Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202183154
Vol. 8, No. 3, Juni 2021, hlm. 457-464 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

INTRUSION DETECTION SYSTEM BERBASIS SELEKSI FITUR DENGAN KOMBINASI FILTER *INFORMATION GAIN RATIO* DAN *CORRELATION*

Nitami Lestari Putri*[1], Radityo Adi Nugroho[2], Rudy Herteno[3]

[1,2,3] Program Studi Ilmu Komputer, Fakultas MIPA, Universitas Lambung Mangkurat
Email: [1]nitamiputri38@gmail.com, [2]radityo.adi@ulm.ac.id, [3]rudy.herteno@ulm.ac.id
*Penulis Korespondensi

(Naskah masuk: 23 Januari 2020, diterima untuk diterbitkan: 08 Juni 2021)

Abstrak

Intrusion Detection System merupakan suatu sistem yang dikembangkan untuk memantau dan memfilter aktivitas jaringan dengan mengidentifikasi serangan. Karena jumlah data yang perlu diperiksa oleh IDS sangat besar dan banyaknya fitur-fitur asing yang dapat membuat proses analisis menjadi sulit untuk mendeteksi pola perilaku yang mencurigakan, maka IDS perlu mengurangi jumlah data yang akan diproses dengan cara mengurangi fitur yang dapat dilakukan dengan seleksi fitur. Pada penelitian ini mengkombinasikan dua metode perangkingan fitur yaitu *Information Gain Ratio* dan *Correlation* dan mengklasifikasikannya menggunakan algoritma *K-Nearest Neighbor*. Dataset yang digunakan pada penelitian ini yaitu dataset 10% KDD CUP 99. Hasil perangkingan dari kedua metode dibagi menjadi dua kelompok. Pada kelompok pertama dicari nilai mediannya dan untuk kelompok kedua dihapus. Lalu dilakukan klasifikasi *K-Nearest Neighbor* dengan menggunakan 10 kali validasi silang dan dilakukan pengujian dengan nilai k=5. Penerapan pemodelan yang diusulkan menghasilkan akurasi tertinggi sebesar 99.61%. Sedangkan untuk akurasi tanpa seleksi fitur menghasilkan akurasi tertinggi sebesar 99.59%.

Kata kunci: *Intrusion Detection System, Seleksi Fitur, Kombinasi Filter*, *Information Gain Ratio*, *Correlation*, *K-Nearest Neighbor*

INTRUSION DETECTION SYSTEM BASED ON FEATURE SELECTION WITH FILTER COMBINATION OF INFORMATION GAIN RATIO AND CORRELATION

Abstract

Intrusion Detection System is a system that was developed for monitoring and filtering activity in network with identified of attack. Because of the amount of the data that need to be checked by IDS is very large and many foreign feature that can make the analysis process difficult for detection suspicious pattern of behavior, so that IDS need for reduce amount of the data to be processed by reducing features that can be done by feature selection. In this study, combines two methods of feature ranking is Information Gain Ratio and Correlation and classify it using K-Nearest Neighbor algorithm. The dataset used in this study is the 10% KDD CUP 99 dataset. The result of feature ranking from the both methods divided into two groups. in the first group searched for the median value and in the second group is removed. Then do the classification of K-Nearest Neighbor using 10 fold cross validation and do the tests with values k=5. The result of the proposed modelling produce the highest accuracy of 99.61%. While the highest accuracy value of the not using the feature selection is 99.59%.

Keywords: *Intrusion Detection System, Feature Selection, Filter Combination, Information Gain Ratio, Correlation, K-Nearest Neighbor,*

1. PENDAHULUAN

Intrusion Detection System telah dikembangkan untuk memantau dan memfilter aktivitas jaringan dengan mengidentifikasi serangan. Ini merupakan suatu solusi yang dapat digunakan pada masalah keamanan jaringan. Berbagai pendekatan IDS telah banyak diperkenalkan dengan menggunakan *data mining*, *machine learning*, analisis statistik dan Teknik kecerdasan buatan seperti algoritma genetika, jaringan saraf tiruan, logika fuzzy, *swarm intelligence* dan lain sebagainya (Khammassi & Krichen, 2017).

Karena jumlah data yang perlu diperiksa oleh IDS sangat besar, sehingga membuat proses analisis menjadi sulit bahkan dengan bantuan komputer. Fitur-fitur asing juga dapat membuat proses analisis menjadi lebih sulit untuk mendeteksi pola perilaku yang mencurigakan. Akibatnya, IDS harus mengurangi jumlah data yang akan diproses. Pengurangan fitur dapat dilakukan

dengan penyaringan data, pengelompokkan data atau seleksi fitur (Hasan, Nasser, Ahmad & Molla, 2016).

Seleksi fitur merupakan hal penting dalam IDS untuk mendapatkan kinerja yang lebih baik. Metode peringkat dan seleksi fitur berguna untuk kepentingan fitur yang ada dalam dataset dan dapat mengkategorikannya menjadi fitur dengan signifikan tinggi atau rendah. Fitur-fitur yang dipilih dapat membantu untuk megklasifikasikan lalu lintas data di jaringan ke kelas normal atau serangan. Fitur yang tidak berkontribusi dalam mendeteksi berbagai jenis serangan harus dihapus untuk mendapatkan akurasi dan kecepatan yang lebih baik dalam IDS. Penghapusan fitur ini akan membuat kinerja IDS menjadi lebih baik dalam hal perhitungan, pengurangan dimensi dan kompleksitas waktu (Akashdeep, Manzoor & Kumar, 2017).

Pada penelitian seleksi fitur, dilakukan juga tahap klasifikasi untuk mendapatkan akurasi dari subset fitur yang dipilih dari metode seleksi fitur. Beberapa algoritma klasifikasi sering digunakan dalam IDS yang salah satunya adalah algoritma *K-Nearest Neighbour* (KNN). algoritma KNN merupakan algoritma yang berbasis *instance.* Pada algoritma ini, model klasifikasi dibangun berdasarkan kesamaan antara *instance* pada data pelatihan dengan nilai k terdekat dari *instance* tertentu. Klasifikasi pada IDS dapat menghasilkan akurasi dan juga dapat mendeteksi serangan ke sistem yang lebih tinggi (Onan & Korukoglu, 2015).

Pada penelitian yang dilakukan oleh Nababan, Sitompul & Tulus dengan judul "*Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio*" membuktikan bahwa dengan menggunakan metode *Information Gain Ratio* sebagai metode seleksi fitur dan algoritma *K-Nearest Neighbor* sebagai algoritma pengklasifikasian membuat nilai akurasi menjadi lebih tinggi daripada hanya menggunakan algoritma *K-Nearest Neighbor* saja. Pengklasifikasian dilakukan dengan menguji k dari algoritma *K-Nearest Neighbor*. K yang diuji pada penelitian adalah k=1 sampai dengan k=10. Hasil akurasi dari masing-masing k yang diuji dengan pendekatan yang diusulkan menghasilkan peningkatan akurasi nilai KNN dengan nilai tertinggi pada nilai k=3 yaitu 9.2%. Dan juga perbandingan akurasi antara algoritma KNN dan kombinasi antara KNN dan *Information Gain Ratio* menunjukkan hasil rata-rata akurasi semua nilai k meningkatkan akurasi sebesar 5.35% terhadap nilai akurasi yang hanya menggunakan algoritma KNN saja (Nababan, Sitompul & Tulus, 2018).

Penelitian yang dilakukan oleh Selvakumar & Muneeswaran dengan judul "*Firefly Algorithm Based Feature Selection For Network Intrusion Detection"* menggunakan algoritma Mutual Information Firefly (MIFA) yang digunakan sebagai strategi pemilihan fitur dalam seleksi fitur berbasis wrapper dengan C4.5 dan *Bayesian Network* sebagai metode klasifikasi. Setelah melakukan seleksi fitur dan didapatkan ada 10 fitur yang dipilih kemudian akurasi tertinggi yang didapatkan menggunakan algoritma C4.5 dengan akurasi 99.98% pada kelas DOS (Selvakumar & Muneeswaran, 2018).

Pada penelitian yang dilakukan oleh Cilia, Stefano, Fontanella, Raimondo & Freca dengan judul "*An Experimental Comparison Of Feature Selection And Classification Methods For Microarray Datasets"* menggunakan lima metode seleksi fitur untuk dapat merangkingkan set fitur pada dataset *microarray*. Lima metode seleksi fitur tersebut antara lain yaitu *Chi-square*, *Relief*, *Gain Ratio*, *Information Gain* dan *Symmetrical Uncertainty*. Untuk tahap pengklasifikasian menggunakan empat metode klasifikasi yang juga dilakukan 10 kali validasi silang pada setiap metode klasifikasi. Empat metode klasifikasi yang digunakan antara lain *Decision Tree*, *Random Forest*, *K-Nearest Neighbour* dan *Artificial Neural Network*. Hasil dari penelitian ini, metode seleksi fitur yang banyak unggul di beberapa dataset *Microarray* adalah metode *Gain Ratio* yang unggul pada dataset *breast, Colon* dan *Ovarian* yang masing masing nilai dari *recognition rate* adalah 84.69, 82.25 dan 87.91 (Cilia, Stefano, Fontanella, Raimondo & Freca, 2019).

Pada penelitian yang dilakukan oleh Akande, Owolabi & Olatunji dengan judul "*Investigating The Effect Of Correlation Based Feature Selection On The Performance Of Support Vector Machines In Reservoir Characterization*" dimana penelitian ini menyelidiki pengaruh pemilihan fitur pada kinerja generalisasi dan kemampuan prediksi *Support Vector Machine* dalam memprediksi permeabilitas reservoir karbonat. Pada tahap seleksi fitur menggunakan algoritma *Correlation Based Feature Selection*. Dengan menggunakan pendekatan yang diusulkan dan hasil yang didapatkan mampu meningkatkan efisiensi algoritma, peningkatan kinerja, waktu yang lebih singkat dan tidak terjadi *overhead* pada komputasi (Akande, Owolabi & Olatunji, 2015).

Pada penelitian ini mengkombinasikan dua algoritma seleksi fitur yaitu *Information Gain Ratio* dan *Correlation* dan juga mengklasifikasikannya menggunakan algoritma *K-Nearest Neighbor*. Penelitian ini menggunakan dataset 10% KDD CUP 99 dan hasil dari penelitian ini akan dibandingkan dengan dataset yang sama tanpa menggunakan seleksi fitur.

2. METODE PENELITIAN

Pada Gambar 1 dapat dilihat alur penelitian yang digunakan sebagai berikut:

a. Pada tahap awal penelitian dataset yang digunakan adalah dataset 10% KDD CUP 99 secara keseluruhan. Dataset ini sudah banyak digunakan untuk penelitian pada *Intrusion Detection System.* Dataset 10% KDD CUP 99 juga menjadi tolak ukur untuk *Intrusion Detection System*. Salah satu penelitian yang menggunakan dataset 10% KDD CUP 99 yaitu pada penelitian Luo & Xia dengan judul "*A Novel Intrusion Detection System Based On Feature Generation With Visualization Strategy"* dimana pada penelitian tersebut melakukan generalisasi fitur di IDS dan tujuannya adalah untuk meningkatkan kinerja deteksi pada IDS (Luo & Xia, 2014).

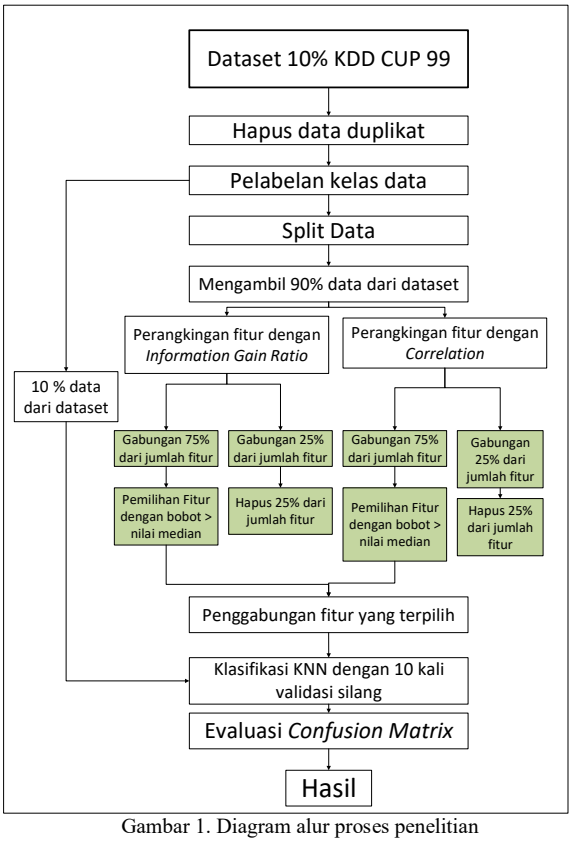
b. Selanjutnya melakukan tahap *Pre-Processing* yang dimana pada tahap ini dilakukan penghapusan data duplikat atau data yang berlebihan pada dataset 10% KDD CUP 99. Kemudian dilakukan pelabelan kelas data yang dimana data dilabelkan menjadi 5 kelas utama normal dan serangan yaitu Normal, U2R, DOS, R2L dan Probe. Setelah dilakukan pelabelan kelas data, kemudian masuk ke tahap split data yang dimana jumlah data dibagi menjadi 90% data *training* dan 10% data *testing*. Untuk 90% data *training* digunakan untuk penelitian seleksi fitur dengan menggunakan metode *Information Gain Ratio* dan *Correlation*. Hasil perankingan dari kedua metode dibagi menjadi dua kelompok. Kelompok pertama yaitu gabungan 75% peringkat fitur dan kelompok kedua gabungan 25% peringkat fitur. Pada kelompok pertama dicari nilai mediannya yang dimana fitur-fitur yang mempunyai bobot lebih besar dari nilai median dipilih untuk dikombinasikan dengan fitur-fitur dari metode seleksi fitur lain. Dan untuk fitur-fitur yang ada di kelompok dua dihapus atau tidak digunakan pada kombinasi fitur. Fitur-fitur yang dipilih dari kedua metode seleksi fitur digabungkan menjadi subset fitur baru.

c. Fitur yang sudah dipilih kemudian dijadikan input pada klasifikasi *K-Nearest Neighbor* dengan menggunakan 10 kali validasi silang. Pada tahap ini dilakukan pengujian dengan nilai k=5 pada algoritma KNN.

d. Evaluasi dari kombinasi filter pada seleksi fitur ini menggunakan *confusion matrix*. Kemudian dengan *confusion matrix* juga dilakukan perhitungan hasil tingkat akurasi, *class precision* dan *class recall*. Pada *confusion Matrix* dicari *Instance* dari kelas positif yang diklasifikasikan secara palsu sebagai negative disebut *false negative* dan *instance* dari kelas negative yang diklasifikasikan secara palsu disebut *false positive*. Jumlah pengamatan *true positive, false positive, true negatve* dan *false negative* dinotasikan dengan TP, FP, TN dan FN. Dari frekuensi ini, kita dapat menghitung indicator kinerja klasifikasi yang mencerminkan bagaimana klasifikasi melakukannya dalam mendeteksi kelas yang diberikan. Indikator yang paling umum adalah:

$$\begin{aligned} &\textit{Precision} = TP/(TP+FP) \\ &\textit{Sensitivity} = TP/(TP+FN) \\ &\textit{Specificity} = TN/(TN+FP) \\ &\textit{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \end{aligned} \quad (1)$$

(Salla, Wilhelmina, Sari, Mikaela, Pekka & Jakko, 2018).



Gambar 1. Diagram alur proses penelitian

3. TINJAUAN PUSTAKA

3.1 *Intrusion Detection System*

Intrusion Detection System (IDS) adalah komponen penting dari sistem informasi yang aman. Biasanya, penyusup dalam jaringan mencoba untuk mengakses sumber daya yang tidak sah dalam sebuah jaringan. Maka dari itu sangat diperlukan untuk memantau dan menganalisis kegiatan pengguna dan perilaku system (Selvakumar & Muneeswaran, 2018)

IDS dapat mendeteksi serangan eksternal dan mengawasi aktivitas pengguna internal yang tidak sah dengan mengidentifikasi dan merespon komunikasi jaringan berbahaya dan perilaku pengguna komputer. IDS bertujuan untuk mendeteksi serangan dengan mempelajari proses dan karakteristik perilaku serangan (Wang, Zhang & Zheng, 2016).

3.2 KDD CUP 99

KDD CUP 99 adalah dataset yang digunakan dalam kontes *Knowledge Discovery and Data Mining* (KDD) yang diadakan pada tahun 1999. Meskipun dataset ini merupakan dataset yang lama, tetapi dataset ini telah diakui secara luas dan digunakan oleh banyak peneliti. Setiap koneksi jaringan dalam dataset KDD CUP 99 ditandai sebagai normal dan serangan. Serangan dibagi menjadi 4 kategori. Keempat jenis serangan tersebut antara lain DOS, R2L, U2R dan *Probing*. Dataset KDD CUP 99 memiliki 42 fitur yang satu diantarannya merupakan fitur dari kategori label (Wang, Zhang & Zheng, 2016).

Dataset KDD CUP 99 memiliki kekurangan yaitu memiliki sejumlah besar catatan yang berlebihan dan berulang. Oleh karena itu, diperlukan untuk menghilangkan catatan yang berlebihan dari dataset KDD CUP 99 sehingga pada tahap klasifikasi dan seleksi fitur tidak akan terjadi bias terhadap catatan yang berulang (Hasan, Nasser, Ahmad & Molla, 2016).

3.3 Seleksi Fitur

Dalam deteksi intrusi, dataset yang digunakan ditandai dengan jumlah yang besar dan dimensi yang tinggi. Dengan demikian, perlu untuk melakukan reduksi dimensi untuk meningkatkan akurasi klasifikasi dan menguransi waktu komputasi. Ada dua pendekatan utama untuk mengurangi dimensionalitas yaitu transformasi fitur dan pemilihan fitur. Transformasi fitur bertujuan mengurangi dimensi data dengan membuat fitur baru dari fitur asli seperti ekstraksi fitur dan kontruksi fitur. Sedangkan pemilihan fitur bertujuan untuk memilih fitur yang relevan dan informatif dengan menghapus fitur yang tidak relevan dan berlebihan. Ketertarikan dalam pemilihan fitur meningkat dalam beberapa waktu terakhir karena meningkatnya jumlah dataset dimensi tinggi terutama di bidang *machine learning* seperti klasifikasi, pengelompokkan dan regresi (Khammassi & Krichen, 2017).

Pemeringkatan dan pemilihan fitur adalah perspektif penting dalam *intrusion detection system* untuk mendapatkan kinerja yang lebih baik. Metode peringkat dan pemilihan fitur berguna untuk mengkategorikan fitur sebagai fitur yang signifikan tinggi atau kurang. Fitur-fitur yang dipilih membantu untuk mengklasifikasikan lalu lintas data di jaringan ke normal atau serangan. Fitur yang tidak berkontribusi dalam mendeteksi berbagai jenis serangan harus dihapus untuk mendapatkan akurasi dan kecepatan yang lebih baik dalam *intrusion detection system*. Penghapusan fitur dapat membuat kinerja IDS lebih baik dalam hal perhitungan, pengurangan dimensi dan kompleksitas waktu (Akashdeep, Manzoor & Kumar, 2017).

Dalam deteksi intrusi, dataset yang digunakan ditandai dengan jumlah yang besar dan dimensi yang tinggi. Dengan demikian, perlu untuk melakukan reduksi dimensi untuk meningkatkan akurasi klasifikasi dan menguransi waktu komputasi. Ada dua pendekatan utama untuk mengurangi dimensionalitas yaitu transformasi fitur dan pemilihan fitur. Transformasi fitur bertujuan mengurangi dimensi data dengan membuat fitur baru dari fitur asli seperti ekstraksi fitur dan kontruksi fitur. Sedangkan pemilihan fitur bertujuan untuk memilih fitur yang relevan dan informatif dengan menghapus fitur yang tidak relevan dan berlebihan. Ketertarikan dalam pemilihan fitur meningkat dalam beberapa waktu terakhir karena meningkatnya jumlah dataset dimensi tinggi terutama di bidang *machine learning* seperti klasifikasi, pengelompokkan dan regresi (Khammassi & Krichen, 2017).

Pemeringkatan dan pemilihan fitur adalah perspektif penting dalam *intrusion detection system* untuk mendapatkan kinerja yang lebih baik. Metode peringkat dan pemilihan fitur berguna untuk mengkategorikan fitur sebagai fitur yang signifikan tinggi atau kurang. Fitur-fitur yang dipilih membantu untuk mengklasifikasikan lalu lintas data di jaringan ke normal atau serangan. Fitur yang tidak berkontribusi dalam mendeteksi berbagai jenis serangan harus dihapus untuk mendapatkan akurasi dan kecepatan yang lebih baik dalam *intrusion detection system*. Penghapusan fitur dapat membuat kinerja IDS lebih baik dalam hal perhitungan, pengurangan dimensi dan kompleksitas waktu (Akashdeep, Manzoor & Kumar, 2017).

3.4 *Information Gain Ratio*

Information Gain Ratio diperkenalkan dalam algoritma *Decision Tree* (C4.5) dan merupakan jenis algoritma pemilihan fitur berdasarkan prinsip *Information Gain* (Liu, Bi & Fan, 2017). Nilai *Information Gain Ratio* dari fitur teks dihitung dengan menormalkan nilai *Information Gain* dari fitur teks. Besar nilai dari *Information Gain Ratio* menunjukkan bahwa fitur teks akan berguna untuk klasifikasi. *Information Gain Ratio* melakukan proses yang berulang untuk memilih subset fitur teks dengan memanfaatkan nilai yang dihasilkan. Iterasi akan berakhir jika jumlah fitur yang ditentukan sebelumnya akan tetap sama. *Split Information* untuk fitur teks *t* dihasilkan dengan memecah kumpulan data *D* menjadi *v* pemisahan , dimana *v* adalah hasil pengujian pada fitur *t. Split Information* fitur teks *t* di definisikan sebagai :

$$\text{SplitInfo(t)} = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|} \quad (2)$$

Dimana $|D_j|$ mewakili jumlah teks milik D_j. Nilai dari *Information Gain Ratio* didefinisikan sebagai:

$$\text{GR}(t) = \frac{\text{IG}(t)}{\text{SplitInfo}\ (t)} \quad (3)$$

Dimana IG(t) menunjukkan nilai dari *Information Gain* dari fitur teks *t*.

3.5 *Correlation*

Correlation adalah metode seleksi fitur yang digunakan untuk menemukan subset fitur yang berpotensi paling relevan dengan tugas klasifikasi yang diberikan. Bagian utama dari algoritma *Correlation* adalah heuristik untuk mengevaluasi utilitas atau kelebihan dari subset fitur. Heuristik ini menunjukkan kegunaan fitur individu untuk memprediksi label kelas Bersama dengan tingkat interkorelasi diantara mereka (Mursalin, Zhang, Chen & Chawla, 2017).

$$Merit_s = \frac{kr_{\bar{c}f}}{\sqrt{k+k(k-1)r\bar{f}f}} \quad (4)$$

Dimana, $Merit_s$ adalah heuristic "merit" dari bagian fitur *S* yang berisi fitur *k*, $r\bar{c}\,f$ adalah korelasi kelas fitur rata-rata, dan $r\bar{f}\,f$ adalah rata-rata interkorelasi fitur-fitur. Heuristik bertujuan untuk menghapus fitur yag tidak

relevan dan berlebihan karena akan menjadi predictor yang buruk dari kelas.

3.6 *K-Nearest Neighbor*

Algoritma *K-Nearest Neighbor* adalah algoritma klasifikasi berbasis *instance*. Pada algoritma ini, model klasifikasi dibangun berdasarkan kesamaan antara *instance* pelatihan *k* terdekat dari *instance* tertentu. *Instance* pelatihan diwakili oleh fitur n-dimensi dan setiap *instance* sesuai dengan satu titik dalam ruang n-dimensi. Setiap *instance* pelatihan disimpan dalam ruang *instance* n-dimensional dan label kelas untuk *instance* baru ditentukan berdasarkan mayoritas *voting* label kelas dari tetangga terdekat *k* (Onan & Korukoglu, 2015).

Untuk mengukur jarak antar data, salah satu opsi yang paling popular adalah model jarak *Euclidean* yang didefinisikan sebagai :

$$d(x_i,x_j)=\sqrt{\sum_{r=1}^{n}(a_r(x_i)-a_r(x_j))^2} \quad (5)$$

dimana, $d(x_i,x_j)$ adalah jarak *Euclidean*, x_i adalah *record* ke-*i* dan x_j adalah *record* ke-*j*, *ar* adalah data ke-*r* (Nababan, Sitompul & Tulus, 2018).

3.7 *Confusion Matrix*

Confusion matrix adalah konsep dari pembelajaran mesin yang berisi informasi tentang klasifikasi aktual dan prediksi yang dilakukan oleh sistem klasifikasi. *Confusion Matrix* memiliki dua dimensi, satu dimensi di indeks oleh kelas aktual suatu objek, yang satu lagi di indeks oleh kelas yang di prediksi oleh pengklasifikasi. Pada Gambar 2 dapat dilihat bentuk dari *Confusion Matrix* untuk klasifikasi lebih dari 1 kelas yaitu kelas A1, A2 dan An. Dalam *Confusion Matrix,* Nij mewakili sejumlah sampel dari kelas Ai dan di klasifikasikan sebagai kelas Aj (Deng, Liu, Deng & Mahadevan, 2016).

		Predicted			
		A_1	... A_j ...		A_n
Actual	A_1	N_{11}	N_{1j}		N_{1n}
	A_i	N_{i1}	... N_{ij} ...		N_{in}
	A_n	N_{n1}	N_{nj}		N_{nn}

Gambar 2. *Confusion Matrix*

4. HASIL DAN PEMBAHASAN

4.1 Hasil

A. Dataset 10% KDD CUP 99

Dataset yang digunakan dalam penelitian ini adalah dataset 10% KDD CUP 99 secara keseluruhan. Jumlah data dari Dataset 10% KDD CUP 99 adalah 494021 data yang memiliki 41 fitur atribut dan ditambah 1 fitur yang merupakan fitur dari tipe serangan dan normal.

B. *Pre-Processing*

Pada tahap awal *pre-processing* adalah melakukan pelabelan kelas data. Seluruh tipe serangan yang ada pada dataset 10% KDD CUP 99 diklasifikasikan ke dalam empat kelas utama. Empat kelas utama serangan tersebut antara lain:

1. *Denial of Service* (DOS) : Penyerang mencoba untuk mencegah pengguna yang sah dari menggunakan layanan.
2. *Remote to Local* (R2L) : Penyerang tidak memiliki akun di mesin korban, tetapi mencoba untuk mendapatkan akses.
3. *User to Root* (U2R) : Penyerang memiliki akses lokal ke mesin korban dan mencoba untuk mendapatkan hak istimewa dari *super user*.
4. *Probe* : Penyerang mencoba untuk mendapatkan informasi tentang host target.

(Selvakumar & Muneeswaran, 2018).

Pengklasifikasian tipe serangan ke dalam empat kelas utama dapat dilihat pada Tabel 1.

Tabel 1. Kategori Serangan Pada Dataset 10% KDD CUP 99

Kategori Serangan	Tipe Serangan
Denial Of Service (DOS)	*Smurf, Neptune, Back, Land, Teardrop, Pod*
Remote to Local (R2L)	*Ftp_write, Guess_Passwd, Multihop, Imap, Warezclient, Warezmaster, Phf, Spy*
User to Root (U2R)	*Buffer_overflow, Rootkit, Loadmodule, Perl*
Probe	*Satan, Portsweep, Nmap, Ipsweep*

Kemudian dilakukan penghapusan data duplikat atau *redundant* pada dataset 10% KDD CUP 99. Penghapusan data duplikat dilakukan pada dataset 10% KDD CUP 99 untuk mencegah adanya bias terhadap banyaknya data yang berulang yang dapat mempengaruhi *Machine Learning* (Khammassi & Krichen, 2017). Setelah dilakukan penghapusan data duplikat, maka diperoleh jumlah data yang jumlah awalnya 494021 data menjadi 145586 data.

C. Perangkingan dan Pembagian Jumlah Fitur

Pada tahap seleksi fitur, tahap pertama yang dilakukan adalah melakukan perangkingan fitur. Pada penelitian ini menggunakan dua metode perangkingan fitur yaitu *Information Gain Ratio* dan *Correlation*. Setelah didapatkan rangking fitur dengan menggunakan kedua metode, maka didapatkan hasil berupa rangking fitur dari rangking tertinggi hingga terendah yang ditentukan berdasarkan tinggi dan rendahnya nilai bobot fitur. Kemudian seluruh fitur yang sudah dirangkingkan di masing-masing metode dan jumlahnya dibagi menjadi dua bagian. Bagian pertama yaitu 75% jumlah fitur yang terdiri dari fitur-fitur yang mempunyai peringkat dan bobot tertinggi di masing-masing metode perangkingan fitur. Dan bagian kedua yaitu 25% jumlah fitur yang terdiri dari fitur-fitur yang mempunyai peringkat dan bobot terendah di masing-masing metode perangkingan fitur.

D. Pemilihan Fitur Menggunakan Median

Pada tahap ini, jumlah fitur yang digunakan yaitu 75% jumlah fitur karena pada 75% jumlah fitur terdiri dari fitur-fitur yang memiliki peringkat dan bobot tertinggi di masing-masing metode perangkingan fitur. Dan untuk 25% jumlah fitur pada penelitian ini tidak akan digunakan untuk ke tahap selanjutnya dikarenakan 25% jumlah fitur terdiri dari fitur-fitur yang memiliki peringkat dan bobot terendah di masing-masing metode pemilihan fitur. Pada penelitian ini median digunakan untuk mencari fitur-fitur mana saja yang dapat digunakan dengan mencari nilai tengah atau nilai median dari seluruh jumlah fitur. Kemudian nilai median tersebut dijadikan patokan untuk mencari fitur-fitur mana saja yang memiliki bobot lebih dari atau sama dengan nilai median. 75% jumlah fitur yang diambil terdiri dari 30 fitur yang memiliki peringkat dan bobot tertinggi. Adapun perhitungan untuk mencari nilai median di masing-masing hasil perangkingan dari metode *Information Gain Ratio* dan *Correlation* dapat dilihat pada Tabel 2.

Tabel 2. Perhitungan Nilai Median

Metode	Perhitungan Median
Information Gain Ratio	$\text{Median} = \frac{x\frac{30}{2} + x(\frac{30}{2}+1)}{2}$ $= \frac{x(15) + x(15+1)}{2}$ $= \frac{x(15) + x(16)}{2}$ $= \frac{0.435 + 0.425}{2}$ $= \frac{0.860}{2}$ $= 0.430$
Correlation	$\text{Median} = \frac{x\frac{30}{2} + x(\frac{30}{2}+1)}{2}$ $= \frac{x(15) + x(15+1)}{2}$ $= \frac{x(15) + x(16)}{2}$ $= \frac{0.234 + 0.212}{2}$ $= \frac{0.446}{2}$ $= 0.223$

Berdasarkan dari perhitungan nilai median pada Tabel 2, maka dapat diketahui nilai median dari masing-masing jumlah fitur dari kedua metode. Untuk nilai median dari jumlah fitur pada metode *Information Gain Ratio* yaitu 0.430 dan untuk nilai median dari jumlah fitur pada metode *Correlation* yaitu 0.223. Kemudian dari hasil tersebut dapat dipilih fitur-fitur apa saja yang memiliki bobot lebih dari dan sama dengan nilai median. Berikut merupakan fitur-fitur terpilih yang memiliki bobot lebih dari sama dengan nilai median di masing-masing hasil perangkingan metode *Information Gain Ratio* dan *Correlation* yang dapat dilihat pada Tabel 3.

Tabel 3. Fitur-Fitur Terpilih Dari Hasil Perhitungan Median

No	Fitur yang Terpilih	
	Information Gain Ratio	***Correlation***
1	Count	Same_srv_rate
2	Same_srv_rate	flag
3	Src_bytes	count
4	Dst_host_serror_rate	Dst_host_srv_count
5	Dst_host_srv_serror_rate	Dst_host_srv_serror_rate
6	Srv_serror_rate	Srv_serror_rate
7	Serror_rate	Loggen_in
8	Dst_host_diff_srv_rate	Dst_host_serror_rate
9	Hot	Serror_rate
10	Dst_host_srv_count	Dst_host_same_srv_rate
11	Dst_bytes	Service
12	Flag	Dst_host_count
13	Num_shells	Dst_host_srv_rerror_rate
14	Loggen_in	Srv_diff_host_rate
15	Num_failed_logins	Srv_rerror_rate

Setelah dilakukan pemilihan fitur dengan menggunakan median dan didapatkan fitur-fitur yang terpilih dari hasil perangkingan di masing-masing metode *Information Gain Ratio* dan *Correlation*, selanjutnya menggabungkan fitur-fitur yang terpilih dari hasil perangkingan kedua metode perangkingan fitur menjadi sebuah subset fitur baru. Gabungan fitur-fitur yang terpilih dapat dilihat pada Tabel 4.

Tabel 4. Gabungan Fitur-Fitur Terpilih

No	Nama Fitur	No	Nama Fitur
1	Count	12	Flag
2	Same_srv_rate	13	Num_shells
3	Src_bytes	14	Loggen_in
4	Dst_host_serror_rate	15	Num_failed_logins
5	Dst_host_srv_serror_rate	16	Dst_host_same_srv_rate
6	Srv_serror_rate	17	Service
7	Serror_rate	18	Dst_host_count
8	Dst_host_diff_srv_rate	19	Dst_host_srv_rerror_rate
9	Hot	20	Srv_diff_host_rate
10	Dst_host_srv_count	21	Srv_rerror_rate
11	Dst_bytes		

E. Klasifikasi *K-Nearest Neighbor*

Tahap selanjutnya adalah tahap klasifikasi menggunakan metode *K-Nearest Neighbor*. Pada tahap ini, seluruh fitur yang telah dipilih dari proses seleksi fitur digunakan sebagai input dari proses klasifikasi. Pada tahap klasifikasi menggunakan 10 kali validasi silang untuk dapat mengevaluasi kinerja dari permodelan yang digunakan dan metode klasifikasi yang digunakan yaitu metode *K-Nearest Neighbor* yang dimana metode tersebut mengharuskan untuk menentukan nilai k tetangga terdekat. Pada penelitian ini ditentukan nilai k nya yaitu k=5. Berikut merupakan perbandingan hasil akurasi pada Tabel 5 dan perbandingan hasil presisi dan *recall* pada Tabel 6.

Tabel.5 Perbandingan Hasil Akurasi

Seleksi Fitur	Akurasi
Permodelan Seleksi Fitur	**99.61%**
Tanpa Seleksi Fitur	99.59%

Tabel 6. Perbandingan Hasil Presisi dan *Recall*

	Presisi		Recall	
Kelas	**Seleksi Fitur**	**Tanpa Seleksi Fitur**	**Seleksi Fitur**	**Tanpa Seleksi Fitur**
Normal	**99.80%**	99.75%	**99.86%**	99.85%
U2R	**75.00%**	72.00%	34.62%	34.62%
DOS	99.51%	**99.52%**	99.83%	**99.88%**
R2L	**96.62%**	96.35%	94.39%	**94.99%**
Probe	96.04%	**96.61%**	**87.71%**	85.50%

4.2 Pembahasan

Data yang digunakan pada penelitian ini adalah dataset 10% KDD CUP 99. Pada tahap *pre-processing,* dataset dilakukan penghapusan data duplikat untuk mencegah adanya bias terhadap banyaknya data yang berulang yang dapat mempengaruhi *Machine Learning*. Kemudian dilakukan pelabelan kelas data yang mana seluruh tipe serangan yang ada pada dataset 10% KDD CUP 99 diberi label U2R, DOS, R2L dan Probe.

Setelah didapatkan 21 fitur baru dari hasil seleksi fitur, maka selanjutnya dilakukan tahap klasifikasi menggunakan metode *K-Nearest Neighbor* dan juga menggunakan 10 kali validasi silang. Setelah dilakukan pengujian tahap klasifikasi maka hasil disajikan dalam bentuk evaluasi *confusion matrix* yang dimana pada *confusion matrix* dapat dilihat nilai dari akurasi, *precision* dan nilai *recall*. Pada perbandingan akurasi pada Tabel 5 dapat dilihat untuk permodelan seleksi fitur memiliki akurasi 99.61% yang mana akurasi tersebut lebih unggul dari tanpa menggunakan seleksi fitur dengan akurasi 99.59%. Kemudian pada Tabel 6 perbandingan nilai presisi untuk permodelan seleksi fitur unggul pada kelas Normal, U2R dan R2L dengan masing-masing nilai presisinya 99.80%, 75.00% dan 96.62%. Sedangkan untuk nilai presisi tanpa menggunakan seleksi fitur unggul pada kelas DOS dan Probe yang nilai presisinya masing-masing 99.52% dan 96.61%. Dan untuk nilai *recall* pada Tabel 6 permodelan seleksi fitur unggul pada kelas Normal dan Probe dengan nilai *recall* masing-masing 99.86% dan 87.71%. Sedangkan untuk nilai *recall* tanpa seleksi fitur unggul pada kelas DOS dan R2L dengan masing-masing nilai *recall* 99.88% dan 94.99%.

5. KESIMPULAN

Dari penelitian yang telah dilakukan, maka kesimpulan yang dapat diambil adalah sebagai berikut:

a. Penerapan kombinasi metode seleksi fitur yaitu metode *Information Gain Ratio* dan metode *Correlation* ternyata mampu mengurangi fitur-fitur yang tidak berguna dan dapat meningkatkan kinerja dari *Intrusion Detection System* untuk dapat mendeteksi perilaku serangan dengan baik. Serta dengan menerapkan metode klasifikasi *K-Nearest Neighbor* dengan menentukan nilai k=5 pada permodelan yang digunakan mampu meningkatkan nilai akurasi yang lebih tinggi daripada tanpa menggunakan seleksi fitur. Penerapan pemodelan yang diusulkan menghasilkan akurasi tertinggi sebesar 99.61%. Sedangkan untuk akurasi tanpa seleksi fitur menghasilkan akurasi tertinggi sebesar 99.59%.
b. Pada hasil nilai presisi, permodelan yang diusulkan mampu meningkatkan nilai presisi pada kelas Normal, U2R dan R2L dan juga pada hasil nilai *Recall* mampu meningkatkan nilai *Recall* pada kelas DOS dan R2L.

DAFTAR PUSTAKA

AKANDE, K.O., OWOLABI, T.O., & OLATUNJI, S.O. 2015. *Investigating The Effect Of Correlation Based Feature Selection On The Performance Of Support Vector Machines In Reservoir Characterization*. Journal of Natural Gas Science And Engineering, 22, pp.515-522.

AKASHDEEP, MANZOOR,I. & KUMAR, N. 2017. *A Feature Reduced Intrusion Detection System Using ANN Classifier*. Expert Systems With Applications, 88, pp.249-257.

CILIA, N.D., STEFANO, C.D., FONTANELLA, F., RAIMONDO, S. & FRECA, A.S. 2019. *An Experimental Comparison Of Feature Selection And Classification Methods For Microarray Datasets*. Information, 10(3), pp.1-13.

DENG, X., LIU, Q., DENG, Y. & MAHADEVAN, S. 2016. *An Improved Method To Construct Basic Probability Assignment Based On The Confusion Matrix For Classification Problem*. Information Sciences, 340-341, pp.250-261.

HASAN, M., NASSER, M., AHMAD, S. & MOLLA, K, I . 2016. *Feature Selection For Intrusion Detection Using Random Forest*. Journal Of Information Security, 7, pp.129-140.

KHAMMASSI, C. & KRICHEN, S. 2017. A GA-LR Wrapper Approach For Feature Selection In Network Intrusion Detection. Journal of Computers & Security, 70, pp.255-277.

LIU, Y., BI, J.W. & FAN, Z.P. 2017. *Multiclass Sentiment Classification : The Experimental Comparisons Of Feature Selection And Machine Learning Algorithms*. Expert Systems With Applications, 80, pp.323-339.

LUO, B. & XIA, J. 2014. *A Novel Intrusion Detection System Based On Feature Generation With*

Visualization Strategy. Expert Systems With Applications, 41, pp.4139-4147.

MURSALIN, MD., ZHANG, Y., CHEN, Y. & CHAWLA, N, V. 2017. *Automated Epileptic Seizure detection Using Improved Correlation Based Feature Selection With RandomForest Classifier.* Neurocomputing, 241, pp.204-214.

NABABAN, A.A., SITOMPUL, O, S,. & TULUS. 2018. *Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio*. Journal of Physic: Conference Series, 1007, pp.1-6.

ONAN, A. & KORUKOGLU, S. 2015. *A Feature Selection Model Based On Genetic Rank Aggregation For Text Sentiment Classification*. Journal Of Information Science, 43, pp.25-38.

SALLA, R., WILHELMIINA, H., SARI, K., MIKAELA, M., PEKKA, M. & JAAKKO, M. 2018. *Evaluation Of The Confusion Matrix Method In The Validation Of An Automated System For Measuring Feeding Behaviour Of Cattle.* Behavioural Processes, 148, pp.56-62.

SELVAKUMAR, B. & MUNEESWARAN, K. 2018. *Firefly Algorithm Based Feature Selection For Network Intrusion Detection*. Computers & Security, 81, pp.148-155.

WANG, X., ZHANG, C. & ZHENG, K. 2016. *Intrusion Detection Algorithm Based on Density, Cluster Centers, And Nearest Neighbors*. Network Coding And Algorithm,13, pp.24-31.