Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 8, Nomor 1, Februari 2023, pp 204-212
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Injection Attack Detection on Internet of Things Device with Machine Learning Method

***Mara Muda Pohan*[1], *Benfano Soewito*[2]**
[1,2] *Master of Information Technology, Bina Nusantara University, Jakarta, Indonesia*
Email : mara.muda@binus.ac.id, bsoewito@binus.edu

Abstract

The Internet of Things (IoT) Industry is growing rapidly, security surrounding this Industry has to be upgraded. This study analyzes which machine learning performs the best in detecting Injection Attacks in IoT devices. The proposed machine learning methods includes Catboost, Decision Tree, Support Vector Machine (SVM), and Multilayer Perceptron (MLP). This study uses Edge-IIoTset dataset. The traffic data obtained in this dataset comes from 13 different types of IoT devices which contains 10 files with normal traffic and 14 files of attack traffics. This study takes normal traffic and injection attacks traffic from Edge-IIoTset. Results shows that Catboost machine learning model performs the best in terms of performance score with 0.95599 score in Accuracy, Precision, F1-Score, and recall score where as Decision Tree model performs the fastest with 0.09 seconds of runtime and achieving 0.93 score in the performance.

***Keywords**: Machine Learning, Internet of Things (IoT), Catboost, Decision Tree, Support Vector Machine (SVM), Multilayer Perceptron (MLP), Injection Attack*

1. INTRIDUCTION

At this time, technological developments have entered a period where devices need a sensor to be able to operate. A device that has a sensor, can provide data or information that is useful to humans, when a device has a sensor or many sensors is connected to the internet, which allows a device to communicate or exchange data with other devices that can produce comprehensive data or information for humans, the device is called the internet of things. IoT devices themselves unconsciously are all around us and are used in everyday life [1]. Based on data [2], it is predicted that in 2030 there will be 29.4 billion devices. From the projection data shown in Figure 1. It can be stated that the growth in the use of IoT devices has almost tripled in a decade or within 10 years. The growth in the use of IoT itself is caused by the increasing demand for smart devices or wearables in our lives. Smart home-based devices, smart cars, smart watches, and others are triggers for the mass use of IoT.

Figure 1. IoT Growth Projection. Reference: (Statista,2022a)

However, in line with the increasing number of IoT devices, this technology certainly has its own challenges. The challenges faced by this technology can be categorized into 4 main challenges, namely security, privacy, data volume and complexity [3]. Security and privacy constraints are the biggest concerns in the IoT world. This is because the ability of IoT devices can directly attack a person's privacy, which can retrieve surrounding data and personal data from devices used by individuals. While security constraints themselves, can directly affect or cause privacy constraints to occur or make IoT devices not work or provide wrong information. With security constraints, both IoT device and service providers or individual users can suffer significant losses [4].

Attacks that are usually carried out by hackers on IoT devices themselves can be categorized into 5 categories. Attacks can be categorized as DoS/DDoS, Injection, Information Gathering, Man in the Middle of attack and Malware [6]. With this attack category, it is also necessary to take preventive and countermeasures when an attack occurs. This leads to a strong need for cyber security, where by increasing cyber security, the losses that can be experienced by companies or individuals can be minimized. Globally, the cybersecurity market alone by 2022 is estimated to reach $159.8 billion in revenue [2]. Meanwhile, based on data from Businesswire [13] as seen below in figure 2, provides an estimate that 41.6 billion devices will be connected to IoT in 2025. In addition, IDC also reports an estimate that there will be an exchange of 79.4 zettabytes of data in 2025 from IoT devices where this figure is up by 3 times more than in 2019 of 18.3 zettabytes.

Figure 2. A Comperhensive Estimation view of IoT Security

DoS / DDoS itself is an attack that most often occurs on IoT devices, where this attack allows loss of functionality in the system [6] Based on [7], in 2020 alone there was a 151% increase in attacks that occurred in the form of DoS / DDoS on the internet network as a whole. Meanwhile, based on [8] in a period of 15 months from 2020 to 2021, DoS/DDoS attacks reached 500 Gbps. However, on the other hand, with many attacks in the form of DoS/DDoS, research and preventive measures for this type of attack are much advanced than other types of attacks.

Therefore, this study will focus on the type of attack that is the second most common, namely the injection attack.

This study will focus on detecting injection attacks with 4 methods of machine learning which is Catboost, Decision Tree, Multilayer Perceptron (MLP), and Support Vector Machine (SVM). Where author will find the best Hyper Parameter for each model and compare each model to find which Machine Learning model has the fastest and most accurate results.

2. RESEARCH METHODOLOGY

There are many websites where users can upload documents (for example, financial documents, resumes, profile pictures, etc.). This can be exploited by attackers by entering files that are not in accordance with system functions. Once the attacker manages to upload the malware program files to the web server, he can gain administrative privileges. This can trigger a very significant risk, where the file when the file is uploaded in the execution, the attack will occur. The consequences of this attack can vary, including a complete system takeover, an overloaded file system or database, forwarding the attack to a back-end system, a client-side attack, or a simple crash, an attacker can also upload and run a web shell, filtrating potentially confidential data, uploading permanent XSS as well as phishing pages [6].

SQL Injection is a technique that sends malicious messages to the DB trying to find an unauthorized channel to gain access. This is an application security flaw that allows different attackers to control database driven applications by letting them access or delete sensitive data.

In this section, we briefly describe the various machine learning algorithms that are frequently used in this domain. Support Vector Machine (SVM) is part of supervised machine learning that is often used, especially in malware attack detection research [14]. Support Vector Machine (SVM) aims to separate data into two classes, namely classes +1 and -1. This can be achieved by finding the maximum distance between the two classes.

Decision tree is one of the most popular classification machine learning algorithms. This is because the concept of a decision tree is easy for humans to understand. Decision tree is a classification method that is shaped like a tree structure, each node in the decision tree represents an attribute and the branch of the node represents the value of the attribute, and the leaf represents the class. Root is the top node of the decision tree.

Multi-layer perceptron is one of supervised machine learning with classification method. Multi-layer perceptron is part of deep learning (DL), the learning process in deep learning is similar to how humans learn, because the structure of deep learning or artificial neural network is taken from the structure of the human brain. The perceptron is the smallest part of the neural network, which consists of a neuron, similar to a human nerve cell.

Categorical Boosting (Catboost) is one of the boosting algorithms which is a tree algorithm (such as a decision tree) that is more sophisticated. CatBoost creates a combination of categorical and numeric attributes. All splits in the tree are used in combination in the same way as categorical splits and are considered as categorical with two values. CatBoost is an algorithm that combines Gradient Boosting Decision Tree (GDBT) with categorical features, this algorithm is a more advanced version of GDBT. CatBoost's main goal is to be able to process categorical features efficiently and rationally

Figure 3. Research Framework

3. RESULTS AND DISCUSSION

a. Support Vector Machine

The results of the accuracy level of the classification performance of the Support Vector Machine model using the values of degree 1, 2, and 3 get the same value, which is 91.83% with the fastest learning time of 82.04 seconds at degree 2 where the performance be seen in in table 6 and the confusion matrix can be seen in figure 5. In Support Vector Machine, Precision, Recall, and F1 Score has the same value for all degree values, which is 0.918332.

Table 1. SVM with degree value of 2 Performance Score

Accuracy	F1- Score	Recall	Precision	Time Complexity
0.918332	0.918332	0.918332	0.918332	82.0401 Seconds

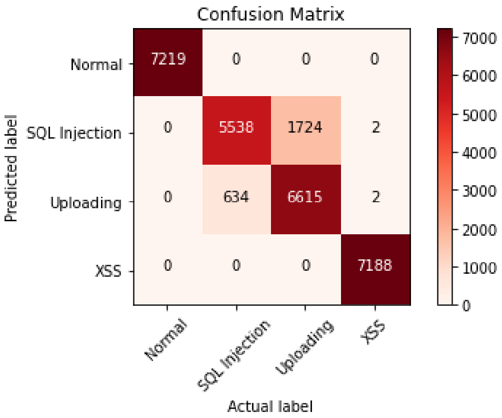


Figure 4. SVM with degree value of 2 Confusion Matrix

b. Decision Tree

The results of the accuracy level of the Decision Tree model classification performance with max_depth values of 3, 5 and 7 get a relatively good score. The maximum accuracy reaches 93.47% at max_depth 7 with a learning time of 0.12 seconds. The fastest learning algorithm is at max_depth 3, which is 0.09 seconds. For Decision Tree, Precision, Recall, and F1, the model gets the highest score with a max_depth 7 value of 0.934756. The performance be seen in in table 7 and the confusion matrix can be seen in figure 6

Table 2. Decision tree with max_depth 7 Performance Score

Accuracy	F1- Score	Recall	Precision	Time Complexity
0.934756	0.934756	0.934756	0.934756	0.123218 Seconds

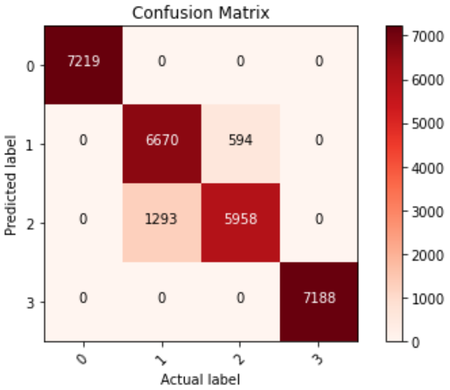


Figure 5. Decision Tree with max_depth 7 Confusion Matrix

c. Multi-Layer Perceptron

The results of the accuracy level of the classification performance of the Multi-Layer Perceptron model using hidden_layer_sizes values of 3,5, and 7 have relatively small differences. The highest accuracy is found in the model which has a hidden_layer_sizes value of 5, which is 93.46%. with a learning time of 24 seconds. The fastest learning time is found in the Multi-Layer Perceptron with the hidden_layer_sizes hyperparameter value of 7, which is 18 seconds. In Multi-Layer Perceptron, Precision, Recall, and F1 Score get the highest value at hidden_layer_sizes 7 where the performance be seen in in table 8 and the confusion matrix can be seen in figure 7, which is 0.932508.

Table 3. MLP with hidden_layer 7 Performance Score

Accuracy	F1- Score	Recall	Precision	Time Complexity
0.932508	0.932508	0.932508	0.932508	18.3246 Seconds

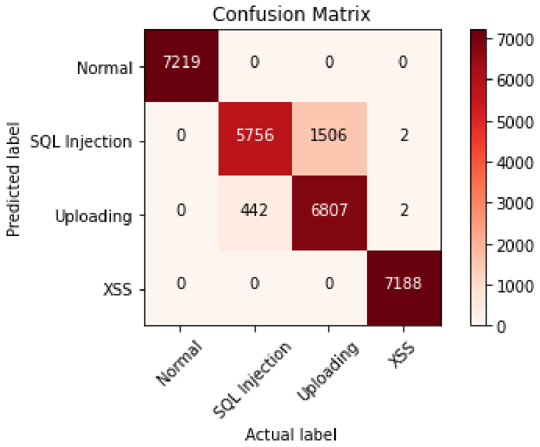


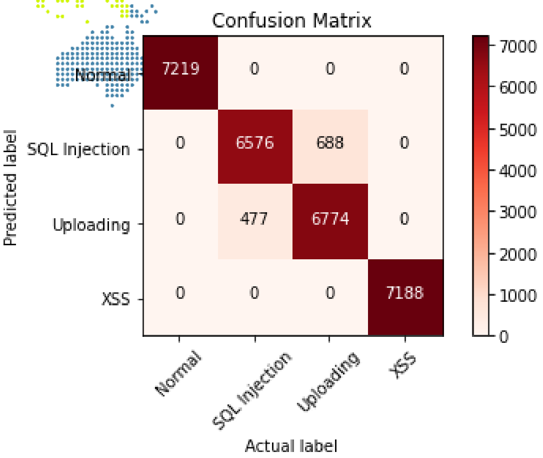
Figure 6. MLP with hidden_layer 7 Confusion Matrix

d. Catboost

The results of the accuracy level of the classification performance of the CatBoost model using the learning_rate value of 0.1; 0.01; 0.001 shows the same accuracy value, which is 95.97% with the fastest learning time, which is 39.29 seconds on the model with a learning rate of 0.1. The performance be seen in in table 9 and the confusion matrix can be seen in figure 8. In CatBoost, Precision, Recall, and F1 Scores get the highest score at learning_rate 0.1, which is 0.959719.

Table 4. Catboost with Learning_rate 0.1 Performance Score

Accuracy	F1- Score	Recall	Precision	Time Complexity
0.959719	0.959719	0.959719	0.959719	39.2995 Seconds



	Normal	SQL Injection	Uploading	XSS
Normal	7219	0	0	0
SQL Injection	0	6576	688	0
Uploading	0	477	6774	0
XSS	0	0	0	7188

Figure 7. Catboost with learning_rate 0.001 Confusion Matrix

e. Model Performance Comparison

Of the four models used in this study and as shown in table 10, the highest accuracy value can be achieved in the CatBoost model with a learning rate of 0.1 achieving an accuracy of 95.97% with a learning time of 39.29 seconds. However, the Decision Tree model managed to achieve the fastest learning time, which was 0.09 seconds. The full comparison can be seen in table 10 below.

Figure 7. Table Performance score from all machine learning models

4. CONCLUSION

Although hyperparameters affect machine learning model learning, the values of different hyperparameters do not necessarily provide different performance, as seen in the Support Vector Machine and Catboost models which are given three different values for the selected hyperparameter, but the performance on both machine models the learning remains the same.

The results of the accuracy level for the classification performance of the Support Vector Machine model using degree values of 1, 2, and 3 give the same value of 91.83%, and the fastest training time is 82.04 seconds with degree 2. For Support Vector Machine, precision, recall, and F1 scores have the same value for all degree values, which is 0.918332.

The results of the accuracy level of the Decision Tree model classification performance for max_depth values 3, 5 and 7 get relatively good scores. Maximum accuracy reaches 93.47% at max_depth 7 and training time is 0.12 seconds. The fastest learning algorithm is max_depth 3 (0.09 seconds). For Decision Tree, Precision, Recall, and F1, the model gets the highest score with a max_depth value of 7, which is 0.934756.

The results of the accuracy level of classification performance for the Multi-Layer Perceptron model using hidden_layer_sizes 3, 5 and 7 values show a relatively small difference. The highest accuracy was found for the model with a Hidden_layer_sizes value of 5 (93.46%). The fastest learning time is 18 seconds on hidden_layer_sizes 7. For Multi-Layer Perceptron, precision, recall, and F1 score get the highest score (0.934617) on hidden_layer_sizes 5.

In this study, Catboost, which is one of the more advanced machine learning models, is able to learn and achieve the best performance from the other three algorithms tested, where this algorithm achieves 95.59% accuracy with F1-score, Recall, and precision. 0.959719. This algorithm is classified as an algorithm that is more advanced than Support Vector Machine, Decision Tree, and Multi-Layer Perceptron and also proves research [8], where Catboost managed to outperform two machine learning algorithms, namely XGBoost and LightGBM.

Decision Tree managed to get the fastest learning time, which was 0.09 seconds. This is because the Decision Tree algorithm does not calculate all the possibilities in the learning process, this makes Decision Tree able to make decisions and learn faster than other algorithms. Although the Decision Tree algorithm has the fastest learning time, the accuracy obtained from learning the Decision Tree model is not the highest.

REFERENCES

[1] Acharya, S. (2022, March 17). Why IoT Security is Important for Today's Networks? SECTRIO. https://sectrio.com/why-iot-security-is-important-for-todays-networks/.

[2] Statista. (2022a, May). Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2030. Statista. https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/.

[3] Weinberg, A. (2021, October 24). Top 4 Challenges in IoT Data Collection and Management. FirstPoint. https://www.firstpoint-mg.com/blog/top-4-challenges-in-iotdata-collection-and-management/.

[4] Tawalbeh, L., Muheidat, F., Tawalbeh, M., & Quwaider, M. (2020). IoT privacy and security: Challenges and solutions. Applied Sciences (Switzerland), 10(12). https://doi.org/10.3390/APP10124102.

[5] Conti, M., Dehghantanha, A., Franke, K., & Watson, S. (2018). Internet of Things security and forensics: Challenges and opportunities. In Future Generation Computer Systems (Vol. 78, pp. 544–546). Elsevier B.V. https://doi.org/10.1016/j.future.2017.07.060.

[6] Ferrag, M. A., Friha, O., Hamouda, D., Maglaras, L., & Janicke, H. (2022). Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning. IEEE Access, 10, 40281–40306. https://doi.org/10.1109/ACCESS.2022.3165809.

[7] Kidd, I. (2021, September 30). The Shocking DDoS Attack Statistics That Prove You Need Protection. Info Security Magazine. https://www.infosecuritymagazine.com/blogs/ddos-attacks-stats-protection/.

[8] Warburton, D. (2021, May 7). DDoS Attack Trends for 2020. F5 Application Threat Intelligence. https://www.f5.com/labs/articles/threat-intelligence/ddos-attack-trendsfor-2020.

[9] Gaber, T., El-Ghamry, A., & Hassanien, A. E. (2022). Injection attack detection using machine learning for smart IoT applications. Physical Communication, 52. https://doi.org/10.1016/j.phycom.2022.101685.

[10] Danardono, E. (2021). Detection of Distributed Denial of Service Attacks from Internet of Things Devices using SVM and NN-MLP methods [Thesis, Bina Nusantara]. http://library.binus.ac.id/Collections/ethesis_detail/RS2-KG-MTI-2021-0011.

[11] Zhou, F., Pan, H., Gao, Z., Huang, X., Qian, G., Zhu, Y., & Xiao, F. (2021). Fire Prediction Based on CatBoost Algorithm. Mathematical Problems in Engineering, 2021. https://doi.org/10.1155/2021/1929137.

[12] Nugroho, K. S. (2019, October 13). Confusion Matrix for Model Evaluation on Supervised Learning. Medium. https://ksnugroho.medium.com/confusion-matrixfor-evaluation-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f.

[13] Businesswire. (2019). The Growth in Connected IoT Devices is Expected to Generate 79.4ZB of Data in 2025, According to a New IDC Forecast. Businesswire. https://www.businesswire.com/news/home/20190618005012/en/The-Growth-inConnected-IoT-Devices-is-Expected-to-Generate-79.4ZB-of-Data-in-2025-Accordingto-a-New-IDC-Forecast.

[14] Hussain, F., Hussain, R., Hassan, S. A., & Hossain, E. (2019). Machine Learning in IoT Security: Current Solutions and Future Challenges. http://arxiv.org/abs/1904.05735.

[15] Garcia-Alfaro, J., & Navarro-Arribas, G. (2009). A Survey on Cross-Site Scripting Attacks. http://arxiv.org/abs/0905.4850.