ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal) | Impact Factor: 7.249|

Website: www.ijirset.com

Vol. 8, Issue 7, July 2019

Safe and Dependable Intrusion Detection Method Designs Created with Artificial Intelligence Techniques

Pankit Arora[1*], Sachin Bhardwaj[2]

Sr. Risk Modeling/Analytics Analyst, Mr. Cooper, USA[1]

Assistant Manager (GRC), EXL Service Pvt Ltd. India[2]

ABSTRACT: In terms of network security, intrusion detection systems are crucial. A prediction model called an intrusion detection model is used to determine if network data traffic is intrusion or normal. Accurate models for prediction, classification, and grouping are created using machine learning methods. This work uses machine learning classification techniques, specifically Logistic Regression, Gaussian Naive Bayes, Support Vector Machine, and Random Forest, to build classification and predictive models for intrusion detection. The NSL-KDD data set is used to test these methods. According to experimental findings, the Random Forest Classifier outperforms the other techniques in determining whether data traffic is an attack or normal.

KEYWORDS: Classification Algorithms; Intrusion Detection; Machine Learning; Network Security; Supervised Learning, Data Analytics; Big Data; Anomaly based algorithm; Classification algorithms; Intrusion detection; Cyber Security; Cloud Security; Network; Cyber; Cyber Threats; Threat Analysis; Information Security; Data security.

I. INTRODUCTION

In the present era, it is difficult to envision world without internet. Every person has dependency on internet. It has become an important model in various applications such as education, business and others. So security of the data that is communicated through internet is necessary. Secure network is maintained by Intrusion Detection System (IDS). IDS observes the data traffic carefully and identifies it as normal or spam. Nowadays most of the applications depends on the advance network technologies namely wireless networks, wireless sensor networks and Bluetooth [1-9]. In case of wireless sensor networks security mechanisms such as key-management protocols, authentication techniques and security protocols cannot be used because of resource constraints. Intrusion Detection System is the ideal security mechanism for wireless sensor networks.

1.1. Intrusion Detection System

A security mechanism used to monitor the abnormal behavior of the network is an Intrusion Detection System (IDS) [10-17]. The IDS identifies and informs that whether the user activity is normal or not. The user's activities are compared by the IDS with the already stored intrusion records to identify the intrusion.

Accurate predictive models can be built for large data sets using supervised machine learning techniques that are not possible by traditional methods. As specified by Tom Mitchell, machine learning based intrusion detection falls under two categories Anomaly and Misuse. IDS learns the patterns by the training data, so the misuse based method is used. Misuse based detection can detect only the known attack, new attacks cannot be identified. Anomaly based IDS observes the normal behavior and if there is a change in the behavior then it considers that behavior as anomaly. So anomaly based IDS can detect new attacks that are not learned from the training model.

Till now different machine learning techniques such as artificial neural networks, Support Vector Machine and Naive Bayes based techniques are proposed for the intrusion detection. A new detection by combining different techniques, a hybrid detection technique is proposed. The literature on comparison of supervised machine learning techniques in

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal) | Impact Factor: 7.249|

Website: www.ijirset.com

Vol. 8, Issue 7, July 2019

intrusion detection is limited [18-27]. Hence this paper aims at understanding the implications of using supervised machine learning techniques on intrusion detection.

The paper is organized as follows. Section 2 discusses some research work on Intrusion Detection System. In Section 3 various supervised machine learning methods used are discussed. Section 4 gives brief introduction about the data- set used. Section 5 discusses the methodology. Conclusion and future scope is given in section 7.

II. RELATED WORK

Recently authors used algorithms namely Random Forest, Naive Bayes, K-means and Support Vector Machine to identify four types of attacks. They also proposed best feature selection method. Concluded that the Random Forest Classifier (RFC) outperforms the other methods. They have mentioned that hierarchical clustering method can be used to improve the performance.

Authors proposed semi supervised machine learning based intrusion detection. Authors have not considered the resource consumption. Combination of different classifiers to identify the intrusion is proposed. They used supervised classification orunsupervised clustering for filtering of the data. They used NSL-KDD data-set and tested with decision tree classifier. But the proposed method works only for binary class classification.
Authors proposed intrusion detection system using supervised machine learning techniques to identify the on line network data as normal or not. The proposed method identifies probe and Denial of Service attacks only, but the other attacks are not considered.

A framework of machine learning approach is proposed. Intrusion is identified by analyzing the local features. Researchers proposed Naive Bayes based multiclass classifier to identify the intrusions. They suggested that intrusion detection is possible by Hidden Naive Bayes (HNB) model. Denial of Service attacks is identified with good accuracy compared to other attacks.

Studies suggested Intrusion detection technique using Support Vector Machine (SVM). They also used feature removal method to improve the efficiency. Using the proposed feature removal method they selected best nineteen features from the KDD-CUP99 data-set. In the proposed method the data set used is very small. A light weight IDS is proposed. The proposed method mainly focused on pre-processing of the data so that only important attributes can be used. The first step is to remove the redundant data so that the learning algorithms give the unbiased result.

A survey on intrusion detection systems was conducted. Information about IDSs such as classification, Intrusion type, computing location and infrastructure are discussed. They discussed about the Mobile Adhoc Networks (MANET) IDS. They compared MANETIDS and the Wireless Sensor Networks (WSN) IDS. Authors suggested that for mobile applications distributed and cooperative IDS schemes are suitable. For stationary applications centralized IDSs are suitable and for cluster based applications hierarchical IDSs are suitable. Authors proposed intrusion detection framework to detect routing attacks [28-36]. Specification based approach is used to detect routing attacks. Authors claim that the proposed method has low False Positive Rate (FPR) and good intrusion detection rate. The proposed method works only for static networks. Investigators developed IDS for Sink, Cluster Head (CH) and for a Sensor Node (SN) separately and combined altogether to identify the intrusion in heterogeneous Cluster Based Wireless Sensor Networks (CWSN) but the detection rate for U2R, R2L and Probe attacks is very low.

III. SUPERVISED MACHINE LEARNINGTECHNIQUES

Our work is to design a network intrusion detection system with the different supervised machine learning classifiers. This paper is to investigate the performance of the classifiers namely Logistic Regression, Support Vector Machine (SVM), Gaussian Naive Bayes (GNB) and Random Forest in intrusion detection. These classifiers are discussed below.

3.1. Logistic regression

To solve the classification problems Logistic Regression (LR) is used. TLR works for both binary classification and

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal) | Impact Factor: 7.249|

Website: www.ijirset.com

Vol. 8, Issue 7, July 2019

multiclass classification. Probability of occurrence of an event is predicted by fitting data to the Logistic function. The values selected by the logistic function are in the range 0 and 1. If the value is 0.5 and above then it is labeled as otherwise 0.

3.2. Support vector machine

Mainly for classification problems Support Vector Machine algorithm is used, but it can be used in regression problems also. N-dimensional feature space is considered to plot each data item as a point with the value of each feature as a particular coordinate. Then classification is made by finding the hyper-plane that differentiates the two classes very well. Support Vectors are the co-ordinates of specific observation that lies closest to the boarder line. In case of SVM training samples are divided into different subsets called as support vectors, the decision function is specified by these support vectors. This paper is based on the liner kernel method of SVM for intrusion detection.

3.3. Gaussian naive Bayes

The Gaussian Naive Bayes algorithm is the supervised learning method. Probabilities of each attribute whichbelongs to each class are considered for a prediction. This algorithm is assumes that the probability of each attribute belonging to a given class value is not depends on all other attributes. If the value of the attribute is known the probability of a class value is called as the conditional probabilities. Data instances provability can be found out by multiplying all attributes conditional probabilities together. Prediction can be made by calculating the each class instance probabilities and by selecting the highest probability class value.

3.4. Random forest classifier

In 2001 Breiman proposed the random forest machine learning classifier. It is a collaborative method which works based on the proximity search. It is decision tree based classifier. It makes use of standard divide and conquers approach to improve the performance. The main principle behind random forest is that strong learner group is created by a group of weak learners. It is applicable to disjunctive hypothesis.
Comparison of above mentioned classifiers based on the Precision, Recall, F1-Score and Accuracy is discussed in section 6.

IV. INTRUSION DETECTION DATA-SET

The standard intrusion detection data set KDDCUP9923 has redundant records. This may lead to unfair result of the machine learning algorithms. So the supervised machine learning algorithms are tested NSL-KDD24 data-set which is the advanced version of the KDDCUP99 intrusion detection data-set. It has 42 features and the four simulated attacks [37-41].

Denial of Service (DoS) attack: Over usage of the bandwidth or non-availability of the system resources leads to the DoS attacks. Examples: Neptune, Teardrop and Smurf.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal) | Impact Factor: 7.249|

Website: www.ijirset.com

Vol. 8, Issue 7, July 2019

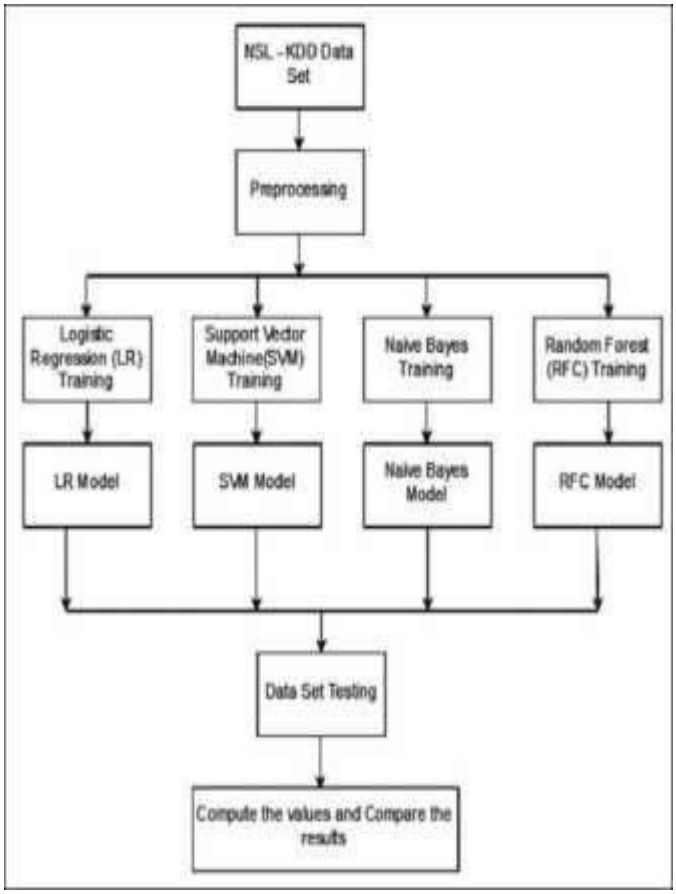

Fig.1. Methodology

User to Root (U2R) Attack: Initially attacker access normal user account, later gain access to the root by exploiting the vulnerabilities of the system.Examples: Perl, Load Module and Eject attacks.

Probe Attack: Have an access to entire network information before introducing an attack. Examples: ipsweep, nmap attacks.

Root to Local (R2L) Attack: By exploiting some of the vulnerabilities of the network attacker gains local access by sending packets on a remotemachine. Examples: imap, guess password and ftp- write attacks.

V. METHODOLOGY

The methodology used is shown in the Fig 1. In pre- processing step all the categorical data which are in textual form are converted to numerical form. Pre-processed data is divided as testing data and training data. The models are built using Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest classifiers. These models are used for predicting the labels of the test data. Actual labels and predicted labels are compared. Accuracy, True Positive Rate (TPR) and False Positive Rate (FPR) are computed. Based on these parameters performance of the models are compared.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal) | Impact Factor: 7.249|

Website: www.ijirset.com

Vol. 8, Issue 7, July 2019

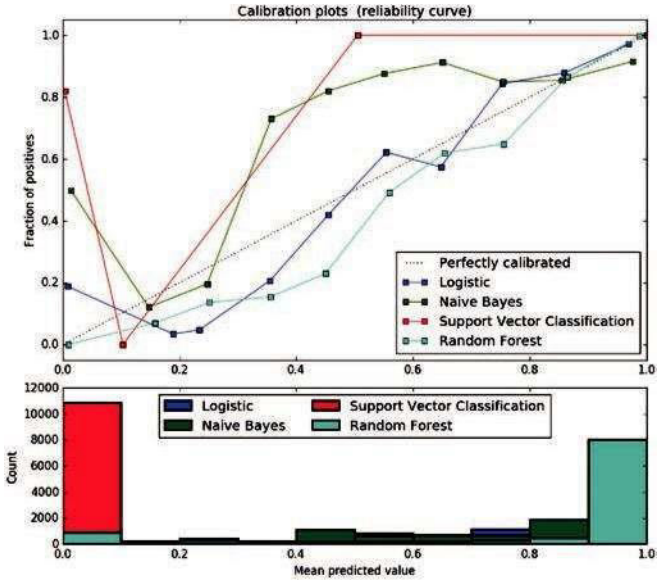

Fig. 2. Calibration Plots

Following steps are used to build the models.
•Pre-process the data set.
•The data set is divided as training data and testing data
•Build the classifier model on training data for
•Logistic Regression
•Support Vector Machine
•Gaussian Naive Bayes
•Random Forest
•Read the test data
•Test the classifier models on training data
•Compute and compare TPR, FPR, Precision, Recall, F1-Score and Accuracy for all the models.

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

Supervised machine learning algorithms namely Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest are tested on NSL-KDD dataset, the newstandard intrusion detection data-set. These algorithms are tested on Intel Core (TM) i5-3230M CPU @2.60 GHZ, 4 GB RAM and coding is done by Python. The result of the experiment is represented as a Reliability curve. In Reliability curve estimated probabilities are plotted against the true empirical probabilities. Figure 2 shows the Reliability Curve for the above mentioned supervised machine learning classifiers. Reliability curve for the ideal classifier falls near the diagonal because the estimated probabilities and empirical probabilities are nearly equal.

X-axis probability space is divided into ten bins as shown in Fig. 2. Estimated probabilities values ranging from 0 to 0.1, to 0.2 and so on. The values 0 to 0.1 belongs to I bin, 0.1 to 0.2 belongs to II bin and similarly the other ranges. From the graph shown in Fig. 2, it can be concluded that theRandom Forest classifier out performs the other methods in identifying the network traffic as normal or an attack. Whereas the SVM identifies the intrusion with the lowest probability estimate.

Quality of the classification models is identified by plotting the Receiver Operating Characteristics (ROC) curve. In ROC curve shows FPR verses TPR. ROC curve for the above mentioned classifiers is shown in the Fig. 3. Random

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal) | Impact Factor: 7.249|

Website: www.ijirset.com

Vol. 8, Issue 7, July 2019

Forest has highest TPR. Hence, the ROC curve for Random forest isplotted separately. By observing the graphs, it can be concluded that the Random forest classifier has lowest FPR and highest TPR in identifying attacks. It outperforms the other techniques. Whereas Support Vector Machine has highest FPR (39%) and minimal TPR (75%) for intrusion detection. This is due to the fact that too many features from the data set is considered15 and SVM's linear kernel functionis used.

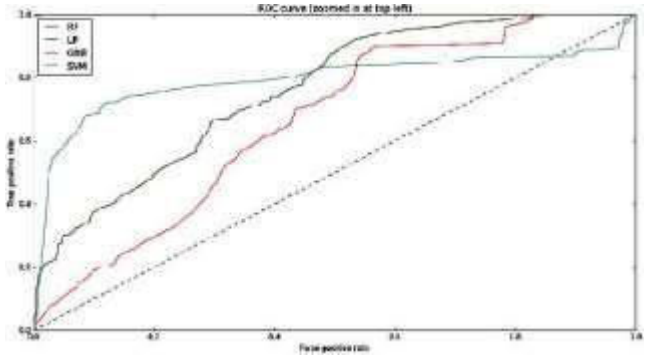

Fig. 3(a). Receiver Operating Characteristics (ROC) curve

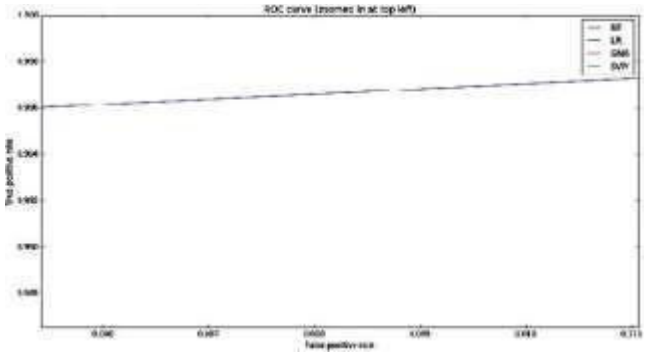

Fig. 3(b). ROC for Random Forest Classifier

Table 1. Performance Measures

–	Precision	Recall	F1-Score	Accuracy
LR	0.83	0.85	0.82	0.84
GNB	0.79	0.81	0.78	0.79
SVM	0.76	0.79	0.77	0.75
RFC	0.99	0.99	0.99	0.99

Table 1 shows Precision, Recall, F1-Score and Accuracy of the supervised machine learning classifiers in identifying the intrusion. Based on the results shown in the Table 1 it can be identified that Random Forest classifier with the highest accuracy, outperforms the other methods. Whereas SVM has the lowest accuracy, Logistic Regression algorithm has the good accuracy than Gaussian Naive Bayes and SVM.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal) | Impact Factor: 7.249|

Website: www.ijirset.com

Vol. 8, Issue 7, July 2019

VII. CONCLUSIONS AND FUTURE SCOPE

An attempt has been made to check the performance of the supervised machine learning classifiers namely Support Vector Machine, Random Forest, Logistic Regression and Gaussian Naive Bayes are compared for intrusion detection. These algorithms are tested with the NSL-KDD data-set. Effective classifier is identified by comparing the performances based on the precision, recall, F1-Score and accuracy. From the observed results it can be concluded that the Random forest classifier outperforms other classifiers for the considered data-set and parameters. It has the accuracy of 99%. The work can be extended by considering the classifiers for multiclass classification and considering only the important attributes for the intrusion detection.

REFERENCES

1. A. Serageldin, H. Alturkostani, and A. Krings, “On the reliability of DSRC safety applications: a case of jamming,” in International Conference on Connected Vehicles and Expo, 2013, pp. 501–506.
2. B.K. Chaurasia, R.S. Tomar, S. Verma, G.S. Tomar, Suitability of manet routing protocols for vehicular ad hoc networks, in: 2012 International Conference on Communication Systems and Network Technologies, IEEE, 2012, pp.334–338.
3. C. Laurendeau and M. Barbeau, “Threats to security in DSRC/WAVE,” in Proc. 5th International Conference on Ad-Hoc Networks & Wireless, LNCS 4104, 2006, pp. 266–279.
4. E. Hamida, H. Noura, and W. Znaidi, “Security of cooperative intelligent transport systems: standards, threats analysis and cryptographic countermeasures,” Connected Vehicles, V2V Communications, and VANET, vol. 4, no. 3, pp. 380–423, 2015.
5. F.K. Karnadi, Z.H. Mo, K.-c. Lan, Rapid generation of realistic mobility models for vanet, in: 2007 IEEE Wireless Communications and Networking Confer-ence, IEEE, 2007, pp.2506–2511.
6. H. Hasbullah, I. Soomro, and J. Ab Manan, “Denial of service (DOS) attack and its possible solutions in VANET,” International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering, vol. 4, no. 5, pp. 813–817, 2010.
7. H.-M. Zimmermann, I. Gruber, C. Roman, A Voronoi-based mobility model for urban environments, in: 11th European Wireless Conference 2005-Next Generation Wireless and Mobile Communications and Services, VDE, 2005, pp.1–5.
8. J. Härri, F. Filali, C. Bonnet, M. Fiore, Vanetmobisim: generating realistic mo-bility patterns for vanets, in: Proceedings of the 3rd International Workshop on Vehicular Ad Hoc Networks, ACM, 2006, pp.96–97.
9. J. Harri, M. Fiore, Vanetmobisim–vehicular ad hoc network mobility exten-sion to the canumobisim framework, Institut Eurécom Department of Mobile Commu 6904 (2006) 1–19.
10. J. Zhao, G. Zucchelli, and M. Roggero, “Design of FMCW radars for active safety applications,” http://embedded-computing.com/articles/design-fmcw-radars-active-safety-applications/, 2015.
11. L. Bononi, M. Di Felice, M. Bertini, E. Croci, Parallel and distributed simulation of wireless vehicular ad hoc networks, in: Proceedings of the 9th ACM In-ternational Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems, ACM, 2006, pp.28–35.
12. M. Brooker, “Mutual interference of millimeter-wave radar systems,” IEEE Transactions on Electromagnetic Compatibility, vol. 49, pp. 170–181, 2007.
13. M. Piorkowski, M. Raya, A.L. Lugo, P. Papadimitratos, M. Grossglauser, J.-P. Hubaux, TRANS: realistic joint traffic and network simulator for vanets, Mob. Comput. Commun. Rev. 12 (2008) 31–33.
14. N. Li and Y. Zhang, “A survey of radar ECM and ECCM,” IEEE Trans. Aerospace and Electronic Systems, vol. 31, no. 3, pp. 1110–1120, 1995.
15. N. Wisitpongphan, O.K. Tonguz, J.S. Parikh, P. Mudalige, F. Bai, V. Sadekar, Broadcast storm mitigation techniques in vehicular ad hoc networks, IEEE Wirel. Commun. 14 (2007) 84–94.
16. Q. Chen, T. Roth, T. Yuan, J. Breu, F. Kuhnt, M. Zollner, M. Bogdanovic, C.Weiss, J. Hillenbrand, and A. Gern, “DSRC and radar object matching for cooperative driver assistance systems,” in IEEE Intelligent Vehicles Symposium, 2015, pp. 1348–1354.
17. Q. Xu, T. Mak, J. Ko, and R. Sengupta, “Vehicle-to-vehicle safety messaging in DSRC,” in Proceedings of the 1st ACM International Workshop on Vehicular ad hoc Networks, 2004, pp. 19–28.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal) | Impact Factor: 7.249|

Website: www.ijirset.com

Vol. 8, Issue 7, July 2019

18. R. Chauhan, "A platform for false data injection in frequency modulated continuous wave radar," http://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=4983&context=etd, 2014.
19. R. Fernandes, P.M. d'Orey, M. Ferreira, Divert for realistic simulation of hetero-geneous vehicular networks, in: The 7th IEEE International Conference on Mo-bile Ad-Hoc and Sensor Systems (IEEE MASS 2010), IEEE, 2010, pp.721–726.
20. R. Mangharam, D.S. Weller, D.D. Stancil, R. Rajkumar, J.S. Parikh, Groovesim: a topography-accurate simulator for geographic routing in vehicular networks, in: Proceedings of the 2nd ACM International Workshop on Vehicular Ad Hoc Networks, ACM, 2005, pp.59–68.
21. S. Senthilkumar, R. Nithya, P. Vaishali, R. Valli, G. Vanitha, & L. Ramachanndran, "Autonomous navigation robot", International Research Journal of Engineering and Technology, vol. 4, no. 2, 2017.
22. S. Jaap, M. Bechler, L. Wolf, Evaluation of routing protocols for vehicular ad hoc networks in typical road traffic scenarios, 2005, pp.584–602.
23. S. Maddio, A. Cidronali, M. Passafiume, G. Collodi, and G. Manes, "Interference cancellation for the coexistence of 5.8 GHz DSRC and 5.9 GHz ETSI ITS," in IEEE MTT-S International Conference on Microwaves for Intelligent Mobility, 2015, pp. 1–4.
24. S. Roome, "Digital radio frequency memory," Electronics & Communication Engineering Journal, vol. 2, no. 4, pp. 147–153, 1990.
25. S.-Y. Wang, C.-L. Chou, Nctuns Simulator for Wireless Vehicular Ad Hoc Net-work Research, Ad Hoc Networks: New Research, Nova Science Publishers, 2009.
26. T. Fujiki, M. Kirimura, T. Umedu, T. Higashino, Efficient acquisition of local traffic information using inter-vehicle communication with queries, in: 2007 IEEE Intelligent Transportation Systems Conference, IEEE, 2007, pp.241–246.
27. T. Jeyaprakash, R. Mukesh, A survey of mobility models of vehicular adhoc networks and simulators, Int. J. Electron. Inf. Eng. 2 (2015) 94–101.
28. T. Zhang, H. Antunes, and S. Aggarwal, "Defending connected vehicles against malware: challenges and a solution framework," IEEE Internet of Things Journal, vol. 1, pp. 10–21, 2014.
29. V. Richard, Millimeter wave radar applications to weapons systems, USA Ballistic Research Laboratories, 1976.
30. V.D. Khairnar, S. Pradhan, Comparative study of simulation for vehicular ad-hoc network, preprint, arXiv:1304 .5181, 2013.
31. W. Zhang, H, Zeng, Y, Li, and X. Wang, "Polarimetric radar performance test of signal processing for anti-active jamming," in IET International Radar Conference, 2009, pp. 1–4.
32. X. Qiao, T. Jin, X. Qi, M. Zhang, S. Yuan, and Q. Zhang, "Anti-millimeter wave polarization agile active jamming," in Proceedings of the International Conference on Microwave and Millimeter Wave Technology, 2007, pp. 1–4.
33. Y.P. Fallah, C. Huang, R. Sengupta, H. Krishnan, Congestion control based on channel occupancy in vehicular broadcast networks, in: 2010 IEEE 72nd Ve-hicular Technology Conference-Fall, IEEE, 2010, pp.1–5.