# ABSTRAK

*Intrusion Detection System* (IDS) merupakan alat penting dalam keamanan jaringan yang berfungsi untuk memantau dan mendeteksi serangan. IDS menjadi krusial dalam mengidentifikasi serangan yang dapat melewati langkah-langkah keamanan tradisional seperti *firewall* dan *antivirus*. Saat ini, penerapan *machine learning* telah banyak dilakukan untuk meningkatkan kinerja deteksi intrusi pada data log sistem jaringan. Namun, beberapa metode masih menghadapi keterbatasan seperti tingginya angka positif palsu, rendahnya akurasi klasifikasi, dan rendahnya angka positif sebenarnya. Selain itu, fitur-fitur asing serta besarnya jumlah data yang perlu diperiksa oleh IDS, membuat proses analisis menjadi sulit bahkan dengan bantuan komputer. Oleh karena itu, pada penelitian ini akan menerapkan seleksi fitur *Information Gain* dengan 3 nilai *threshold* berbeda yang akan dievaluasi untuk menemukan nilai *threshold* terbaik pada seleksi fitur untuk *Random Forest* pada case *Intrusion Detection System* (IDS).

Metodologi dari penelitian ini akan melibatkan beberapa tahap. Pertama, dilakukan pengumpulan data, penelitian ini menggunakan *dataset NSL-KDD,* dimana data berupa log jaringan yang sudah di proses. Lalu akan dilakukan *preprocessing* berupa *class label mapping, label encoding, splitting data, information gain,* dan *standardization*. Lalu pelatihan dan evaluasi model akan dilakukan. Data yang sudah melalui *preprocessing* akan digunakan untuk melatih model, dan hasilnya akan dievaluasi dengan metrik akurasi, presisi, *recall* dan f1-*score.* Dalam proses ini, akan dilakukan pemodelan untuk model *multiclass* dan model *binary class,* dimana masing – masing model akan dibangun dengan 3 *threshold* yang berbeda pada seleksi fitur *information gain*-nya.

Hasil dari pengujian pada model yang dibangun menunjukan hasil positif. Dengan menguji 3 *threshold* berbeda (standar deviasi, 0,05, dan 0,07), model *multiclass* dengan *threshold* 0,05 mencapai akurasi tertinggi sebesar 99,95%, dengan *precision, recall,* dan *F1-score* yang berkisar antara 99,70% hingga 99,99% di seluruh kelas. Pada model *binary class*, penggunaan *threshold* standar deviasi menghasilkan akurasi sebesar 99,88%, dengan *precision* sebesar 99,85%, *recall* 99,93%, dan *F1-score* 99,89% untuk kelas "*Attacking*". Namun, masih terdapat misklasifikasi pada kelas *Probe,* yang disebabkan oleh kemiripan data kelas *Probe* dengan kelas *Normal*. Hasil pengujian ini menunjukkan bahwa penerapan seleksi fitur *Information Gain* dengan *threshold* yang sesuai dapat meningkatkan efektivitas deteksi serangan pada sistem IDS dengan algoritma *Random Forest*.

**Kata Kunci**: *Intrusion Detection System, Information Gain, Random Forest,* seleksi fitur, *threshold*.

# ABSTRACT

*An Intrusion Detection System (IDS) is an important tool in network security that functions to monitor and detect attacks. IDS can be crucial in identifying attacks that can bypass traditional security measures such as firewalls or an antivirus software. Currently, the application of machine learning has been implemented to further enhance the performance of intrusion detection in network system log data. However, most of the methods used still face limitations such as a high rate of false positives, low classification accuracy, and a low rate of true positives. In addition, the foreign features and the large amount of data that need to be examined by the IDS make the analysis process difficult, even with the help of computers. Therefore, this study will apply Information Gain feature selection with 3 different threshold values that will be evaluated to find the best threshold value for feature selection in Random Forest for the Intrusion Detection System case. (IDS).*

*The methodology of this research will involve several stages. First, data collection was carried out; this research uses the NSL-KDD dataset, which consists of processed network log data. Then, the data will goes to the preprocessing process, which consist of class label mapping, label encoding, data splitting, information gain, and standardization. After that, model training and evaluation will be conducted. The data that has been preprocessed will be used to train the model, and the results will be evaluated using accuracy, precision, recall, and F1-score metrics. In this process, modeling will be conducted for both multiclass and binary class models, where each model will be built with three different thresholds based on information gain for feature selection.*

*The test results from the constructed models shows positive outcomes. By testing three different thresholds (standard deviation, 0.05, and 0.07), the multiclass model with a threshold of 0.05 achieved the highest accuracy of 99.95%, with precision, recall, and F1-score ranging from 99.70% to 99.99% across all classes. In the binary class model, using the standard deviation threshold resulted in an accuracy of 99.88%, with precision at 99.85%, recall at 99.93%, and an F1-score of 99.89% for the "Attacking" class. However, there is still misclassification in the Probe class, which is caused by the similarity of the Probe class data to the Normal class. The results of this test indicate that the application of Information Gain feature selection with an appropriate threshold can enhance the effectiveness of attack detection in IDS systems using the Random Forest algorithm.*

***Keywords***: *Intrusion Detection System, Information Gain, Random Forest, feature selection, threshold.*