме факторизации больших чисел и вычислении дискретных логарифмов, а квантовый компьютер может легко решить эту проблему[1,2].

Ведется активная работа над созданием криптосистем, которые защищены от атак квантового компьютера. Такими являются системы электронной подписи, основанные на хешировании. Безопасность данных крипто систем основывается на стойкости к коллизиям хеш функций.

Схема одноразовой подписи Лэмпорта

Была предложена схема одноразовой подписи Лэмпорта (Lamport–Diffie one-time signature scheme), данная схема является электронной подписью, основанной на хешировании. В этой схеме генерация ключа и генерация подписи являются эффективнымы, но размер подписи является довольно большим.

Схема одноразовой подписи Винтерница

Для уменьшения размера подписи была предложена схема одноразовой подписи Винтерница (Winternitz one-time signature scheme). В данной схеме одной строчкой ключа подписываются одновременно несколько битов хешированного сообщения, этим уменьшается длина подписи.

Схемы одноразовой подписи неудобны в использовании, потому что для подписи каждого сообщения нужно использовать уникальную пару ключей.

Крипто система Меркле

Была предложена крипто система Меркле для решения проблемы одноразовой пары ключей. В Merkle используется бинарное дерево, для замещения большого количества ключей верификации одним открытым ключом, корнем бинарного дерева. Данная криптосистема использует схему одноразовой подписи Лэмпорта или Винтерница и криптографическую хеш функцию

Нами был реализован алгоритм данной системы, в данном алгоритме использована рекурсия.

Рекурсия:

1. Importing necessary libs
2. Define class
3. Defining “ alt_hashes(hashes) “ method
4. Set list “ arr ”
5. If hashes == “ ”, raise Exception
6. Foreach loop
   6.1. sorting hashes and appending into arr
7. Length_of_block == length of arr
8. While loop, if length is odd, copy last element in list
   8.1. append it into arr list
9. Set list “ another_arr ”
10. Foreach loop
    10.1. For loop with range from 0 to length of “ arr “ and iteration by 2
        10.1.1. Define variable with “ sha512() “ value
        10.1.2. Hash elements that are in “ arr “ list
        10.1.3. Apennd them into new “ another_arr “ list
        10.1.4. Return this list in hex
11. Set list “ hash_arr ”
12. Foreach loop
    12.1. Generate Hex and put it into “ hash_arr ” list
13. Create message put it in “ st “ variable
14. Convert “ st “ value in binary
15. First_secret_key = hash_arr[0]
16. Second_secret_key = hash_arr[1]

17. Generate “ one-time signature ”
    17.1. If st == 0
        17.1.1. Choose “ First_secret_key “ bit
    17.2. Else
        17.2.1. Choose “Second _secret_key “ bit
18. First_pub_key = hash(hash_arr[0])
19. Second_pub_key = hash (hash_arr[1])
20. Encryption
    20.1. Concatenate “ one-time signature ” with message’s hash
21. Verification of “ one-time signature ”
    21.1. If bit of “ one-time signature ” == 0
        21.1.1. Compare with “ First_secret_key “ bit
    21.2. Else
        21.2.1. Compare with “Second _secret_key “ bit
22. Verification of “ signature ”
    22.1. Concatenate siblings with each other
    22.2. If this equals to public key
        22.2.1. Sign is correct
    22.3. Else
        22.3.1. Sign is not correct

В данном примере показана реализация алгоритма „Меркле“ с помощью рекурсии. Время подсчета “public key” для 8 элементов составляет 0.0159 секунд, время шифрования - 0.01684, время подтверждения - 0.0288883 .

Мы изменили рекурсию на циклы для улучшения эффективности.

Реализация с помощью циклов:

1. Importing necessary libs
2. Define class
3. Defining “ loop_hashes(hashes) “ method
4. Set list “ arr ”
5. If hashes == “ ”, raise Exception
6. Foreach loop
    6.1. sorting hashes and appending into arr
7. Length_of_block == length of arr

8. While loop, if length is odd, copy last element in list
    8.1. append it into arr list
9. Set list “ another_arr ”
10. Set i = 0
11. While loop, Length_of_block > 1
    11.1. Set to hash_f sha512()
    11.2. Concatenate arr[i] and arr[i+1]
    11.3. append it into “another_arr” list
    11.4. append arr[i+1] to “auth_list” list
    11.5. i = i + 2
    11.6. if i equal to “Length_of_block”
        11.6.1. set to “Length_of_block” “Length_of_block / 2”
        11.6.2. i = 0
        11.6.3. set “another_arr” to “arr”
        11.6.4. empty “another_arr”
    11.7. return “arr”
12. Set list “ hash_arr ”
13. Foreach loop
    13.1. Generate Hex and put it into “ hash_arr ” list
14. Create message put it in “ st “ variable
15. Convert “ st “ value in binary
16. First_secret_key = hash_arr[0]
17. Second_secret_key = hash_arr[1]
18. Generate “ one-time signature ”
    18.1. If st == 0
        18.1.1. Choose “ First_secret_key “ bit
    18.2. Else
        18.2.1. Choose “Second _secret_key “ bit
19. First_pub_key = hash(hash_arr[0])
20. Second_pub_key = hash (hash_arr[1])
21. Encryption
    21.1. Concatenate “ one-time signature ” with message’s hash
22. Verification of “ one-time signature ”
    22.1. If bit of “ one-time signature ” == 0

22.1.1. Compare with “ First_secret_key “ bit

22.2. Else

22.2.1. Compare with “Second _secret_key “ bit

23. Verification of “ signature ”

23.1. Concatenate siblings with each other

23.2. If this equals to public key

23.2.1. Sign is correct

23.3. Else

23.3.1. Sign is not correct

В данном примере показана реализация алгоритма „Меркле“ с помощью цикла. Время подсчета “public key” для 8 элементов составляет 0.0061761 секунд, время шифрования - 0.0080878, время подтверждения - 0.0181923. Как мы видим изменения реализации дало довольно хорошие результаты.

БИБЛИОГРАФИЯ:

1. Guang Hao Low, Artur Scherer, and Dominic W. Berry , Black-Box Quantum State Preparation without Arithmetic Yuval R. Sanders, Phys. Rev. Lett. 122, 020502 – Published 16 January 2019
2. Liu J. et al. (2019) Formal Verification of Quantum Algorithms Using Quantum Hoare Logic. In: Dillig I., Tasiran S. (eds) Computer Aided Verification. CAV 2019. Lecture Notes in Computer Science, vol 11562. Springer, Cham