13. Dezember 2024;
Zürich, Schweiz

ABSCHNITT 18.
INFORMATIONSTECHNOLOGIEN UND –SYSTEME

DOI 10.36074/logos-13.12.2024.050

ANALYSIS OF MACHINE LEARNING METHODS FOR DETECTING MALICIOUS PROCESSES IN A CORPORATE NETWORK

Vladyslav Samoilenko[1], Sergii Gakhov[2]

1. student at the Faculty of Cybersecurity and Information Protection
State University of Information and Communication Technologies, UKRAINE
ORCID ID: 0009-0001-6367-1755

2. Candidate of Military Sciences, Associate professor, Associate professor of Department of cybersecurity systems and technologies
State University of Information and Communication Technologies, UKRAINE
ORCID ID: 0000-0001-9011-8210

In the modern digital environment, cybersecurity has become critically important for corporate networks. According to NETSCOUT, approximately 7.9 million DDoS attacks were recorded in the first half of 2023, which is 31% more than in the previous year [4]. In Ukraine, the situation has also worsened: according to the State Service of Special Communications and Information Protection, the number of cyberattacks increased by 62.5% in 2023, reaching 1,105 incidents [2]. Although most attacks were aimed at government institutions, corporate networks have also been significantly affected.

Botnets pose a particular threat because they can be used for various malicious activities - from sending spam to large-scale DDoS attacks. Traditional protection methods are not always effective against modern botnets, which continuously evolve and become more sophisticated. Therefore, there is a need to develop new methods for detecting and preventing such attacks, particularly through the application of machine learning.

Traditional intrusion detection systems (IDS) and intrusion prevention systems (IPS) are based on signatures of known attacks or anomaly detection in traffic. Signature-based methods are effective only against known attacks and cannot detect new or modified threats. Anomaly-based methods, while capable of detecting unknown attacks, often generate a large number of false positives, reducing the system's effectiveness.

Machine learning offers new opportunities for automating malicious process detection systems. Models can learn from large volumes of data, identify complex patterns, and adapt to new types of attacks, which reduces the number of false positives and increases the overall system efficiency.

The following algorithms were used in this study:

• **Random Forest**: an ensemble method that uses multiple decision trees to improve accuracy and prevent overfitting.

• **XGBoost**: an advanced gradient boosting of trees, known for its high performance and efficiency.

• **Support Vector Machine (SVM)**: A method that seeks an optimal hyperplane to separate classes in the feature space.

The CICIDS2018 dataset was used to train and evaluate the models, specifically the file *Friday-02-03-2018_TrafficForML_CICFlowMeter.csv*, which contains network traffic labeled as "BENIGN" and botnet attacks [5].

At the initial stage of the study, data loading and preliminary analysis were conducted. Missing values in numerical columns were filled with the mean value to preserve the statistical properties of the data. The "Timestamp" column was removed as it did not carry significant information.

Class distribution analysis showed significant imbalance: the "BENIGN" class prevails over the botnet attack class. To equalize the number of samples, the SMOTE (Synthetic Minority Over-sampling Technique) method was applied [1], which synthetically increases the number of samples of the less represented class. This improved the models' ability to detect botnet attacks.

The first model built was Random Forest. Initial training with default parameters yielded high results. To improve performance, hyperparameter optimization was conducted using the Hyperopt library [3], which employs Bayesian optimization for efficient parameter tuning. After optimization, the model achieved an average F1 Score of 0.99995 with a training time of approximately 1,181 seconds.

The next model is XGBoost, known for its high performance and training speed on large datasets [6]. After hyperparameter optimization, the model showed excellent results: an average F1 Score of 0.99997 with a training time of only 16 seconds. This is significantly less than Random Forest, making XGBoost more suitable for practical use.

The third model is Support Vector Machine (SVM) with a radial basis function (RBF) kernel. This model required significantly more training time due to the large volume of data. After optimizing the regularization parameter C, the model achieved an average F1 Score of 0.708, but the training time was about 2,393 seconds. The low F1 Score and high computational costs make SVM less effective for this task.

After training and optimizing all models, their performance was compared using key metrics: accuracy, precision, recall, F1-score, and training time (Table 1).

Table 1

Table of model training results

Model	accuracy	precision	recall	F1-score	training time (s)
Random Forest (Optimized)	0.9998	0.9996	1.0	0.9998	1181
XGBoost (Optimized)	0.9999	0.9999	0.9999	0.9999	16
SVM (Optimized)	0.8695	0.7706	1.0	0.8702	2393

[author's development]

It is evident from the table that the optimized Random Forest and XGBoost models demonstrate the highest F1-score and accuracy. However, XGBoost has a much shorter training time (about 16 seconds) compared to Random Forest (over 1,180 seconds). Although the SVM model shows acceptable results, it has significantly lower quality metrics and a very long training time, making it less suitable for practical use.

Conclusions. The study confirmed the effectiveness of applying machine learning methods for detecting botnet attacks in corporate networks. XGBoost proved to be the most effective, achieving high accuracy and F1-score metrics with minimal training time. Based on the results of this study, it can be concluded that the XGBoost model is the best choice for use in malicious activity detection systems.

REFERENCES:

[1] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

[2] Державна служба спеціального зв'язку та захисту інформації України. (2024, січень 13). Кількість кібератак в Україні зросла на 62% у 2023 році. Мінфін. https://minfin.com.ua/ua/2024/01/13/119569727/

[3] Hyperopt. (n.d.). *Hyperopt Documentation*. http://hyperopt.github.io/hyperopt/

[4] NETSCOUT. (2023). NETSCOUT виявили майже 7,9 млн DDoS-атак у першій половині 2023 року. https://netscout.bakotech.com/ua/ddos-attacks-report-2023

[5] University of New Brunswick. (2018). *CSE-CIC-IDS2018 on AWS*. https://www.unb.ca/cic/datasets/ids-2018.html

[6] XGBoost Developers. (n.d.). *XGBoost Documentation*. https://xgboost.readthedocs.io/