International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S, August 2019

Evasion Attack on Text Classified Training Datasets

D. Suja Mary, M. Suriakala

ABSTRACT--- Machine learning algorithms are widespread used in real world training data classification and detection malware. The learning algorithms to detect malware adversarial manipulated training datasets in evasion. The evasion attacker has certain knowledge on training datasets either internal in deploying time attack or external attack do based on adversarial knowledge. Evasion attack targeted document properties features malware. To present this paper, to do an evasion attack on collected text documents using extraction keyword and find mean words using Naive Bayes models . Also to analyses different machine learning algorithms classification on evasion attacked training datasets and discussed defense methods to prevent training dataset from evasion attack.

Keywords Adversarial learning, Machine learning, malware, evasion attack.

I. INTRODUCTION

The adversary trained the sample datasets to make fool the machine learning algorithm accepting wrong decisions is known as Evasion Attack [1]. An attacker to make a small crafted noise in the machine learning classification testing time, the classifier prediction lead incorrect [2]. The adversary brought the normal clean training datasets. He launched the sample training datasets to the online classification T and observed its prediction of each sample s trained as T(s). The adversary paired(s,T(s)) to trained in the machine learning classification T′ make its functionality as T. Adversary produce an evasion attack on the sample training datasets T′. The PDF files attacked by Malware injection [4] the attacker injects the malicious data. The vulnerabilities used in most important PDF documents file formats [5].

An evasion attack targeted to misclassify training dataset samples [23]. Let's we assume M is a machine learning system and C be a benign input training dataset samples. The input sample C classified correctly by the ML system, and then the classification of M(C) has the correct decision maker. The adversary added to small noise A to the clean normal training datasets, and then the misclassification of M(A) has incorrect decision.

Manuscript published on 30 August 2019.
***** Correspondence Author (s)
D. Suja Mary.,M.Sc.,M.Phil., Part time Research Scholar, University of Madras, Assistant Professor, Department of Computer Applications, J.H.A Agarsen College,Madhavaram, Chennai-60, T.N, India (Email: dsuja2004@yahoo.com)
Dr. M. Suriakala., M.Sc.,M.Phil.,Ph.D., Assistant Professor, Department of Computer Science, Government Arts College for Men, Nandanam, Chennai-35. T.N, India (Email: suryasubash@gmail.com)

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

The security of evasion attack in machine learning training datasets has lot of challenges. The privacy preserving data mining [3] stated the security lacks in machine learning. This paper stated how to classified text document and need of the security for prevent machine learning algorithms performance. To separate the evasion attacked data from the adversarial samples and extract new classes from the original document file, then combine to the collected adversarial sample.

II. EVASION ATTACKS IN MACHINE LEARNING

An evasion attack performed by the adversary [7][8] providing attacked datasets as input that produced an incorrect output label. The Machine learning training dataset classifications take certain decisions in real world industries, economics, spam email filter etc. Evasion attack becomes to successful when no information known about the attack model and classification algorithm and training datasets has no longer access [9]. But the evasion attack classification system, adversary has the knowledge on the training datasets, Feature affected datasets and classification algorithm.

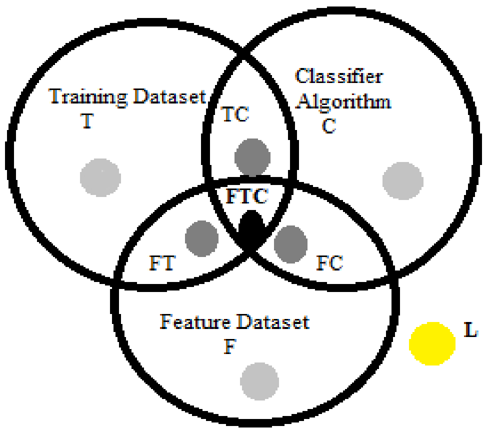


Figure 1 Adversarial Knowledge of evasion attack on classification system

The adversarial knowledge of evasion attack in the machine learning classification described in figure1. The letter L refers low knowledge of adversary about the three sets. The letter F refers the modified features with evasion attack. From the black market, the adversary brings malicious data and combines with normal benign samples.

This attacked feature dataset performed classification result in offline, then the final classification systems result submitted to the future work. The letters FT refers the adversary known about the feature evasion attacked classifier datasets results and the benign training dataset classification results. The letters FC refers to the adversary no knowledge about training dataset but know about feature datasets and classification algorithms execution results. The letters FTC refers the adversary has the chance to do evasion attack based on the three classification system datasets.

Malware Evasion PDF

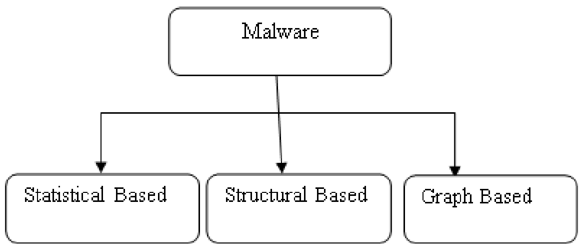


Figure 2. Malware Detection Categories

PDF file formats are the most popular vulnerable targeted attacks [5]. Detecting malicious PDFs in Machine-learning system techniques are using the file's logical structure to accurately identify the malware [6]. Detection of malware classified into three categories [10].

Statistical properties of different PDF files generated in the statistical based malware. The Hidden Markov Models (HMM) detect gradually changed benign files [11] and similarity index [12]. Simple Substitution Distance (SDD) method used opcode sequence for unspecified executable file [13]. Metamorphic malware detection techniques find the Structural Based malware of internal structure modification [14]. Graph based malware techniques used to detect the opcode malware graph similarities [15].

Evasion Text Classification

Text classification techniques used to handle and arrange text data in a specific order [16]. Insertion attacks [17] are the exploited vulnerabilities to evasion attack. This attack to checking pattern match for the evaluations, it leads to incorrect decision. Code morphing technique changed the text and morphed in benign dataset [11].

Genetic Evasion

The adversary known about the training dataset model, but he doesn't know how many benign sample require to attack [23]. The genetic programming algorithm [24], find over fitting learning classifier and fix legal samples.

Black Box Evasion

In black box evasion the attackers don't know the learning algorithms, so he not specifies the modification of training datasets. Adversaries independently collected benign training set and substitutes misclassified [27] by targeted malicious samples. The adversary created black-box gradient for the alternate of white-box attack generation through gradient [26] method. The black-box evasion attack defense use substitute model gradient masking.

III. BACKGROUND AND RELATED WORK

An evasion attack is one of the well known machines learning attack [20]. The attacker adds small modifications to the benign samples such that the machine learning classifier predicts incorrect data with the benign samples. The attacks are not affected in the machine learning models, its produced false output while using attacked training datasets [21]. The adversarial modified inputs in the target model leads to misclassification [9]. Attacks are categorized into white box and black box attacks. The fast gradient sign(FGS) method include in the white box attack. MIMICUS [8] is another evasion attack algorithm to transform attacked training dataset in such a way of changed into benign training datasets, making hard detection of mimicry attack. The first-order approximation [28] to affect the output based on the changes of input training datasets. The defense against evasion attack on machine learning system use Dimensionality Reduction [29]. It has the technique Principal Component Analysis to show high dimensional data projects as low dimension.

IV. TEXT CLASSIFICATION USING NAIVE BAYS

Text can be lot of information, but scattered unstructured in nature. Using text documents for decision making and time consuming in business level, we have to turning the unstructured text into structuring text. The suitable algorithm to classify text documents into string of characters based on word stem technique. To set an attribute value for each classified text. To detect mean number of word in the training dataset text documents [18], the word W_i chosen from the document and compare to all training set data. In this paper for text classification experiments, data collected from "Reuters-21578 text categorization test collection Distribution 1.0". In this datasets contains totally 90 classes, training documents 7769 and test documents 3019. The training dataset words appear 13332 in the whole reuters documents.

4.1 Extract words from text documents

To extract the word from the collection documents of reuters training datasets in the following way.

$W_i \leftarrow$ "earn"

Most_similar(Positive) $\leftarrow W_i$

Count word in training set= $\sum W_i$

The text classification and mean word in training set for 90 classes sample are given below in the figure3.

© Copyright: All rights reserved.

International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S, August 2019

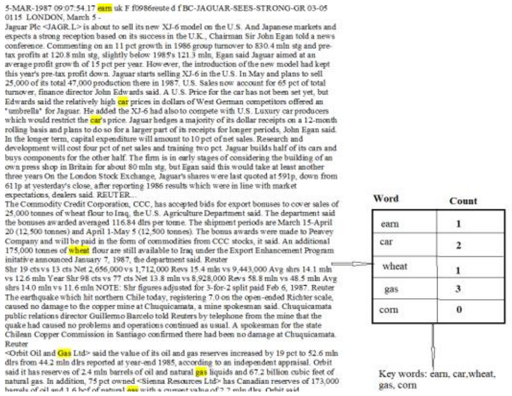


Figure 3 Extract word from text documents

The text document represented X and Y. X is a count of words and Y is a number of documents to collect training data. The label set to all classified text and the text are group by class names. For example the word wheat has the label number 4, also the word grain has the label 4, because wheat related to grain.

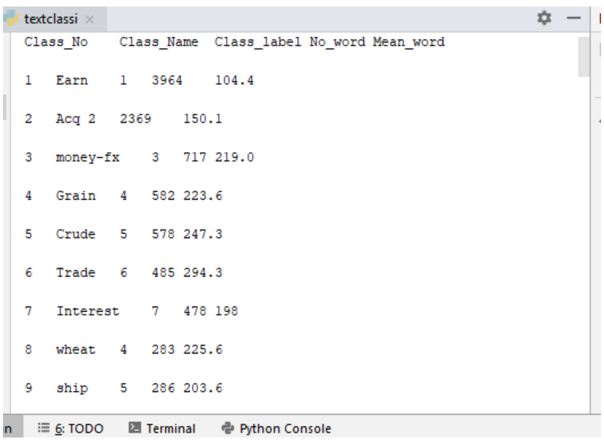
4.2 Mean word Calculation

Using Naive Bays method to find mean words. The probabilistic model refers to:

$\Pr(X/Y) \rightarrow$ X refer words W1,W2,..

$$\Pr(W_1,\dots,W_n/Y)=\prod_{i=1}^{n}\Pr(W_i/Y)$$

The result of word count, class label and mean word list out in figure 4.



Class_No	Class_Name	Class_label	No_word	Mean_word
1	Earn	1	3964	104.4
2	Acq	2	2369	150.1
3	money-fx	3	717	219.0
4	Grain	4	582	223.6
5	Crude	5	578	247.3
6	Trade	6	485	294.3
7	Interest	7	478	198
8	wheat	4	283	225.6
9	ship	5	286	203.6

Figure 4 Mean word using Naïve Bayes

4.3 Visualization of text classification

In this section explains the experiments conducted on classified text training set datas with machine learning algorithms. The experiments applied on reuters text training dataset on machine learning using python programming language. The reuters training datasets evaluated on different machine learning algorithms. Each numerical variable gives as input and the distribution of Box plot before and after evasion attack is shown in figure

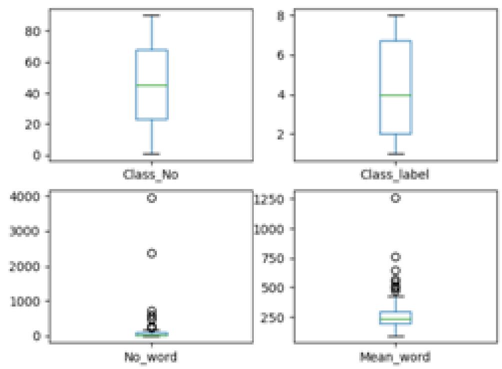
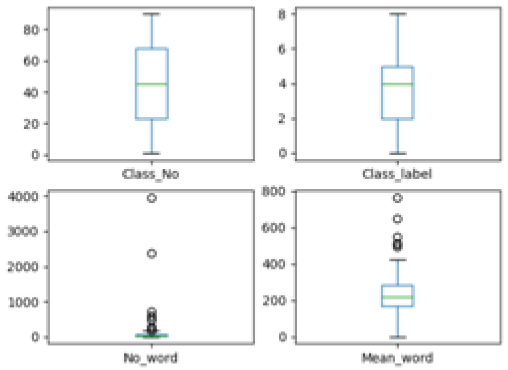


Figure 5a. Box Plot for each input variable before evasion attack



5b. Box Plot for each input variable after evasion attack

The histogram explains the difference of before and after the evasion attack on training datasets data distribution.

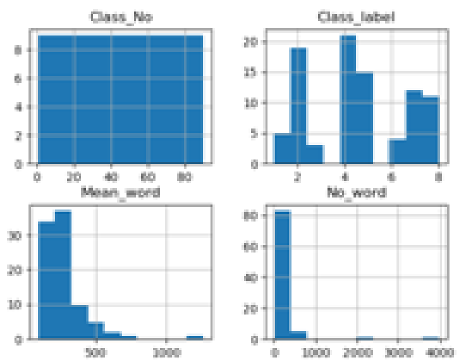
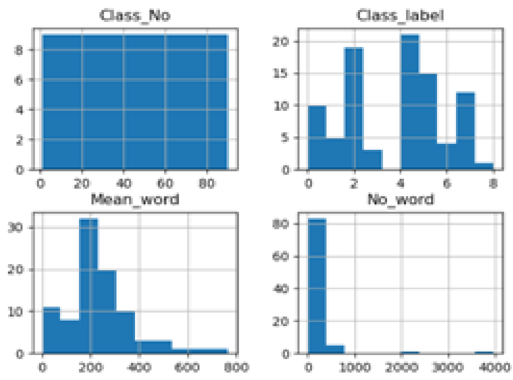


Figure 6a. Histogram before evasion attack

Retrieval Number: F10090886S19/2019©BEIESP
DOI:10.35940/ijeat.F1009.0886S19
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.



6b. Histogram after evasion attack

The Figure 7a & 7b shows scatter plots of all pairs of attributes helpful to spot structured relationships between input variables.

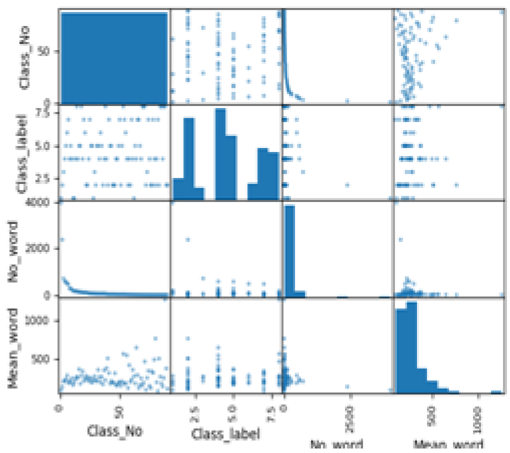
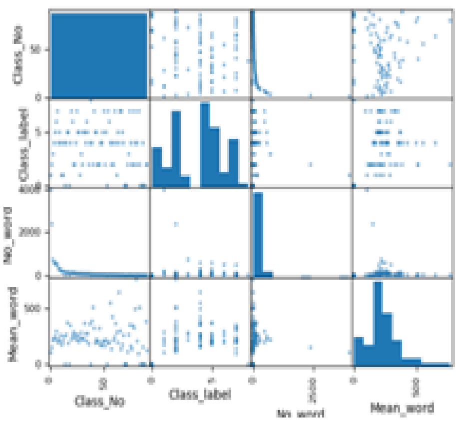


Figure 7a Scatter matrix for each input variable before evasion attack



7b Scatter matrix for each input variable after evasion attack

Experiment of Evasion Attack and ML Accuracy

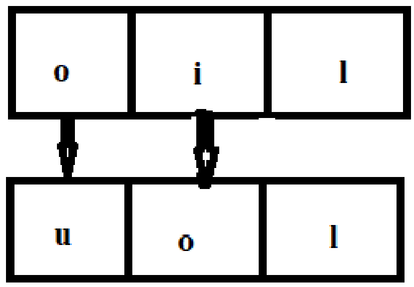
The evasion attacker goal is to simply modify on the training set to misclassify and the machine learning gives worst performance [19]. Evasion attack considered to pattern matching scheme [17]. In this paper, evasion attack done by text classification training datasets. The training datasets considered as D. The word W_i selected from D for doing attack. The evasion attack algorithm D : $W_i \leftarrow W_j$ to inject the replace keyword to the selected word in the text file. The algorithm1 will be representing the way of attack.

Algorithm 1. Evasion attack on text file

Input: Text classified training dataset with manual attribute names

1. Output: Evasion attacked data with training datasets
2. D=Obtain text classified file
3. For each W_i replaced in text do
4. $W_i \leftarrow W_j$
5. Append modified word in D
6. End for
7. Set keyword ← K
8. E=Obtain modified text file
9. For line in E
10. If K in line
11. Print line
12. End for

The above algorithm logic applied on the python programming and the attacked training datasets results display in figure8.



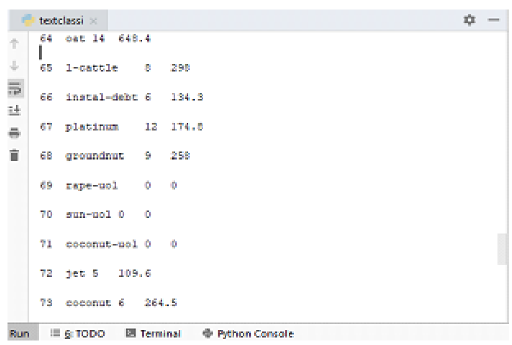


Figure 8. Evasion attacked training datasets

V. MEASURING EVASION ATTACK & RESULTS

The SVM and Logistic regression classification gives the high performance in classification text. They worked on training datasets and determine the text file datasets are malicious or benign.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

In the training dataset text classification both of the algorithms gives over fit protection. The evasion attack injecting small change in training dataset, it will damage the overall performance of SVM classifier [24]. The experiment result of table1 shows the performance of various machine learning algorithms.

Algorithm	Before Attack %	After Attack %
Logistic regression	63	49
Decision Tree	100	100
K-NN classifier	96	91
SVM	42	27

Table1 Comparison of Classification

The Decision tree and K-NN machine learning algorithms gives better performance in text classification. But After attack they support for better classification. Logistic regression and SVM machine learning algorithms not suited for the evasion attacked text. The comparison of learning algorithms accuracy before and after evasion attack explains in figure9.

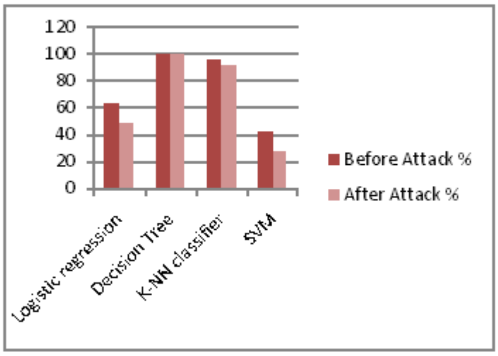


Figure 9 Performance of Learning algorithm after evasion attack.

Security against Evasion attack

The security against training dataset is a challenge for the researcher, because text dataset are unformatted. For the security purpose in this section, the collected training datasets are formatted in table form. The learner using the keyword to search the feature classes from the formatted table. The feature class has the word's count value is zero, then the learner can identify evasion attack happened on the training datasets. Evasion attack as detect the following steps.

Algorithm 2. Detect Evasion Attack

Input: D(x+δ)= t evasion attack text file. x is a class word, δ is an attack, t is an changed class label.

1. Output: classified evasion attacked datasets only.
2. Obtain evasion attack text file
3. Search t←0
4. If t←0 then
5. The class name text is attacked
6. Print attacked text
7. End if

The algorithm2 implemented and the result shown in figure10.

Figure 10 evasion attacked text data

To find out the difference between x and x+δ, then we rectify the attacked training datasets. To protect this attack, we not refer the text classified training datasets. The training set extract from the original documents and apply to the machine learning algorithms.

VI. FUTURE WORK AND CONCLUSION

The Large number of training dataset collection and large amount of unique features makes difficult to classify text documents. The learners use Deep Neural Networks (DNN) to train the training dataset. DNN classifier achieves best classification accuracy without adversarial interaction. To explore new type of evasion attack and make prevention methods for protect training datasets. Evasion attack on speech reorganization datasets are challenge on voice air attack and prevention. Random Forest algorithm in machine learning implements the training dataset different levels and produce more accurate. This paper presented how to evasion attack doing on text classified training dataset and how to change the performance of machine learning algorithms. The attacked datasets identified by using class labels. The original text training set extracted through python program from the document files, so the original performance of the learning algorithms prevented.

REFERENCES

1. Yi Shi and Yalin E. Sagduyu, "Evasion and Causative Attacks with Adversarial Deep Learning" Cyber Security and Trusted Computing, Milcom 2017 Track 3.
2. XiaoyuCao, Neil Zhenqiang Gong, "Mitigating Evasion attacks to Deep Neural Networks via Region-based Classification" Annual Computer Security Applications Conference, ACSAC 2017, USA.
3. Yehuda Lindell and Benny Pinkas "Privacy Preserving Data Mining", in International Cryptology Conference on Advances in Cryptology (CRYPTO), 2000.
4. Dhivya 1, Dharshana R, Divya V, "Security Attacks Detection in Cloud using Machine Learning Algorithms" International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 02 | 2019.
5. Charles Smutz, Angelos Stavrou , "Malicious PDF Detection using Metadata and Structural Features" ACSAC '12 Dec. 3-7, 2012, Orlando, Florida USA.

© Copyright: All rights reserved.

6. Srndic, N. Laskov, P. “Detection of malicious pdf files based on hierarchical document structure”. In: Proc. 20th Annual Net. & Dist. Sys. Sec. Symp. (2013)
7. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. “ Adversarial examples in the physical world” Workshop track - ICLR 2017.
8. Nedim Srndi.c and Pavel Laskov, “Practical Evasion of a Learning-Based Classifier: A Case Study” IEEE Symposium on Security and Privacy, 2014.
9. Kenneth T. Co, “Bayesian Optimization for Black-Box Evasion of Machine Learning Systems” Imperial College London, Sep 2017.
10. Tanuvir Singh, Fabio Di Troia, Visaggio Aaron Corrado, Thomas H. Austin, Mark Stamp, “Support vector machines and malware detection” J Comput Virol Hack Tech, Springer-Verlag France 2015.
11. Annie H. Toderici, Mark Stamp, “Chi-squared distance and metamorphic virus detection” J Comput Virol (2013) 9:1–14, Springer-Verlag France 2012.
12. Wing Wong and Mark Stamp, “Hunting for Metamorphic Engines” Journal in Computer Virology 2(3):211-229, November 2006.
13. Gayathri Shanmugam, Richard M. Low,Mark Stamp, “Simple substitution distance and metamorphic detection” Journal of virology and Hacking techniques, Volume 9,pp 159-170, 2013.
14. Jared Lee, Thomas H. Austin and Mark Stamp, “Compression-based analysis of metamorphic malware” Int. J. Security and Networks, Vol. 10, No. 2, 2015.
15. Neha Runwal , Richard M. Low , Mark Stamp, “Opcode graph similarity and metamorphic detection” J Comput Virol (2012), 8:37-52.
16. Thorsten Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features” University at Dortmund, Germany.
17. Thomas H. Ptacek, Timothy N. Newsham, “Insertion, Evasion, and Denial of Service: Eluding Network Intrusion Detection” Secure Networks, Inc., 1998
18. https://www.freecodecamp.org
19. Paolo Russu, AmbraDemontis, Battista Biggio, Giorgio Fumera, Fabio Roli, “Secure Kernel Machines against Evasion Attacks” AISec’16, 2016.
20. Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and JD Tygar, “Adversarial Machine Learning” 4th ACM Workshop on Artificial Intelligence and Security, October 2011.
21. Katja Auernhammer, Ramin Tavakoli Kolagari, Markus Zoppelt, “Attacks on Machine Learning: Lurking Danger for Accountability” CEUR-WS.org. Vol-2301.
22. Fabiano Dalpiaz, Elda Paja, and Paolo Giorgini. Security Requirements Engineering: Designing Secure SocioTechnical Systems. MIT Press, 2016, p. 224.
23. Octavian Suciu Radu Marginean, Yigitcan Kaya, Hal Daum, “Technical Report: When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks” Arxiv.org, March 2019.
24. Battista Biggio, Blaine Nelson, Pavel Laskov, “Poisoning Attacks against Support Vector Machines” Proceedings of the 29th International Conference on Machine Learning”, UK, 2012.
25. Weilin Xu, Yanjun Qi, and David Evans, “Automatically Evading Classifiers” In Network and Distributed System Security Symposium 2016 (NDSS), San Diego, February 2016.
26. Arjun Nitin Bhagoji , Warren He, Bo Li , and Dawn Song, “Practical Black-box Attacks on Deep Neural Networks using Efficient Query Mechanisms” published ECCV 2018.
27. Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, “Practical Black-Box Attacks against Machine Learning” ASIA CCS ’17, April 02 - 06, 2017.
28. Ian Molloy, Mathieu Sinn, and Irina Nicolae, “Adversarial Machine Learning”, ECML/PKDD, September 2018.
29. Arjun Nitin Bhagoji, Daniel Cullina, Prateek Mittal, “Dimensionality Reduction as a Defense against Evasion Attacks on Machine Learning Classifiers” scmanticscholor.org, Apr 2017.

Retrieval Number: F10090886S19/2019©BEIESP
DOI:10.35940/ijeat.F1009.0886S19
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.