International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.201.01

Effectiveness of Machine Learning Based Network Security

Sarvottam Dixit
Professor in Mewar University, Chittorgarh, Rajasthan, India

Aaina
Research Scholar in Mewar University, Cittorgarh, Rahasthan, India

Abstract—I and II examinations held by the employment opportunities.We are concerned with the safety of knowledge technology mainly because certain of our knowledge must, for legal and competitive reasons, be safeguarded only against illegitimate access and all the knowledge that we store and pertain to must be safeguarded against accidental or intended alteration and must be made available promptly. The genuineness (proper attribution) of everything we generate, distribute and acquire must be established and maintained. Lastly, if inadequate safety procedures can disrupt our networks, we might face litigation actions; if we neglect to enable other parties to suffer damage through our systems that have been compromised, there could be further serious legal difficulties. Today networks operate mission-critical corporate operations both internally and externally dangers need security. In this paper we discussed and analysis about normal and anomaly parameters use in logged_in networks.

Index Terms—Case base reasoning, decision tree, knowledge base system, neural network, WEKA,

I. INTRODUCTION

Machine Learning (ML) attractiveness and generality are on the rise. Existing processes are being enhanced and they are widely recognized for their understanding and response to actual challenges. This has contributed to the use of machine learning in many fields, such as computerized recognition healthcare analytics, gaming, and the commercialization of social media[1]. Machine learning approaches are the greatest alternative for traditional rules and even connected devices in some scenarios[2]. This development also affects the field of network protection as ML components are added to some detecting systems[3]. While designing a fully automated computer securities strategy is still a far aim, Security and Network Operating Center (NOC and SOC) first level operators may benefits from machine-based monitoring and transformational leadership enhances the motivation. This document is aimed particularly to private investigators and tries to evaluate their present development, evaluate their principal shortcomings and suggest potential areas for development [4]. Our analysis is based on a comprehensive existing literature and comprehensive tests on real, major companies and networking devices. Other research publications evaluate ML internet safety systems, evaluating one particular requirements and tend to focus rather than on protection operations on artificial intelligence (AI). In the assessment, we omit commercial goods based on machine education (or the misused AI term), as manufacturers do not disclose their methods and prefer to ignore problems and limits [5]. First, we offer a categorization of safety techniques in the machine learning network. In addition, we correlate the algorithms categories found on three issues, where machine learning is now used: intrusion detection, analytics of malware, image classification and spoofing. Finally, the key constraints of traditional technologies are analyzed. Our study emphasizes the advantages and disadvantages of various techniques, notably for false positively or negatively alarms. Furthermore, we note a common belief in the complexities of maintaining ML information security systems due to an absence of public information and labeled training data and to the length of time necessary to complete operations in a field defined by continual changes. Recent findings highlighting the efficacy of adverse events[6] in avoiding ML detection are also being considered. The disadvantages identified pave the roadmap for future changes required by ML parts in order to be completely embraced in security management technologies [7].

II. LITERATURESURVEY

Pattern architecture and selecting are the highest quality stage in the processing of data to turn connectivity transport information into relevant attribute matrices that enhance detectability to successfully use machine learning to intravention detection or other network security applications [8]. A mix of technical knowledge and automating approaches is often used to clean it up, engineer, minimize and pick the most important

characteristics. Some research has been reported to evaluate engineering approaches in domains other than video surveillance, however most effort focuses on dimensionality reduction [9]. Work in investigates the influence on numerical characteristics of numerous inclusive growth and sustainable development. The studies identify the Probability Sketch Array, which permits their system to anticipate the capability of a change of the characteristics to increase predictive performance[11]. The study carried out in[4] investigates several approaches for improving fraudulent credit card detection. They are trying a range of approaches here that employ flexible time frames and other techniques to include user behavior. Both investigations deviate from our domainwork and the data types we develop. For the purpose of data retrieval and technology, numerous studies conducted standardization, encryption and other methods of transformations (index variables, conditions, etc.) for transforms network traffic characteristics into data that is useful for machine learning algorithms [12]. The authors utilize standardizing of the Z-score in the numerical values of KDD'99 and employ one-hot encode for continuous and multivariate information The issue is the expansion of huge input matrices in this technique. In, writers integrate these quantities with intensity values. This leads to more condensed selected features but depends substantially in comparing to the community on the quality of the underlying specimens. Davis derives embedded characteristics from string data that convert them into average components (frequency vector of ASCII values). No other technical method is addressed in all of these publications and the suitability of these techniques is thus not assessed [13]. In the authors experimented with the KDD'99 dataset on features technique for special fields. In three approaches are compared to a 'arbitrary' methodology analogous to the encoding of labels. For assessments, they employ indication (one-hot) parameters, conditional likelihoods (N-dimensional vector) and separation split value [14]. All three approaches are more efficient than the one-hot method, although the supremacy of all three is not demonstrated. They function in two mentioned categories only: methodology, services and the flags. They are restricted compared to ours. A compare of the technology in IPv4 addresses was conducted by the authors in. They contrast a description of 32-bit vectors, a description of 4-bytes and an extension-dectet. In the machine-learning technology (technique in which cross-octet information is integrated). To test their approaches, they utilize a self-created dataset of good and harmful webpages [15]. They do not use conventional approaches such as one-hot programming or combine additional functionalities into their known malicious detection technique. Efforts have been performed in to assess how the accuracy of four classificator is affected by ten approaches of quantitative modeling (e.g. counts, logarithms, square roots and polynomials) [16]. The results show that some approaches were used by neural networks and SVM classificators, and Similar approaches benefitted from variable selection and Random Forest algorithms. Our research varies from existing research by extending it to quantitative and linguistic characteristics and by using a new data set. A lot of tasks were done to conduct function selection beyond functional engineering. Primary component analyzes for feature extraction to relieve the dimensionality curse have been frequently utilized [17]. For example, Li et al. [18] utilized the random method of ascending mutations, altered, whereas the Markov blanket model was applied by Chebrolu et al. [19] The identification of functionalities is an essential constituent of machine learning systems but relies on well-entered features to ensure accurate outcomes.

III. ML TECHNIQUES

1.3.1 Naïve Bayes: - The naive model from Bayes is a much reduced Bayesian model of probability. The naive Bayes classification is based on a substantial assumption of independence. This indicates that one attribute's likelihood does not impact the other. The naive Bayes classifier makes distinct predictions of 2n!, given a number of n characteristics [19]. The findings of the naïve classification of the Bayes are, unfortunately, often right. The report explores how the naïve classification of the Bayes works effectively and why. The inaccuracy is shown by three factors: data noise, partiality and inconsistency. The distortion level of data sets can simply be reduced if excellent training data are selected. By using the machine learning method the training data should be split into several categories. The mistake is because groups are excessively big in the learning algorithm. Due to the groups that were too tiny, variability is the mistake [10].

1.3.2 Multilayer Perceptron:- The media access control protocol (MAC) is the linear transformation (MLP) to protect wireless sensor networks based in CSMA against adverse denial of services operations. The MLP increases the security of a WSN by persistent surveillance of a parameters that indicate unexpected fluctuation in the event that an attacker, and when a suspected factor surpasses a predetermined threshold level MLP alarm the MAC layer and physical layer in sensor mode. The MLP training is carried out using back spread and Radial Basis Function Network (RBFN) [20].

1.3.3 Instance Based Learning (IBK):- Machine learning is a class of clustering models that compares examples of issue (also referred to as "memory-based learning") with occurrences that are kept in memory. Instead of explicit generalization [22]. As a specific product is delayed, the calculation is frequently referred to as "lazy," and this technique is no longer used. These record (a subset of) their learning package and calculate distances or commonalities between current instance and the training examples when estimating a value/class for a subsequent observation [21].

1.3.4 Random Forest:- Random Forest is an ensemble technique based on the findings of a Randomized Decisions gathering. The selection of each tree in the "forest" is done using the bootstrapping technique. A subset of characteristics is also picked randomly for each node division, with the divisive variable generated via this subset. For categorization and averaging reversal, the projected value is the prevailing decision [23].
1.3.5 J48:- C4.5 is an ID3 successor created by Ross Quinlan and integrated with Java in WEKA as J48. They all take a gloomy and top-down decision-making style. It is used to classify incoming data thus according previous information in which (training data set). The decision tree induction starts with a dataset (training set) that is divided into smaller divisions at each and every nodes and therefore follows a recursive division and a technique of conquest. A set of characteristics is also given in additional to a data set, which would be a collection of elements. The knowledge relating to the item might be an event, an action and the characteristics. A classification accuracy that specifies if an item corresponds to a specific class is assigned to each tuple in the data collection. Further separation may only be done if the duplicates in various classes decrease [24].
1.3.6 Decision Stump:- A stumps is a one stage clustering algorithm machine learning model. In other words, it is a predictive model with one individual element (root) linked to the output layer instantly (its leaves). A stump for a decision predicts the importance of just one input characteristic. They are quite often sometimes known as 1-rules [25].

IV. ML BASED N/W SECURITY

The machine-learning techniques will perform vector representation data to identify the greatest compatible response with the user's query, meaning the response of ViEA to the user's inquiry.

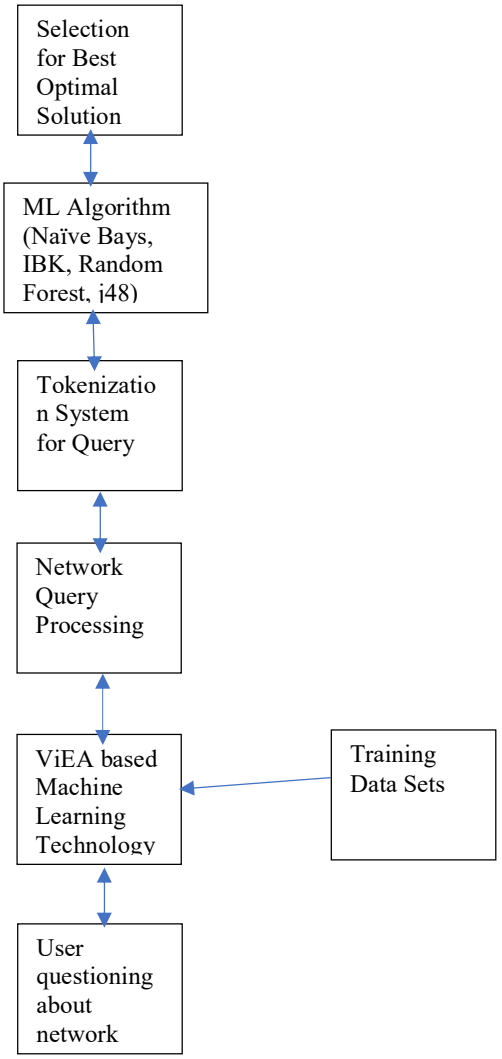


Figure 1.1 Structure of machine learning based network security

It comprises KNN, Naïve Bayes and SVM for the issue of classification. For KNN, the next (or similar) topic to the user query is found and the response is selected. The Multilayer perceptron Bayes system is utilized for the reply choosing for Naïve Bayes since this approach is suited for the sequence in Figure 1.1.

V. RESULT ANDDISCUSSION

This result compile in machine learning tools weka

A Training Data Sets

The Training Data sets KDDTrain collected from this link https://www.unb.ca/cic/datasets/dohbrw-2020.html. To analyses the datasets most of system login to calculate the class functional value (anomaly and normal). 25192 instances used in Table 1 describe below

Table 1.1 Parameters of data sets

class
count
diff_srv_rate
dst_bytes
dst_host_count
dst_host_diff_srv_rate
dst_host_rerror_rate
dst_host_same_src_port_rate
dst_host_same_srv_rate
dst_host_serror_rate
dst_host_srv_count
dst_host_srv_diff_host_rate
dst_host_srv_rerror_rate
dst_host_srv_serror_rate
duration
flag
hot
is_guest_login
is_host_login
logged_in
num_access_files
num_compromised
num_file_creations
num_outbound_cmds
num_root
num_shells
protocol_type
rerror_rate
root_shell
same_srv_rate
serror_rate
service
src_bytes
srv_count
srv_diff_host_rate
srv_rerror_rate
srv_serror_rate
su_attempted
urgent
wrong_fragment

B Naïve Bayes based confusion matrix

```
=== Confusion Matrix ===

     a     b   <-- classified as
 12272  1177 |     a = normal
  1445 10298 |     b = anomaly
```

C Lazy IBK based confusion matrix

```
=== Confusion Matrix ===

     a     b   <-- classified as
 13393    56 |     a = normal
    85 11658 |     b = anomaly
```

D Random Forest based confusion matrix

```
=== Confusion Matrix ===

     a     b   <-- classified as
 13439    10 |     a = normal
    41 11702 |     b = anomaly
```

E J48 based confusion matrix

```
=== Confusion Matrix ===

     a     b   <-- classified as
 13389    60 |     a = normal
    51 11692 |     b = anomaly
```

F Decision Stump based confusion matrix

```
=== Confusion Matrix ===

     a     b   <-- classified as
 12517   932 |     a = normal
  1031 10712 |     b = anomaly
```

G Information Gain

The attribute information gain, A is computed as follows:

$$gain = info(T) - \sum_{i=1}^{s} \frac{|T_i|}{|T|} \times info(T_i)$$

Where, T and Ti (i=1 to s) are the T subsets that include for the attribute A positive implication. Info (T) has been described as the entropy capability

$$info(T) = -\sum_{j=1}^{Nclass} \frac{freq(C_j,T)}{|T|} \times log_2\left(\frac{freq(C_j,T)}{|T|}\right)$$

In fact, the decision tree created is enormous, making it impossible to interpret. In C4.5, by changing the trust level, we may optimize the decision tree.

Table 1.2 Confusion Matrix

	Predicted Class +ve	Predicted Class -ve
Actual Class +ve	TP	FP
Actual Class -ve	FN	TN

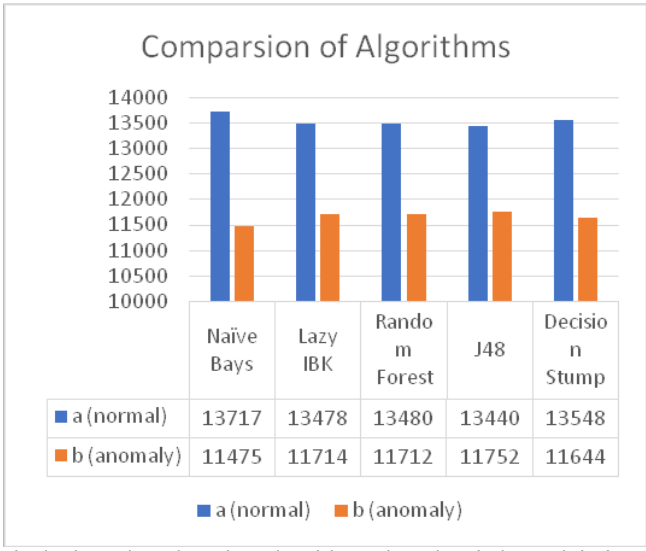
H Confusion Matrix

The effectiveness table of an algorithms is the confusion matrix. The confusion matrix is shown in Table 1.2: A confusion matrix includes four measurement elements, e.g. true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

- True Positive: TP displays the right prediction number of the instance to be classified.
- True Negative: TN displays the number of wrong forecasts of another class instances
- False Positive: FP displays the number of wrong predictions of an element of the same class.
- False Negative: FN displays the right number of predictions of a different class instance.
- Accuracy: - Accuracy of decision tree denotes the proportion of occurrences properly categorized. Precision is computed using the following confusion matrix:

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN}$$

When the result derived from confusion matrix to show in graphical representation



In Naïve Bays algorithm is the best than the other algorithm when data is logged_in in well form and show less anomaly compare than the other machine learning algorithm

VI. CONCLUSION

Machine training techniques for various technologies are widely used and are also taken for internet national security reasons. It is thus necessary to assess whether or which technique categories may provide sufficient results. Three significant information security challenges are discussed in the following ways: intrusion detection, vulnerability scanning and sentiment analyzation we offer originally an innovative classification of the most prominent ML strategies and illustrate which of them now apply. We discuss then a number of problems which impact information security implementation of ML. Our findings show that current machines are still impacted by a number of weaknesses, which impair information security efficiency. Every technique is subject to adverse assaults and requires constant retraining and careful tweaking of parameters that cannot be automated. In addition, the detection accuracy is embarrassingly low, particularly when the same classification is used to identify various risks; a potential mitigating may be done utilizing various ML classifiers to detect distinct threats. Important advancements may be predicted, particularly in view of recent and prospective domain adaptation developments. Our goal is to help the protection manufacturer's operations by using machine-learning technologies to automate some jobs but benefits and disadvantages must be understood. The autonomy of Existing methods must not be underestimated because the lack of human oversight may also allow the infiltration, robbery and even the destruction of a company by qualified adversaries.

CONFLICT OF INTEREST

I have collected data from this link (https://www.unb.ca/cic/datasets/dohbrw-2020.html). It is available and freeware on google data set. No any type of conflict generate in future.

AUTHOR CONTRIBUTIONS

Prof. Sarvottam Dixit deep knowledge in machine learning. They have imply dataset in ML techniques to get better results to learn Ms. Aaina from them to use their learning techniques than to write this paper.

ACKNOWLEDGMENT

I highly thanks to google data sets and my guide Prof. Sarvottam Dixit and they help me in this research paper.

REFERENCES

[1] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” Science, 2015.
[2] A. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” IEEE Communications Surveys & Tutorials, 2015.
[3] E. Blanzieri and A. Bryl, “A survey of learning-based techniques of email spam filtering,” Artificial Intelligence Review, 2008.
[4] J. Gardiner and S. Nagaraja, “On the Security of Machine Learning in Malware C8C Detection,” ACM Computing Surveys, 2016.

[5] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial Machine Learning," in ACM workshop on Security and artificial intelligence, 2011.
[6] F. Pierazzi, G. Apruzzese, M. Colajanni, A. Guido, and M. Marchetti, "Scalable architecture for online prioritization of cyber threats," in International Conference on Cyber Conflict (CyCon), 2017.
[7] J. Kim, J. Kim, H. L. T. Thu, and H. Kim, "Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection," in IEEE International Conference on Platform Technology and Service (PlatCon), 2016.
[8] P. Torres, C. Catania, S. Garcia, and C. G. Garino, "An analysis of Recurrent Neural Networks for Botnet detection behavior," in IEEE Biennial Congress of Argentina (ARGENCON), 2016.
[9] G. E. Dahl, J. W. Stokes, L. Deng, and D. Yu, "Large-scale malware classification using random projections and neural networks," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.
[10] Kuldeep Singh Kaswan, Jagjit Singh Dhatterwal"The Use of Machine Learning Sustainable and Resilient Buildings" in book entitled "Digital Cities Roadmap: IoT-based Architecture and Sustainable Buildings", Published by John Wiley & Sons, July 2020, , ISBN No. 978-1-119-79159-1.
[11] R. Pascanu, J. W. Stokes, H. Sanossian, M. Marinescu, and A. Thomas, "Malware classification with recurrent networks," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.
[12] M. Z. Alom, V. Bontupalli, and T. M. Taha, "Intrusion detection using deep belief networks," in IEEE National Aerospace and Electronics Conference (NAECON), 2015.
[13] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), 2016.
[14] Y. Li, R. Ma, and R. Jiao, "A hybrid malicious code detection method based on deep learning," International Journal of Security and Its Applications, 2015.
[15] W. Hardy, L. Chen, S. Hou, Y. Ye, and X. Li, "DL4MD: A Deep Learning Framework for Intelligent Malware Detection," in International Conference on Data Mining (DMIN), 2016.
[16] G. Tzortzis and A. Likas, "Deep belief networks for spam filtering," in IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2007.
[17] G. Mi, Y. Gao, and Y. Tan, "Apply stacked auto-encoder to spam detection," in International Conference in Swarm Intelligence, 2015.
[18] M. Stevanovic and J. M. Pedersen, "An efficient flow-based botnet detection using supervised machine learning," in IEEE International Conference on Computing, Networking and Communications (ICNC), 2014.
[19] S. Ranjan, Machine learning based botnet detection using real-time extracted traffic features, Google Patents, 2014.
[20] B. Rahbarinia, R. Perdisci, A. Lanzi, and K. Li, "Peerrush: mining for unwanted p2p traffic," in International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, 2013.
[21] A. Feizollah and e. al, "A study of machine learning classifiers for anomaly-based mobile botnet detection," in Malaysian Journal of Computer Science, 2013.
[22] M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou, S. Abu-Nimeh, W. Lee, and D. Dagon, "From throwaway traffic to bots: detecting the rise of DGA-based malware," in USENIX Security Symposium, 2012.
[23] Kuldeep Singh Kaswan, Jagjit Singh Dhatterwal "Machine Learning and Deep Learning Algorithms for IoD" in book entitled "Internet of Drones: Opportunities and Challenges" in "Apple Academic Press (AAP), Canada, Publishing date Feb 2022, Hard ISBN: 9781774639856
[24] C. Annachhatre, T. H. Austin, and M. Stamp, "Hidden Markov models for malware classification," Journal of Computer Virology and Hacking Techniques, 2015.
[25] J. Demme, M. Maycock, J. Schmitz, A. Tang, A. Waksman, S. Sethumadhavan, and S. Stolfo, "On the feasibility of online malware detection with performance counters," in ACM SIGARCH Computer Architecture News, 2013.
[26] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in ACM Proceedings of the Anti-Phishing Working Groups, 2007.
[27] G. Xiang, J. Hong, C. P. Rose and, L. Cranor, "Cantina+: A feature-rich machine learning framework for detecting phishing web sites," ACM Transactions on Information and System Security (TISSEC), 2011.
[28] G. Apruzzese, M. Marchetti, M. Colajanni, G. Gambigliani Zoccoli, and A. Guido, "Identifying malicious hosts involved in periodic communications," in IEEE International Symposium on Network Computing and Applications (NCA), 2017.
[29] F. S. Tsai, "Network intrusion detection using association rules," International Journal of Recent Trends in Engineering, 2009.
[30] F. Bisio, S. Saeli, L. Pierangelo, D. Bernardi, A. Perotti, and D. Massa, "Real-time behavioral DGA detection through machine learning," in IEEE International Carnahan Conference on Security Technology (ICCST), 2017.