Tugas VI (UTS)

Johnny
Faculty of Information Technology
Institut Teknologi Batam
Batam, Indonesia
1822004@student.iteba.ac.id

Rian Sanjaya
Faculty of Information Technology
Institut Teknologi Batam
Batam, Indonesia
1822012@student.iteba.ac.id

Muhammad Riduan
Faculty of Information Technology
Institut Teknologi Batam
Batam, Indonesia
1822001@student.iteba.ac.id

***Abstract*—Abstract—Pada era transformasi digital, keamanan jaringan menjadi semakin penting dan menarik untuk dikaji. Sistem deteksi intrusi (IDS) merupakan bagian integral dari keamanan jaringan. Untuk meningkatkan keamanan jaringan, algoritma pembelajaran mesin dapat diterapkan untuk mendeteksi dan mencegah serangan jaringan. Pemanfaatan kumpulan data (dataset) seperti NSL-KDD, menjadi salah satu pendekatan untuk melatih model guna mendeteksi berbagai serangan jaringan. Pada pekerjaan ini mahasiswa diharapkan dapat membandingkan performa dari beberapa algoritma yaitu Random Forest, K-Neighbors, SVM dan Ensemble Learning Algoritma Random Forest, K-Neighbors, SVM dan Ensemble Learning dari masing-masing algoritma tersebut memiliki cara klasifikasi yang berbeda, di sini kami akan membandingkan kinerja masing-masing algoritma.**

***Index Terms*—Keywords—intrusion detection, Random Forest, K-Neighbors, SVM, Ensemble Learning, NSL-KDD dataset.**

I. PENDAHULUAN

pada era saat ini keamanan sistem maupun jaringan sangat penting khusunya pengamanan informasi. Informasi pribadi maupun informasi perusahaan sangat penting untuk dijaga jika keamanan sistem atau jaringan rusak, maka harus segera dilakukan perbaikan Pengguna internet yang terus meningkat dari tahun ke tahun tidak terlepas dari serangan yang timbul dari teknologi jaringan seperti serangan malware infection. Oleh karena itu, diperlukan keamanan dalam sistem komputer untuk mencegah dari serangan. IDS (Intrusion Detection System) merupakan sebuah aplikasi yang mampu mencatat kegiatan dalam suatu jaringan dan menganalisa paket-paket yang dikirim melalui lalu lintas jaringan secara realtime. Tujuan dari sistem ini yaitu mengawasi jika terjadi penetrasi ke dalam sistem, mengawasi traffic yang terjadi pada jaringan, mendeteksi anomaly terjadinya penyimpangan dari sistem yang normal atau tingkah laku user. Dalam melakukan deteksi serangan, dapat digunakan beberapa algoritma yaitu Random Forest, K-Neighbors, SVM dan Ensemble Learning. Dan tujuan dari penelitian ini adalah untuk membandingkan performa dari masing-masing algoritma.

Identify applicable funding agency here. If none, delete this.

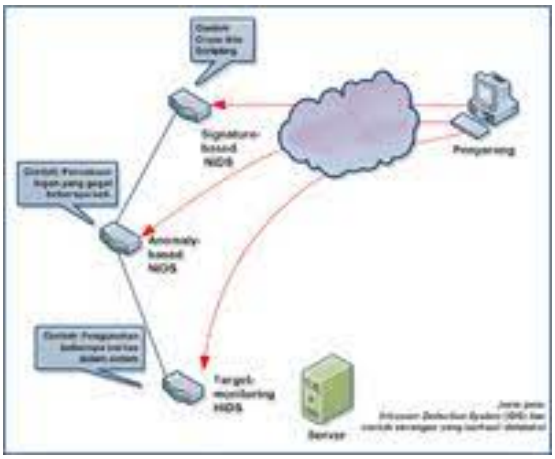

Fig. 1. Jaringan IDS

II. PEMBAHASAN

A. Intrusion Detection System (IDS)

Intrusion Detection System adalah sebuah metode yang dapat digunakan untuk mendeteksi aktivitas yang mencurigakan dalam sebuah sistem atau jaringan. IDS dapat melakukan inspeksi terhadap lalu lintas inbound dan outbound dalam sebuah sistem atau jaringan, melakukan analisis dan mencari bukti dari percobaan intrusi dan IDS merupakan sistem untuk mendeteksi adanya “intrusion” yang dilakukan oleh “intruder” atau “pengganggu atau penyusup” di jaringan. IDS (Intrusion Detection System) sangat mirip seperti alarm, yaitu IDS (Instrusion Detection System) akan memperingati bila terjadinya atau adanya penyusupan pada jaringan. IDS (Intrusion Detection System) dapat didefinisikan sebagai kegiatan yang bersifat anomaly, incorrect, inappropriate yang terjadi di jaringan atau host. IDS (Intrusion Detection System) adalah sistem keamanan yang bekerja bersama Firewall untuk mengatasi Intrusion.

B. Random Forest

Random forest adalah suatu algoritma yang digunakan pada klasifikasi data dalam jumlah yang besar. Klasifikasi random forest dilakukan melalui penggabungan pohon dengan

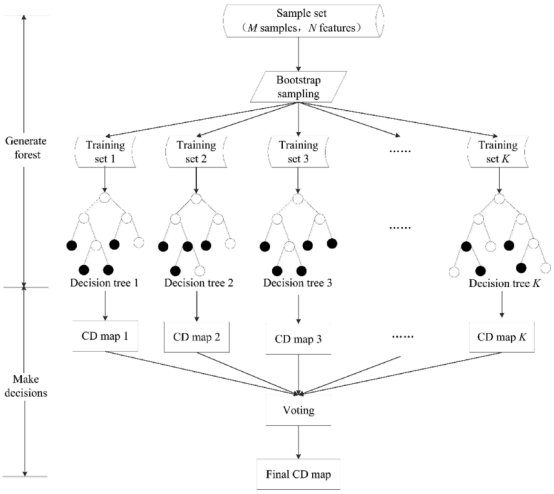

Fig. 2. Flowchart Algoritma Random Forest

melakukan training pada sampel data yang dimiliki. Keuntungan Random Forest adalah sebagai berikut [1]:

1) Hutan yang dihasilkan dapat disimpan untuk referensi di masa mendatang.

2) Hutan acak mengatasi masalah penyesuaian.

3) Dalam akurasi RF dan kepentingan variabel secara otomatis dihasilkan.

Flowchart dari proses pemodelan algoritma Random Forest dapat dilihat pada "Gambar. 1".

C. K-Nearest Neighbors

Algoritme k tetangga terdekat adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pemelajaran yang jaraknya paling dekat dengan objek tersebut. Data pemelajaran digambarkan ke ruang berdimensi banyak dengan tiap-tiap dimensi mewakili tiap ciri/fitur dari data.*Euclidean Distance*
Untuk mendefinisikan jarak antara dua titik yaitu titik pada data training (x) dan titik pada data testing (y), maka digunakan rumus *Euclidean* [**?**], yaitu:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

dimana: d = jarak antara 2 titik x = data uji y = data latih i = merepresentasikan nilai atribut n = merupakan dimensi atribut. *City Block Distance*

City Block Distance umumnya dihitung antara 2 koordinat objek yang berpasangan. Ini adalah penjumlahan dari perbedaan absolut antara 2 koordinat. *City Block Distance* 2-titik a dan b dengan dimensi k dihitung secara matematis menggunakan rumus berikut ini:

$$d_{ij} = \sum_{i=1}^{k}|a_i - b_i|$$

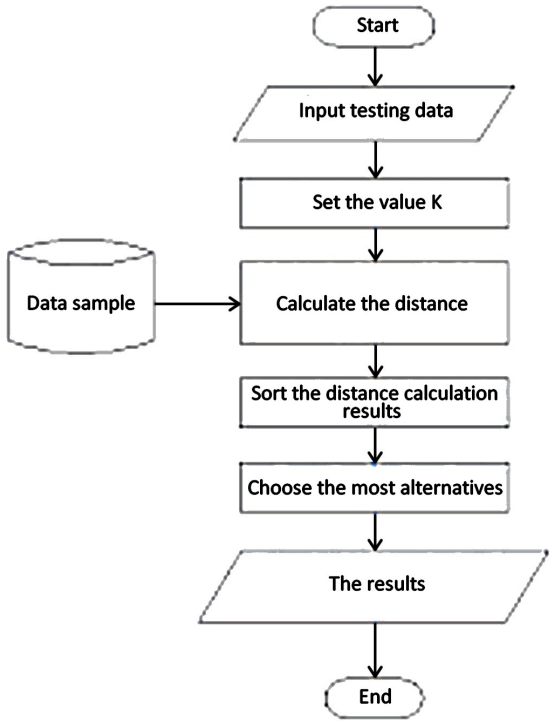

Fig. 3. Flowchart Algoritma K-Nearest Neighbors

Manhattan Distance
Manhattan Distance merupakan salah satu pengukuran yang paling banyak digunakan meliputi penggantian perbedaan kuadrat dengan menjumlahkan perbedaan *absolute* dari variabel-variabel. Fungsi ini hanya akan menjumlahkan selisih nilai x dan y dari dua buah titik. *Minkwoski Distance*

Minkwoski Distance adalah metrik dalam ruang vektor bernorma yang dapat dianggap sebagai generalisasi dari kedua jarak *Euclidean* dan jarak *Manhattan*. Jarak *Minkowski* antara dua variabel X dan Y didefinisikan sebagai:

$$d = (\sum_{i=1}^{n}|X_i - Y_i|^p)^{1/p}$$

Kasus di mana p = 1 setara dengan jarak *Manhattan* dan kasus di mana p = 2 setara dengan jarak *Euclidean*. Flowchart dari proses pemodelan algoritma *K-nearest neighbors* dapat dilihat pada "Gambar. 2" [**?**].

D. Algoritma SVM

Support Vector Machine (SVM) adalah algoritme supervised yang berupa klasifikasi dengan cara membagi data menjadi dua kelas menggunakan garis vektor yang disebut hyperplane (Octaviani, et al., 2014). Pada permasalahan yang kompleks atau permasalahan dengan parameter yang banyak, metode ini sangat baik untuk digunakan.

Teori SVM berasal dari statistik dan prinsip dasar SVM adalah menemukan *hyperplane* linier yang optimal dalam

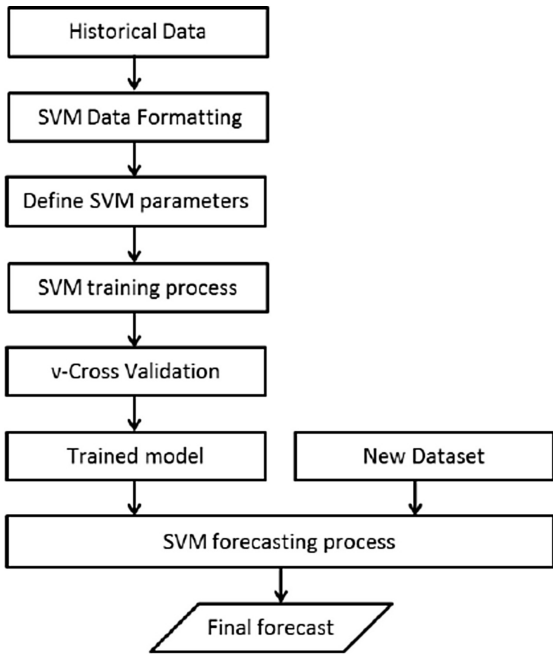

Fig. 4. Flowchart Algoritma SVM

ruang fitur yang secara maksimal memisahkan dua kelas target [**?**].

Dalam kaitannya dengan fungsi kernel, fungsi diskriminan mengambil bentuk berikut:

$$f(x) = \sum_{i}^{n} \alpha_i k(x, x_i) + b$$

Dalam pekerjaan ini, kernel *Gaussian* telah digunakan untuk membangun pengklasifikasi SVM.
Gaussian kernel:

$$K(x_i, x_j) = exp\left(-\frac{||x_i - x_j||^2}{2\sigma}\right)$$

dimana σ adalah lebar fungsi.

Fungsi kernel dan parameternya harus dipilih untuk membangun pengklasifikasi SVM. Melatih SVM menemukan *hyperplane* margin besar, yaitu menetapkan parameter α.

Flowchart dari proses pemodelan algoritma SVM dapat dilihat pada "Gambar. 3" [**?**].

E. Ensemble Learning

Metode ensemble atau metode ansamble adalah algoritma dalam pembelajaran mesin (machine learning) dimana algoritma ini sebagai pencarian solusi prediksi terbaik dibandingkan dengan algoritma yang lain karena metode ensemble ini menggunakan beberapa algoritma pembelajaran untuk pencapaian solusi prediksi yang lebih baik daripada algoritma yang bisa diperoleh dari salah satu pembelajaran algoritma kosituen saja. Tidak seperti ansamble statistika didalam mekanika statistika biasanya selalu tak terbatas. Ansemble Pembelajaran hanya terdiri dari seperangkat model alternatif yang bersifat

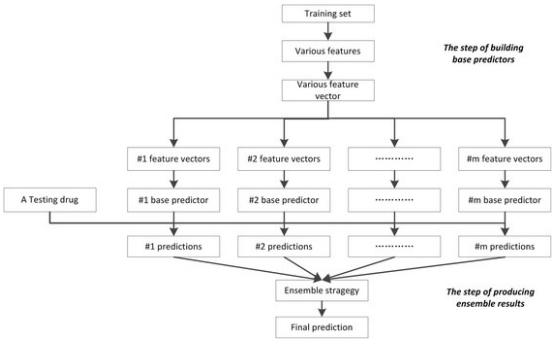

Fig. 5. Flowchart Algoritma ensemble learning

terbatas, namun biasanya memungkinkan untuk menjadi lebih banyak lagi struktur fleksibel yang ada diantara alternatif model itu sendiri.

Metode ini adalah menggabungkan beberapa fitur dengan pembelajaran *Ensemble*.

Flowchart dari proses pemodelan algoritma *Ensemble Learning* dapat dilihat pada "Gambar. 4" [**?**].

III. METODOLOGI

Metodologi adalah ilmu-ilmu/cara yang digunakan untuk memperoleh kebenaran menggunakan penelusuran dengan tata cara tertentu dalam menemukan kebenaran, tergantung dari realitas yang sedang dikaji. Metodologi tersusun dari cara-cara yang terstruktur untuk memperoleh ilmu.

A. Random Forest

Random Forest merupakan salah satu metode yang digunakan untuk menyelesaikan permasalahan. Metode ini merupakan metode pohon gabungan yang berasal dari metode classification and regression tree (CART) dan didasarkan pada teknik pohon keputusan (decision tree), sehingga mampu mengatasi masalah non-linier. Dalam random forest, banyak pohon ditumbuhkan, sehingga terbentuk suatu hutan (forest). Analisis selanjutnya akan dilakukan pada kelompok hutan tersebut.

B. K-Nearest Neighbor

Algoritma K-Nearest Neighbor (K-NN) adalah sebuah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumya. Termasuk dalam supervised learning, dimana hasil query instance yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam K-NN. Tahapan Langkah Algoritma K-NN

- Menentukan parameter k (jumlah tetangga paling dekat).
- Menghitung kuadrat jarak eucliden objek terhadap data training yang diberikan.
- Mengurutkan hasil no 2 secara ascending (berurutan dari nilai tinggi ke rendah)
- Mengumpulkan kategori Y (Klasifikasi nearest neighbor berdasarkan nilai k)
- Dengan menggunakan kategori nearest neighbor yang paling mayoritas maka dapat dipredisikan kategori objek.

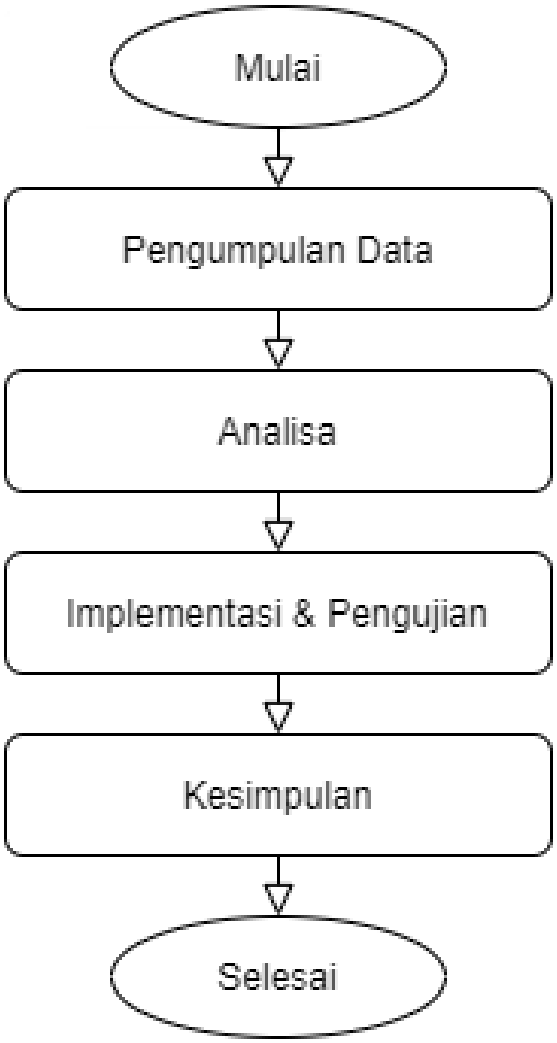

Fig. 6. contoh tahapan

C. Structural Risk Minimization (SVM)

SVM digunakan untuk mencari hyperplane terbaik dengan memaksimalkan jarak antar kelas. Hyperplane adalah sebuah fungsi yang dapat digunakan untuk pemisah antar kelas. Dalam 2-D fungsi yang digunakan untuk klasifikasi antar kelas disebut sebagai line whereas, fungsi yang digunakan untuk klasifikasi antas kelas dalam 3-D disebut plane similarly, sedangan fungsi yang digunakan untuk klasifikasi di dalam ruang kelas dimensi yang lebih tinggi di sebut hyperplane.

Tahapan pada penelitian ini dapat dijelaskan sebagai berikut:

Pengumpulan Data

Pengumpulan data yang dilakukan dengan membaca dan mempelajari penelitian sebelumnya yang berhubungan dengan IDS.

Analisa

Pada tahap ini adalah menganalisa data yaitu data latih yang digunakan untuk standarisasi melakukan pengujian, dan data uji yang digunakan untuk mengetes penilaian yang dihasilkan dari data latih.

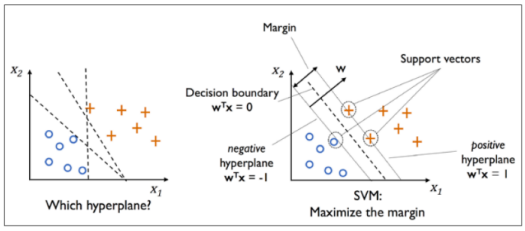

Fig. 7. contoh hyperplane

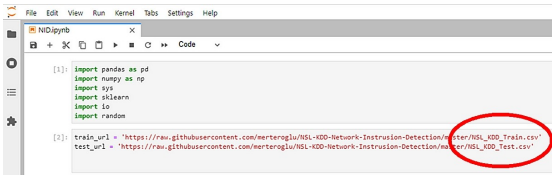

Fig. 8. contoh hyperplane

Tahap ini juga menganalisa metode yang digunakan dalam penelitian yang berkaitan dengan sistem yang digunakan.

IV. HASIL DAN PEMBAHASAN

Untuk melakukan pelatihan/pengujian digunakan dataset NSL-KDD dimana NSL KDD Train sebagai data latih dan NSL KDD Test sebagai data uji seperti pada “Gambar. 6” [8].

Dengan *confusion matrix* dilakukan penghitungan *Accuracy*, *Precision*, *Recall*, dan *F-measure* dari nilai masing-masing dalam matriks dengan menerapkan persamaan berikut:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - measure = 2 * \frac{precision * recall}{precision + recall}$$

dimana:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Dari hasil evaluasi kinerja dari masing-masing model atau algoritma, dapat dilihat pada tabel-tabel berikut:

V. KESIMPULAN

Dalam penelitian ini, kami membandingkan beberapa model untuk sistem deteksi trusi menggunakan Random Forest, KNeighbors, Support Vector Machine, dan Ensemble Learning dengan ketiga model diatas. Performa keempat pendekatan ini telah diamati berdasarkan accuracy, precision, recall, dan fmeasure (F1-score). Dari hasil pengujian dari masing-masing

algoritma yang ada pada tabel, menunjukkan kemampuan klasifikasi algoritma Ensemble Learning lebih tinggi tingkat akurasi dan ketepatan. Hasil penelitian ini sangat berguna untuk penelitian masa depan dengan cara memaksimalkan tingkat kinerja serta meminimalkan tingkat false negative.

Pada saat keamanan komputer sangat penting. salah satu cara meningkatkan keamanan komputer adalah kita menggunakan IDS. Ada beberapa algoritma yang dapat digunakan, masing-masing algoritama memiliki kelebihan dan kekurangan. Dengan tersebut keamanan komputer akan terus terjaga.

VI. REFERENSI

[1] N. Farnaaz and M. Jabbar, "Random forest modeling for network intrusion detection system," Procedia Computer Science, vol. 89, pp. 213–217, 2016.

[2] Y. Liao and R. Vemuri, "Use of k-nearest neighbor classifier for intrusion detection," Computers and Security, vol. 21, pp. 439–448, 10 2002.

[3] N. Nurhadi, Aplikasi Intelligence Intrusion Detection System (IIDS) Dengan Menggunakan Metode K-Nearest Neighbor Untuk Mendeteksi Serangan Pada Jaringan. PhD thesis, Universitas Islam Negeri Sultan Syarif Kasim Riau, 2017.

[4] Z. Lubis, P. Sihombing, and H. Mawengkang, "Optimization of k value at the k-nn algorithm in clustering using the expectation maximization al- gorithm," in IOP Conference Series: Materials Science and Engineering, p. 012133, IOP Publishing, 2020.

[5] M. A. Hasan, M. Nasser, B. Pal, and S. Ahmad, "Support vector machine and random forest modeling for intrusion detection system (ids)," Journal of Intelligent Learning Systems and Applications, vol. 06, pp. 45–52, 01 2014.

[6] E. Xydas, C. Marmaras, L. Cipcigan, A. Sani Hassan, and N. Jenkins, "Forecasting electric vehicle charging demand using support vector ma- chines," Proceedings of the Universities Power Engineering Conference, pp. 1–6, 09 2013.

[7] W. Zhang, F. Liu, L. Longqiang, and J. Zhang, "Predicting drug side effects by multi-label learning and ensemble learning," BMC Bioinfor- matics, vol. 16, 11 2015.

[8] Mamcose, "Nsl-kdd-network-intrusion-detection," 2019.