Implementasi Metode Seleksi Fitur Mutual Information pada Algoritma Random Forest untuk Deteksi Serangan Siber pada Jaringan Wi-Fi = Implementation of Mutual Information Feature Selection Method on Random Forest Algorithm for Cyber Attack Detection on Wi-Fi Networks

Dewita Oktavia Nuur Marwan, author

Abstrak

Internet of Things (IoT) merupakan sebuah konsep di mana berbagai perangkat komputasi saling terhubung melalui internet dan memiliki kemampuan untuk mengumpulkan atau mengirimkan data. Perluasan dan kecepatan perangkat komputasi menggunakan jaringan Wi-Fi dapat menghasilkan data yang kompleks dan berdimensi tinggi pada sistem IoT. Data yang berdimensi tinggi dapat menimbulkan beberapa kendala dan perangkat IoT akan menghindari untuk melakukan tugas yang komputasinya berat. Semakin kompleksnya sistem IoT, semakin sulit bagi sistem untuk mengidentifikasi dan menemukan serangan siber. Salah satu upaya yang paling umum digunakan untuk melindungi sistem IoT adalah Intrusion detection system (IDS). Pada penelitian ini dilakukan model berbasis machine learning untuk mengembangkan IDS menggunakan dataset AWID2 dengan tipe “CLS” yang berisikan 2 juta lalu lintas trafik pada jaringan WI-Fi yang dikelompokkan ke dalam empat kelas yaitu, normal, impersonation, injection, dan flooding. Random forest merupakan salah satu teknik ensemble atau gabungan dari sejumlah model decision tree yang memiliki keunggulan-keunggulan dibandingkan dengan metode machine learning lainnya, yaitu dapat mencegah terjadinya overfitting, memiliki waktu komputasi yang rendah, dan memiliki kemampuan lebih baik dalam mengelola dataset yang tidak seimbang. Untuk mengatasi data berdimensi tinggi, dilakukan seleksi fitur mutual information pada algoritma random forest untuk mendapatkan hasil model klasifikasi yang optimal. Hasil dari penelitian menunjukkan bahwa metode seleksi fitur mutual information dengan menggunakan 30 fitur terbaik pada algoritma random forest dengan hyperparameter-tuning random search terbukti dapat meningkatkan performa model klasifikasi dan efisiensi waktu jika dibandingkan menggunakan algoritma random forest tanpa seleksi fitur. Nilai metrik yang diperoleh oleh kombinasi tersebut adalah dengan nilai accuracy = 99,95276%, macro average F1-score = 99,76335%, macro average recall = 99,97962%, dan macro average presicion = 99,54935% dengan waktu prediksi 6,112 detik.

.....The Internet of Things (IoT) is a concept where various computing devices are interconnected via the internet and have the capability to collect or transmit data. The expansion and speed of computing devices using Wi-Fi networks generate complex and high-dimensional data in IoT systems. High-dimensional data in datasets pose several challenges, as IoT devices tend to avoid tasks that are computationally intensive. As IoT systems become more complex, it becomes increasingly difficult for the system to identify and detect cyber attacks. One of the most common efforts to protect IoT systems is the Intrusion Detection System (IDS). In this study, a machine learning-based model is developed to create an IDS using the AWID dataset with the “CLS” type, which contains 2 million network traffic records on Wi-Fi networks categorized into four classes: normal, impersonation, injection, and flooding. Random forest is an ensemble technique or a combination of multiple decision tree models that has advantages over other machine learning methods, such as preventing overfitting, having low computational time, and having better capabilities in handling

imbalanced datasets. To address high-dimensional data, mutual information feature selection is applied to the random forest algorithm to achieve optimal classification model results. The results of the study indicate that the mutual information feature selection method using the top 30 features in the random forest algorithm with random search hyperparameter tuning can improve the performance of the classification model and time efficiency compared to using the random forest algorithm without feature selection. The metrics obtained by this combination are accuracy = 99.95276%, macro average F1-score = 99.76335%, macro average recall = 99.97962%, and macro average precision = 99.54935% with a prediction time of 6.112 seconds.