Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Malicious Network Traffic Detection Using Machine Learning Algorithms

Dr. N. Mangathayaru[1] and L. Sreenidhi[2]

[1]Professor, VNR Vignana Jyothi Institute of Engineering & Technology mangathayaru_n@vnrvjiet.in
[2]MTech-(PG Student), VNR Vignana Jyothi Institute of Engineering & Technology sreenidhilakkampalli@gmail.com

September 10. 2024, submitted,
November 15. 2024, accepted,
December 18th 2024, published

The misuse of networks and the detection of malicious traffic is a primary role of Anomaly detection systems, since these sing out the flows of data which is considered to be unusual in relation to the normal. In parallel with the establishment of the advanced network technology, getting over the complicated attack pattern is the weakness that the traditional detection systems have. These challenges are resolved in this study with the use of machine learning and deep learning algorithms to improve the classification of impacts from malicious traffic. Using the CIC IDS 2017 dataset for binary classification and the CIC IDS 2019 dataset for multiclass classification, we test the effectiveness of various algorithms: K- Nearest Neighbors (KNN), Gradient Boosting, Random Forest, and a Voting Classifier that compares a Boosted Decision Tree Classifier and a Bagging Classifier based on Random Forest. Also, models like Deep Neural Network (DNN) and Torch-based neural networks (Torch NN) are also incorporated. This is based on performance evaluation parameters of accuracy, precision, recall and F1 score. Such findings show that the Voting Classifier performs the best across both sets, which asserts the classifier's ability to identify more simple and complex network threats in a more efficient and accurate manner.

“**Index Terms –** Machine Learning, Deep Learning, Malicious Network Traffic Detection, KNN, GB, RF, VC, DNN, Torch NN”.

1. INTRODUCTION

While with the emergence of more digital communication channels and online service providers, networks' susceptibility and vulnerability to attacks are proportional accordingly. Malicious network traffic has been named as a common cause or threat to computer security and data integrity detrimental to user's privacy and organization's financial loss. The current threat of these attacks has been addressed by using Anomaly Detection systems which focus on recognizing the difference between the abnormal traffic and the actual traffic. This forces the need for better traffic analysis than what traditional detection systems can detect, offer and or provide because as attackers change their tactics, these system will be forced to lag behind.

The function of the IDS's most crucial element, the anomaly detection, is to identify the activity that is different from the normal one indicating possible intrusion. Signature-based detection of known threats seems to be one of the most traditional IDS methods that employ certain rules in order to detect known threats; however, traditional methods struggle when they have to face new and unknown threats [GL4 9]. Due to this limitation, use of machine learning (ML) and deep learning (DL) has become more important as they are in a position to analyze huge amount of network traffic, other factors such as identifying the fine grained anomalous activities and enhance the performance of detection and accuracy over time. These techniques have been used by the researchers to create models that tend to change with time as different attacks are practiced in the networks making it very useful in the present network systems [10].

Anomaly detection using ML involves feeding large amounts of data and training the system to develop models that can make good predictions on other unseen data hence detect an array of anomalous behaviours that could imply threats [11]. Some of the most used ML algorithms used for DDoS detection are decision trees, SVM, and ensemble methods such as AdaBoost and Random Forest especially for classification in intrusion detection in computer networks [12]. Most baseline performance techniques and algorithms in the areas of ML have been empowering but CNNs and RNNs give further utilization potential of network traffic applications due to its additional proficiency in pattern revolution for Internet traffic, which has some shortcoming than conventional algorithms [13].

For conducting anomaly detection research, both CIC IDS 2017 and CIC IDS 2019 datasets provide large volumes of raw network traffic on which binary as well as multiclass classification can be performed. These datasets cover variety of attack types like DoS attack, brute force attack and botnet traffic that makes it useful for training and testing models. By using such datasets, the scientifically grounded creation and calibration procedures of the systems and the anomaly detection models, are achieved under realistic conditions, and the models' resistance against various types of contemporary attacks is guaranteed [14].

Here in this project, we are trying to improve the performance of network anomaly detection using a combined approach of both ML and DL. In this framework, it is an attempt to develop a framework that can detect various types of malicious activities in a complex network environment and can contribute towards effective proactive cybersecurity measures and enhancing the methods of anomaly detection [15].

2. LITERATURE SURVEY

The recent events show that computer networks became more and more targeted by attacks so, this work underlines the need for an improved anomaly detection to protect computer networks from threats like data breaches, financial loss, and loss of privacy. Anomaly detection systems are quite able to use various methods in order to segregate normal from abnormal network traffic, and the use of machine learning has dramatically helped advanced this area because of its ability to identify newer, more complex and constantly evolving patterns of malicious traffic.

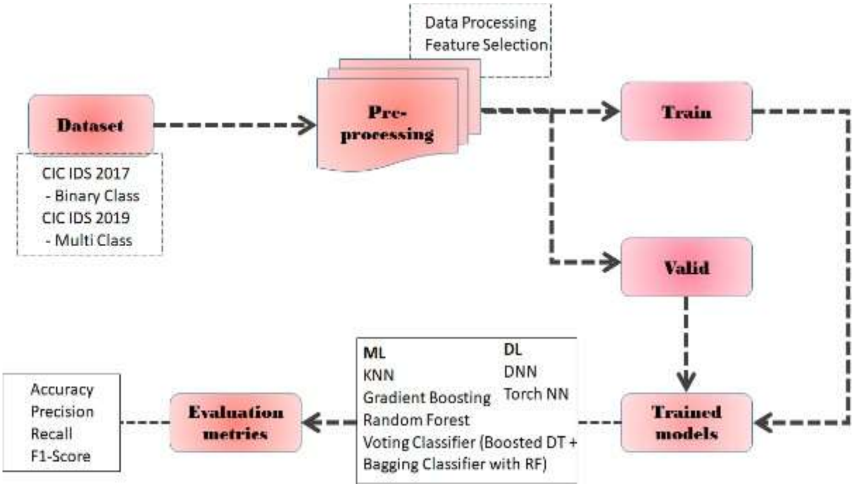
ML for intrusion detection in network traffic was investigated in [1], where the authors fitted supervised learning models to distinguish between normal traffic and traffic containing code that was deemed malicious. It established the basis for a significant number of improvements to network anomaly detection research by confirming that ML could still identify malicious traffic patterns despite their disguise. Wang, Fok, and Thing [2] built upon those studies by highlighting the approach of encrypted malware traffic identification, with a specific stress laid on the feature selection as well as extraction pre-requisites for the different model. It compared many works that proposed to detect encrypted traffic and showed that deep learning models are effective in cases where traffic fluctuations are intense. The study also noted that the limitations in the coverage of the dataset in relation to practical applications increases the false negatives though the work emphasized the need for diverse dataset.

In the context of cloud computing, Alshammari and Aldribi [3] studied the related problem of malicious network traffic detection. What they discovered was that conventional rule-based approaches did not work because the cloud environments they were implemented in immensely depended on shared resources and are highly dynamic. In Random Forest, and SVM, they were able to show that there was enhanced accuracy in detection when the datasets were balanced and sufficiently sampled cloud-based network traffic. Rajesh & Satyanarayana [4] similarly extended classification efforts in SCADA application domain specifically to evaluate the efficiency of the decision tree and ensemble learner techniques for recognizing anomalous operations in real-time and often, constrained computing milieu. However, their findings stressed the high detection rate should be sustained while the extra computational cost should be minimized in order to apply the developed techniques in industrial systems.

Shafiq et al. [5] discussed the issues arising with IoT where low power devices, different types of devices make it difficult to implement traditional anomaly detection. They developed their CorrAUC approach which was augmented with correlation-based feature selection to improve the identification of botnet traffic in IoT networks. The authors are reporting a high model accuracy with low complexity, probably due to the use of only core features, enabling real-time IoT application. The study pointed out that the demands for analytics algorithms mean that anomaly detection must be both highly efficient in terms of energy consumption and low in latency where IoT is becoming standard on networks.

The concept of IDS was introduced by Mukherjee, Heberlein and Levitt [6] in the prior study, where they have defined the different categories of IDS namely anomaly-based and signature based. Some of them pointed out that, despite high efficiency of signature-based systems in addressing known threats, the latter are inefficient when it comes to discovering new forms of threats, which is why designers and engineers insisted on using anomaly- based detection in addition. Thus, extending this concept, Garcia-Teodoro et al. [7] provided a survey of anomaly-based IDS and classified them with regard to the statistical, knowledge, and machine learning models. They observed that traditional statistical anomaly detection approaches lost their effectiveness outcompared to ML-based systems in constantly evoloving network condition, particularly when supported with large and incorporation datasets containing various types of attack.

Anomaly detection has been benefited from recent developments in semi-supervised learning. , for instance, Ashfaq et al. [8] proposed a semi- supervised fuzzy logic-based IDS that tries to solve the lack of training data by making the system learn from few labeled and even more unlabeled instances. This method was proven particularly beneficial in situations where attack patterns were less rigorous, such as in small and medium firms, where labeled data sets were scarce. Semi- supervised learning also contributed to diminishing false positives, as the function could distinguish between quiet specific benign and much noisier malicious network activities, even when only a few examples are available.



3. METHODOLOGY

i) Proposed Work:

The proposed system on the other hand focuses on prevention of malicious network traffic through the adoption of a sound anomaly detection framework that employs both machine learning and deep learning architectures. This system uses the CIC IDS 2017 and CIC IDS 2019 datasets for the binary and multiclass network traffic classification tasks. For our machine learning, we use K-Nearest Neighbors (KNN), Gradient Boosting and Random Forest as these algorithms have been reputed in the identification of anomalous sections. Moreover, a Voting Classifier is developed using the concepts of boosting a decision tree classifier and a Bagging classifier with random forest classifier for implementing more number of perspectives towards the detection of the same say output. For high dimensional data deep learning, the system uses Deep Neural Networks (DNN) and Torch-Based Neural Networks (Torch NN) for more accurate results. It helps in evaluating the performance measuring factor such as accuracy, precision, recall, F1 score and thereby gives a detailed view of each and every model that how effectively they are protecting the network from probable threats.

Fig 1 Proposed Model Architecture

It began with a dataset (CIC IDS 2017 or CIC IDS 2019), which has been preprocessed and divided into training and validation datasets. In this paper, the training set is used to train a number of ML and DL models including KNN, Gradient Boosting, Random Forest, DNN and Torch NN. The performance of the trained models is presented in terms of performance data on IT with the validation metrics being the accuracy, precision, recall, and the F1-score measures on the validation set. Last of it, the best model is chosen and it is saved as the trained one.

ii) Dataset:

From literature, it is established that CIC-IDS 2017 and CIC-IDS 2019 are the most popular datasets for implementing and comparing the performance of methods for network intrusion and anomaly detection. They are solutions developed by the Canadian Institute for Cybersecurity, which contains real-life network traffic and includes different types of cybers attacks such as DoS, Brute force, botnet, and infiltration. While CIC-IDS 2017 contains application-level traffic details derived from an ordinary corporate network, it offers both fundamental and advanced flow-based features for binary and multiclass classification. CIC-IDS2019 builds on this with further attack categories, richer traffic variability, and more recent protocols appended to it, making it more realistic. In combined, these datasets support the ethically sound training and validation necessary for using machine learning and deep learning algorithms focused on cybersecurity defense.

iii) Data Processing:

The processing of datasets from CIC-IDS 2017 and CIC-IDS 2019 starts with joining all files in the dataset in the dataset type. This merging step allow uniform data structure for extensive analysis on both datasets. Third, labeling mapping gives different attack types and normal traffic numeric labels for easy input and processing by the machine learning models. This process involves a process of association of categories of attacks with integer values in order to ease the training and classification. Lastly, it is converted to the Torch tensors, which is the basic data structure of PyTorch and facilitating GPU training. This tensor conversion guarantees compatibility with current PyTorch-based neural network models and enhances the models themselves and training processes.

iv) Feature Selection:

During attribute ranking we use the Select Mutual Information (MI) Classification model because it measures the dependency between each feature and the target class based on mutual information. This approach is used to sort out features that depend maximum on the output labels and therefore it minimizes the amount of data to be used in the modelling as much as possible and discards all other unnecessary data. MI quantifies the amount of information that two variables have in common: the extent to which knowing one reduces uncertainties about the other. Choosing features conveying the greatest amount of mutual information lead to the development of a simpler model that comprehensively captures the selected dataset's features and may be more accurate. It enhances the input space with a compression technique so as to direct the unnecessary inputs for classification towards machine learning models.

v) Training & Validation:

To prepare the dataset for model training and evaluation, we split it into training and validation sets using an 80:20 ratio. This also guarantees that 80% of the collected data is employed to train the model to identify patterns and relationship in the data. To achieve this, the model is trained on attack and normal traffic samples from various network conditions so as to allow generalization across the network settings.

That leaves 20% of the entire dataset for the validation set, which gives an additional data sample for checking the model's efficacy. It will allow us to calculate its performance and check whether it will overfit on new data, on the validation set.

vi) Algorithms:

The study utilized both machine learning and deep learning algorithms, as detailed below:

A) Machine Learning:

K-Nearest Neighbors (KNN) is an easy and basic example of instance base learning which categorizes data points in relation to those which are closest to it. In our case, for our malicious network traffic detection project, KNN has been used since it captures the ability to learn from patterns in the dataset to classify the network traffic as either normal or malicious. In this context, KNN makes it possible to estimate distances between discrete points in feature space, and thereby offers a 'perceived' method of flagging abnormal traffic situations.

Gradient Boosting is one of the ensemble learning techniques and learns to create models step by step, step by step correcting the mistakes of the models that have been built before. We use Gradient Boosting to improve the performance of identifying malicious traffic using a number of weak learners for building a strong classifier. It works well with respect to data distribution and owns good aspiration for precision and recall, making it possible to identify tiny abnormalities in the network traffic. This can be particularly great when it comes to working on classification because of the capabilities of the algorithm of enhancing on iterations.

Random Forest is an ensemble method that builds many decision trees as it is trained and returns a class in case of classification problems, which was voted on by most of the trees. Random Forest is used for increasing the detection rate of malicious network traffic due to the fact is less prone to overfitting and is capable of handling larger amount of randomly generated features. By collecting results of different trees, it improves the accuracy and possibility of generalization and is more applicable in situations with multiple sets of network traffic with different and rather complicated pattern.

Voting Classifier is an aggregation of various models through which an end result is produced aiming at high accuracy. To do this, we decide to employ a Voting Classifier that will use the power of a Boosted Decision Tree and a Bagging Classifier, then Random Forest. This alignments improves the detection of malicious traffic since it integrates their approaches, decreases probability of misclassification, and provides better picture of the network traffic behavior, thereby improving the performance of the detectors.

B) Deep Learning:

DNN: Conventional neural networks are composed of many interconnected layers of nodes, known as Deep Neural Networks (DNN) that learn features in data sets by using backpropagation. In the present study, DNNs are used due to their ability to learn complex interdependencies in large digital network traffic data dimensions. DNNs expand the potential of the system, as several innovative parameters such as non-linear activation functions, multiple hidden layers improve the perception of malicious activities that can be overlooked by standard algorithms. This capability is instrumental in comprehending and categorizing numerous patterns that characterise the behaviour of the network.

Torch Neural Networks (Torch NN), in particular, is an optimized project for building and training neural networks based on the PyTorch platform. In the present work, Torch NN is used due to its highly flexible computation graph and in order to quickly test different architectures for detection of malicious traffics in the network. The boosted support for GPU increases the training pace and quality of models in the framework. With Torch NN, we want to increase the performance of the detection by using the deep learning to enhance the understanding of intricate patterns in the network data.

4. CONCLUSION

Finally, this paper confirms that anomaly detection systems remain essential in detecting malicious network traffic when considering modern attacks. In doing so, the research shows how the CIC IDS 2017 and CIC IDS 2019 datasets can benefit from a properly tuned advanced machine learning and deep learning algorithms in terms of network security. Among the several classifiers described above the Voting Classifier, which was used with a Boosted Decision Tree and a Bagging Classifier built on the Random Forest algorithm, was the most efficient here. Due to their capability of staking out prediction from two or more models, these systems perform better and are more reliable when identifying various network threats. The selected performance indicators – accuracy, precision, recall, and F1- score, confirm the advantage of the Voting Classifier in the consideration of the modern network traffic challenges. In this way the presented ensemble approach can be applied to improve the network security measures of organizations providing fast and accurate detection of new varieties of malicious actions. This study is therefore a clear pointer to the need for organizations to develop and adopt modern ways of detecting threats to prevent the loss of valuable segments of their networks.

5. FUTURE WORK

Future work includes extending the project to identify several other advanced techniques for detecting bad network traffic. This includes using a combination of different machine learning and deep learning techniques and trying out the unsupervised learning model for surprise detection, apart from introducing new and complex feature engineering techniques to the method. Further, the effectiveness and the capacity of improving the model robustness and the performance will also be examined by the generative adversarial networks (GANs). These developments will apply to the overall improvement of the anomaly detection system for network security.

REFERENCES

[1] Elovici, Y., Shabtai, A., Moskovitch, R., Tahan, G., & Glezer, C. (2007). Applying machine learning techniques for detection of malicious code in network traffic. In KI 2007: Advances in Artificial Intelligence: 30th Annual German Conference on AI, KI 2007, Osnabrück, Germany, September 10-13, 2007. Proceedings 30 (pp. 44-50). Springer Berlin Heidelberg.

[2] Wang, Z., Fok, K. W., & Thing, V. L. (2022).

Machine learning for encrypted malicious traffic detection: Approaches, datasets and comparative study. Computers & Security, 113, 102542.

[3] Alshammari, A., & Aldribi, A. (2021). Apply machine learning techniques to detect malicious network traffic in cloud computing. Journal of Big Data, 8(1), 90.

[4] Rajesh, L., & Satyanarayana, P. (2022). Evaluation of machine learning algorithms for detection of malicious traffic in scada network. Journal of Electrical Engineering & Technology, 17(2), 913-928.

[5] Shafiq, M., Tian, Z., Bashir, A. K., Du, X., & Guizani, M. (2020). CorrAUC: a malicious bot-IoT traffic detection method in IoT network using machine-learning techniques. IEEE Internet of Things Journal, 8(5), 3242-3254.

[6] Mukherjee, B.; Heberlein, L.T.; Levitt, K.N. Network intrusion detection. IEEE Netw. 1994, 8, 26–41.

[7] Garcia-Teodoro, P.; Diaz-Verdejo, J.; Macia- Fernandez, G.; Kim, I. Anomaly-based network intrusion detection: Techniques, systems and challenges. Comput. Secur. 2009, 28, 18–28.

[8] Ashfaq, R.A.R.; Wang, X.Z.; Huang, J.Z.; Abbas, H.; He, Y.L. Fuzziness based semi- supervised learning approach for intrusion detection system. Inf. Sci. 2017, 378, 484–497.

[9] Sabeti, E.; Host-Madsen, A. Data Discovery and Anomaly Detection Using Atypicality for Real-Valued Data. Entropy 2019, 21, 219.

[10] Aloqaily, M.; Otoum, S.; Ridhawi, A.I.; Jararweh, Y. An intrusion detection system for connected vehicles in smart cities. Ad Hoc Netw. 2019, 90.

[11] Lu, H.M.; Li, Y.J.; Mu, S.L.; Wang, D.; Kim,

H.; Serikawa, S. Motor Anomaly Detection for Unmanned Aerial Vehicles Using Reinforcement Learning. IEEE Internet Things J. 2018, 5, 2315–

42322.

[12] Podgorelec, B.; Turkanovic, M.; Karakatic, S. A Machine Learning-Based Method for Automated Blockchain Transaction Signing Including Personalized Anomaly Detection. Sensors 2020, 20, 147.

[13] Wang, J.; Yang, Q.; Ren, D. An intrusion detection algorithm based on decision tree technology. In Proceedings of the 2009 Asia-Pacific Conference on Information Processing, Shenzhen, China, 18–19 July 2009.

[14] Farid, D.M.; Harbi, N.; Rahman, M.Z. Combining Nave Bayes and Decision Tree for Adaptive Intrusion Detection. Available online: https://arxiv.org/abs/1005.4496 (accessed on 15 January 2020).

[15] Hu, W.; Hu, W.; Maybank, S. Adaboost-based algorithm for network intrusion detection. IEEE Trans. Syst. Man Cybern. Part B (Cybern) 2008, 38, 577–583.