PENERAPAN *NAIVE BAYES* PADA *INTRUSION DETECTION SYSTEM* DENGAN DISKRITISASI VARIABEL

I Nyoman Trisna Wirawan[1)], Ivan Eksistyanto[2)]
[1, 2)]Institut Teknologi Sepuluh Nopember Surabaya
Jalan Teknik Kimia, Gedung Teknik Informatika
Kampus ITS Sukolilo, Surabaya, 60111
e-mail: wirawan14@mhs.if.its.ac.id[1)], ivan14@mhs.if.its.ac.id[2)]

ABSTRAK

Intrusion Detection System (IDS) merupakan sebuah perangkat lunak atau perangkat keras yang dapat digunakan untuk mendeteksi adanya aktivitas yang tidak wajar dalam jaringan. Teknik data mining telah banyak diterapkan dalam proses deteksi seperti decision tree, naive bayes, algoritma genetika, dan teknik machine learning lainnya. IDS membutuhkan performansi yang relatif cepat dengan tingkat false positif yang rendah sehingga hal ini menjadi masalah yang menarik untuk dipecahkan. Penerapan algoritma naive bayes pada masalah i dapat dilakukan namun kelemahan dari naive bayes sendiri adalah memerlukan atribut dengan nilai diskrit sehingga diperlukan proses diskritisasi untuk merubah atribut kontinu kedalam bentuk diskrit.

Pada penelitian ini akan dibahas mengenai penerapan *naive bayes classifier* dengan menggunakan pemilihan atribut berdasarkan pada korelasi serta preprocessing data dengan diskritisasi dengan menggunakan metode mean/standar deviasi untuk atribut kontinu dengan menggunakan 3-interval dan 5-interval. Hasil percobaan menunjukan bahwa penerapan naive bayes pada klasifikasi data yang telah melewati proses diskritisasi mampu memberikan akurasi hingga 89% dengan running time rata-rata adalah 31 detik.

***Kata Kunci**: Intrusion Detection System, Diskritisasi, Naive Bayes.*

ABSTRACT

Intrusion Detection System (IDS) is a software or hardware that can be used to detect any unusual activity in the network. Data mining techniques have been widely applied in the detection process such as decision tree, naive Bayes, genetic algorithms, machine learning and other techniques. IDS requires a relatively fast performance with a low rate of false positives so that this becomes an interesting problem to solve. Application of Naive Bayes algorithm on this problem can be done but the weakness of Naive Bayes itself is require attributes with discrete values so that the necessary processes to change the attributes of continuous dikritisasi into discrete shapes.

In this study will be discussed on the application of Naive Bayes classifier using attribute selection based on correlation and data preprocessing discretization by using the mean / standard deviation for continuous attributes using the 3-and 5-interval. The result showed that the application of the Naive Bayes classification of data which has passed through the process of discretization able to provide accuracy until 89% with an average running time is 31 seconds.

***Keywords**: Intrusion Detection System, Discritization, Naive Bayes*

I. PENDAHULUAN

Perkembangan komputer dan jaringan internet menjadikan keamanan jaringan menjadi hal yang sangat penting dalam teknologi komputer. Denning pada tahun 1987 mengusulkan tentang mendeteksi dan mengidentifikasi adanya intrusi (Intrusion Detection System), kemudian dikembangkan kembali dengan membagi IDS kedalam dua jenis yaitu *misuse detection* dan *anomaly detection*. *Misuse detection* memanfaatkan pencocokan pola berdasarkan database yang telah didefinisikan sebelumnya, sedangkan *anomaly detection* memanfaatkan ketidaknormalan dari aktivitas yang dilakukan pada jaringan jika dibandingkan dengan kondisi jaringan pada saat normal.

Intrusion Detection System (IDS) merupakan sebuah sistem perangkat lunak atau perangkat keras yang dapat digunakan untuk mendeteksi adanya aktivitas yang mencurigakan dalam sistem atau jaringan computer [1]. Permasalahan IDS ini telah didekati dengan menggunakan beberapa algoritma dalam kecerdasaran buatan seperti *decision tree*, *naive bayes*, *support vector machine* (SVM), jaringan syaraf tiruan dan algoritma lainnya. Pada penelitian [2] performansi dari setiap algoritma tersebut telah dibandingkan dengan menggunakan KDD99 dataset. Skenario pengujian yang digunakan adalah mendeteksi besarnya akurasi, *detection rate*, *false alarm rate*, *running time* berdasarkan pada 1 jenis anomali diantara DoS, Probe, R2L, U2R.

Penelitian ini bertujuan untuk menerapkan dan menganalisa *naive bayes* dengan dataset yang memiliki 4 jenis serangan yaitu DoS, Probe, R2L, dan U2R. Pemilihan fitur dari dataset yang akan digunakan menggunakan teknik *corellation feature selection* (CFS). Algoritma *naive bayes* didasarkan pada tingkat probabilitas nilai dari suatu atribut terhadap kelasnya, teknik ini akan menghasilkan nilai probabilitas yang sangat kecil jika terdapat

sangat banyak nilai yang berbeda dalam suatu atribut, sebagai contoh penerapan pada IDS dataset yang digunakan adalah dengan tipe kontinu, sehingga nilai probabilitas menunjuk kemungkinan nilai yang sama keluar pada suatu kelas namun pada sisi lain rentang nilai pada atribut tersebut sangat besar sehingga nilai probabilitas dari nilai tersebut muncul kembali dalam suatu kelas akan sangat kecil, hal ini akan menyebabkan lemahnya performansi dari *naive bayes*, untuk mengatasi hal tersebut dilakukan pendekatan teknik diskritisasi dengan menggunakan mean/standar deviasi yang berfungsi untuk mengelompokkan data kedalam beberapa kategori. Performansi dari penerapan teknik ini akan dinilai dari 4 aspek yang telah digunakan pada penelitian sebelumnya.

Penyusunan penulisan dilakukan sebagai berikut, Bagian 2, menjelaskan penelitian terkait *Intrusion Detection System*. Bagian 3, menjelaskan teknik yang akan digunakan pada pemilihan fitur, normalisasi, dan dikritisasi dalam KDD99 dataset. Bagian 4, menjelaskan hasil ekseperimen dan analisa hasil yang telah dilakukan. Bagian Intrusion Detection System (IDS) dan penelitian terkait

A. *Intrusion Detection System* (IDS)

Intrusion Detection System adalah sistem yang membantu sistem informasi mempersiapkan dan melakukan pengawasan menghadapi serangan terhadap traffic jaringan. Terdapat 2 teknik Intrusion detection yaitu *Anomaly Detection* dan *Misuse detection* [1] dimana,

- Misuse Detection
  Misuse detection merupakan teknik mendeteksi adanya serangan dengan mengamati aktifitas sistem, mencari pola dari aktifitas tersebut dan mencocokannya dengan pola perilaku serangan yang telah didefinisikan dalam database, pola dari serangan tersebut dapat disebut sebagai *signatures*. Tingkat deteksi dari teknik ini sangat tergantung pada pola serangan yang telah didefinisikan dalam database. Teknik ini membutuhlan waktu eksekusi yang relatif singkat untuk proses pencocokan pola, namun kelemahan dari teknik adalah terhadap jenis serangan yang mampu memodifikasi dirinya sendiri maka pola dari setiap variasi serangan tersebut harus didefinisikan dalam jaringan.
- Anomaly detection
  Anomaly Detection meruapakan teknik kedua yang dapat digunakan dalam permasalahan IDS dimana teknik ini akan mendeteksi adanya aktifitas yang tidak wajar dari aktifitas jaringan pada kondisi biasanya teknik ini dapat mendeteksi adanya serangan dengan tipe baru namun definisi dari keadaan normal harus dapat dipastikan terlebih dahulu. Kelemahan dari teknik ini adalah jika pola serangan yang terjadi masuk kedalam aktifitas jaringan normal sehingga serangan tersebut tidak akan terdapat terdeteksi.

Untuk penerapannya sendiri IDS dapat dibedakan kedalam 2 jenis yaitu host based IDS (HIDS) dan network based IDS (NIDS) kedua teknik ini memiliki kemiripan yang sangat besar namun hanya dibedakan pada perangkat keras tempat dari sistem IDS ini terpasang.

B. Studi Literatur

Penelitian dalam bidang *Intrusion Detection System* (IDS) telah sangat banyak dilakukan baik dalam bidang keilmuan kecerdsan buatan ataupun teknik dalam bidang ilmu lainnya. Pada penelitian [2] telah dilakukan sebuah survei dengan menggunakan beberapa algoritma pada lingkungan kecerdasan buatan seperti naive bayes, J48, NBTree, dan Random Forent. Penelitian ini menggunakan dataset KDD99 dengan memodifikasi dimana untuk proses *learning* dan *testing* dataset yang digunakan adalah dataset yang mengandung barisan data dengan kelas normal dan sebuah jenis serangan diantara DoS, Probe, R2L, dan U2R. Performansi dari naive bayes pada penilitian ini sangat beragam tergantung pada jenis kombinasi dataset yang digunakan. Tingkat akurasi deteksi dari naive bayes tertinggi diperoleh dengan menggunakan data dengan kombinasi antara kelas normal dan U2R, namun naive bayes hanya mampu mencapai tingkat akurasi 80% pada data dengan kombinasi kelas normal dan DoS. Sedangkan detection rate untuk naive bayes terendah berada pada 39% yang diperoleh dengan menggunakan dataset kelas normal dan R2L, sedangkan detection rate tertinggi diperoleh sebesar 84% untuk dataset kelas normal dan kelas probe. Penelitian ini menggunakan *feature selection* dengan algoritma CFS yang didasarkan pada korelasi antara atributnya yang sangat kecil dan korelasi antar atribut dan kelasnya sangat tinggi. Akurasi yang diperoleh hanya menggambarkan kemampuan sistem untuk mendeteksi hanya sebuah jenis serangan.

Penelitian [3] digunakan dataset NSL.KDD99 yang telah mengalami proses pembersihan sebelumnya. Pada penelitian ini digunakan kolaborasi dari metode *feature selection* dengan menggunakan korelasi, kemudian dilakukan diskritisasi terhadap kebutuhan atribut dengan nilai diskrit dengan menggunakan *equal width discritization*. Teknik *preprocessing* dan *naive bayes* ini memberikan nilai akurasi 88.20% dan *error rate* sebesar 11,20%. Teknik dikritisasi dengan *equal width discritization* menggunakan besarnya rentang data yang sama untuk membentuk suatu nilai diskrit. Teknik ini meninggalkan sedikit kelamahan dalam pembentukan nilai diskrit dimana

terdapat terdapat kemungkinan beberapa nilai yang sebenarnya tidak digunakan dalam proses *diskritization* sehingga probabilitas dari atribut ini akan tetap dihitung untuk menentukan kelas dari suatu baris data.

II. METODE PENYELESAIAN

A. NSL.KDD99 *Dataset*

Penelitian [4] membahas dan mengusulkan penggunaan dataset NSL-KDD untuk pengujian metode pada permasalahan Intrusion Detection System. NSL-KDD Cup 99 merupakan data yang diusulkan untuk mengurangi masalah mendasar pada KDD 99 yang telah dibahas pada [5]. KDD 99 sangat mempengaruhi kinerja dari sebuah sistem, dan hasil estimasi dari pendekatan *anomaly detection*. Untuk mengatasi masalah ini, dataset yang akan digunakan adalah NSL-KDD99 yang terdiri dari *record* yang dipilih dari dataset KDD Cup 99. [6] Keuntungan menggunakan dataset ini adalah :

- Tidak ada *record* yang berlebihan di dalam *train set*, jadi classifier tidak akan menghasilkan hasil yang bias.
- Tidak ada duplikat *record* pada *test set* yang memiliki *reduction rates* yang lebih baik.
- Jumlah record yang dipilih dari setiap level grup yang berbeda berbanding terbalik dengan persentasi *record* didalam dataset KDD asli.

Data training terdiri dari 21 jenis serangan yang berbeda. Pada penelitian [7] 21 jenis serangan yang terdapat pada data training dikelompokkan menjadi 4 kategori yaitu DoS, R2L, U2R, dan Probe, seperti yang terdapat pada Tabel I dibawah ini.

Tabel I Kategori serangan pada IDS

Normal (13449)	**DOS (9422)**	**R2L (46)**	**U2R (19)**	**Probing (2289)**
Normal (13449)	Land (8), Pod (38) Teardrop (188) Back (196) Neptune (8282) Smurf (529) Warezclient (181)	Spy (19) Phf (2) Multihop (2) ftp_write (1) Imap (5) Warezmaster (7) Guess_passwd (10)	Buffer_overflow (6) Rootkit (4) Loadmodule (9)	Nmap (301) Portsweep (587) Ipswep (710) Satan (691)

B. *Naive Bayes Classifier*

Teorema bayes adalah teorema yang digunakan dalam statistika untung menghitung peluang dari suatu hipotesis, Bayes menghitung peluang suatu kelas berdasarkan pada atribut yang dimiliki dan menentukan kelas yang memiliki probabilitas paling tinggi. Pada machine learning, naive bayes mengklasifikasikan kelas berdasarkan pada probabilitas sederhana dengan mangasumsikan bahwa setiap atribut dalam data tersebut bersifat saling terpisah. Naive bayes telah dipelajari secara mendalam sejak tahun 1950, namun diperkenalkan dengan nama yang berbeda pada tahun 1960 [*8*]. Metode *naive bayes* meruapakan salah satu metode yang banyak digunakan berdasarkan beberapa sifatnya yang sederhana, metode naive bayes mengklasifikasikan data berdasarkan atribut data yang dinyatakan dengan $x = (x_1, x_2, x_3, \ldots, x_n)$ pada model probablitias setiap kelas k yang dapat dituliskan seperti Persamaan 1 [2] berikut

$$p(C_k|x_1, x_2, \ldots, x_n) \tag{1}$$

dimana n menyatakan jumlah atribut pada data tersebut dan k merupakan jumlah kelas pada data tersebut. Klasifikasi merupakan skema penentuan sebuah data tertentu masuk kedalam sebuah kelas yang dilihat dari sudut pandang peluang kedalam aturan Bayes dapat dituliskan seperti Persamaan 2 [2] berikut:

$$p(C_k|x_n) = \frac{p(C_k)p(x_n|C_k)}{p(x_n)} \tag{2}$$

untuk menentukan pilihan kelas yang optimal maka akan dipilih nilai peluang terbesar dari setiap kemungkinan kelas yang ada dengan menggunakan Persamaan 3 namun karena nilai $p(x_n)$ yang selalu sama untuk setiap kelas makan nilai tersebut dapat diabaikan, sehingga persamaan dapat dituliskan seperti pada Persamaan 4 [2]berikut :

$$argmax_{c \in C} \frac{p(C_k)p(x_n|C_k)}{p(x_n)} \quad (3)$$

$$argmax_{c \in C} \ (C_k)p(x_n|C_k) \quad (4)$$

Teorema bayes pada umumnya membutuhkan perkalian kartesius antara setiap atribut yang mungkin, jika misalkan terdapat 16 atribut dengan tipe *boolean* maka data latih minimal yang dibutuhkan oleh teorama bayes adalah sebanyak 2^{16} atau 65.536 baris data sehingga beberapa masalah yang mungkin muncul dari keterbatasan tersebut adalah : 1) kebanyakan data latih tidak memenuhi standar data minimal yang dibutuhkan oleh teorema bayes karena menggunakan sample data, 2) jumlah atribut dalam dataset mungkin lebih besar dari 16, 3) tipe data untuk setiap atribut mungkin memiliki jenis lebih dari 2 seperti tipe data diskrit, kontinu, nominal, dan lainnya, 4) jika pola data X tidak terdapat dalam dataset maka besar kemungkinan data tersebut tidak akan dapat diklasifikasikan dengan tepat karena probabilitas X untuk setiap kelas adalah sama.

Pada perkembangannya untuk mengatasi masalah yang mungkin terjadi tersebut, diusulkan sebuah jenis dari teorama Bayes yang dikenal dengan nama Naive Bayes yang menggunakan Teorema Bayes dengan mengasumsikan bahwa setiap atribut bersifat saling bebas. Asumsi tersebut akan menghilangkan kebutuhan akan banyaknya data latih yang dibutuhkan oleh Teorema Bayes, sehingga persamaan untuk menentukan kelas dari sebuah data dapat ditulisakan seperti Persamaan 5 [2]berikut :

$$c(x_i) = \arg\max P(c) \prod_{i=1}^{n} p(x_i|c) \quad (5)$$

C. Normalisasi

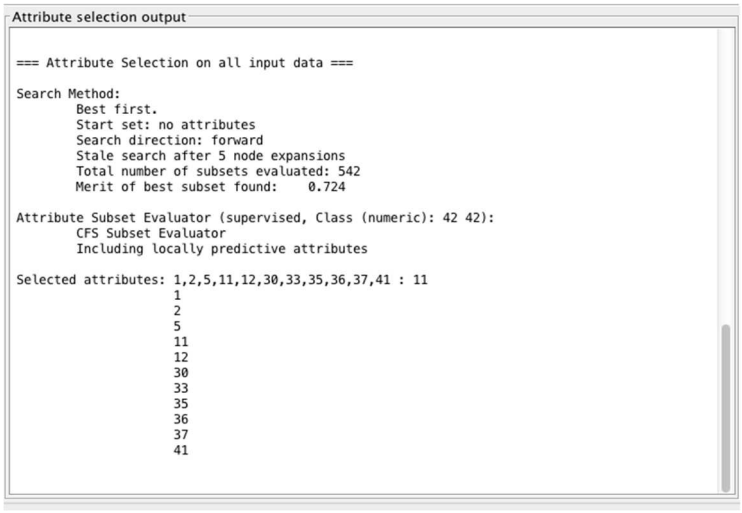
Normalisasi merupakan proses penskalaan nilai atribut dari suatu data tertentu menjadi nilai dalam rentang tertentu. Normalisasi dilakukan bertujuan untuk mengurangi adanya kesalahan pada proses data mining. Pada umumnya teknik normalisasi ini dapat dibagi kedalam 5 jenis yaitu : 1) *Min-Max*, 2) *Z-Score*, 3) *Decimal Scaling*, 4) *Sigmoidal*, 5) *Softmax*. Metode *Max-Min* merupakan metode paling sederhana dengan melakukan transformasi linier terhadap data asli, kelebihan dari metode ini adalah keseimbangan nilai perbandingan antara nilai sebelum melewati proses normalisasi dan nilai setelah melewati proses normalisasi. Secara umum fungsi normalisasi *min-max* dapat dituliskan sebagai Persamaan 6 [4]berikut :

$$new_{data} = \frac{(old_{data} - \text{Min})}{(\text{Max} - \text{Min})} \quad (6)$$

Dengan menggunakan Persamaan 6 diatas maka rentang nilai baru yang dihasilkan akan berada pada rentang 0 dan 1, dimana *min* merupakan nilai terkecil yang muncul dari atribut tersebut sedangkan *max* merupakan nilai terbesar dari atribut tersebut. Kelemahan utama dari metode ini adalah kemungkinan adanya nilai baru yang lebih tinggi dari nilai *max* atau lebih kecil dari nilai *min* yang telah ditetapkan sehingga akan memunculkan *error*.

D. *Feature Selection*

Proses seleksi atribut merupakan tahapan untuk memilih atribut atribut dalam *dateset* yang memiliki korelasi terhadap kelasnya [*9*]. *Feature selection* dapat didekati dengan beberapa metode, antara lain adalah *wrapper*, *filter*, dan *embedded*. Fitur yang biasanya digunakan dalam pemilihan atribut adalah nilai korelasi dan *mutual information*. *Corellation Feature Selection* (CFS) merupakan teknik yang menerapkan fitur korelasi dan *mutual information*. CFS memiilih atribut dengan menghitung korelasi atara atribut dan kelas serta korelasi antara atribut dan atribut lainnya, CFS akan memilih atribut dengan nilai korelasi dengan kelas yang sangat tinggi namun nilai korelasi antar atribut memiliki nilai yang kecil. Berdasarkan pada sifat tersebut CFS dapat dedifinisikan pada Persamaan 7 [9].



Gambar 1 Feature Selection dengan Algoritma CFS

$$CFS = \max\left(\frac{\left((r_{cf1} + r_{cf2} + \cdots + r_{cfn}\right)}{\sqrt{k + 2(r_{f1f2} + \cdots + r_{fifj} + \cdots + r_{fkf1})}}\right) \quad (7)$$

CFS akan memilih sejumlah atribut yang memiliki nilai terbesar yang menyatakan kelayakan dari atribut tersebut untuk dipilih, r_{cf1} menyatakan korelasi atribut f1 dengan kelas c, dan r_{f1f2} adalah nilai korelasi antara fitur 1 dan fitur 2 dalam dataset tersebut. Pada penelitian ini proses *feature selection* akan dilakukan dengan menggunakan *tools* WEKA. WEKA menyediakan algoritma CFS pada modul *feature selection*. Pada hasil percobaan menggunakan *tools* WEKA dengan *dataset* yang telah melewati proses normalisasi diperoleh hasil yang disajikan dalam Gambar 1.

Pada Gambar 1 terlihat bahwa hasil proses *feature selection* menunjukan bahwadari 42 atribut yang dilibatkan dalam *dataset* NSL.KDD99 terpilih 11 buah atribut dengan nilai korelasi antara atribut yang sangat kecil namun memiliki nilai korelasi yang besar antara atribut dan kelas. Atribut yang terpilih dari proses *feature selection* ini dengan menggunakan CFS adalah atribut { 1,2,5,11,12,30,33,35,36,37,41 } dengan kelas atribut 42.

E. Diskritisasi Variabel

Discretization (pendiskritan) atribut merupakan teknik untuk merubah sebuah fungsi atau nilai kontinu kedalam bentuk diskrit. Teknik ini dilakukan sebagai penyesuai terhadap kemungkinan kemunculan nilai kontinu dalam fitur dataset yang sangat kecil sehingga akan mempengaruhi proses klasifikasi dengan menggunakan metode *naive-bayes*. Pediskritan dari sebuah nilai kontinu tentu akan menimbulkan kesalahan berupa hilangnya beberapa informasi, pada penelitian ini akan dicoba dengan menggunakan teknik *binning* yang telah diimplementasikan pada WEKA atau Clementine 12.0. Teknik binning yang digunakan pada penelitian ini adalah standar deviasi 3 interval dan 5-interval.

Bin 1	*Bin* 2	*Bin* 3
x < (*Mean* – Std. Dev)	(*Mean* – Std. Dev) <= x <= (*Mean* + Std. Dev)	x > (*Mean* + Std. Dev)

Dengan menggunakan pembagian tersebut maka atribut dengan nilai continous tersebut akan dirubah kedalam bentuk diskrit dengan nilai 1,2 atau 3. Pada *binning* dengan menggunakan 3 atau 5 interval maka *binning* yang akan dibentuk sesuai dengan jumlah interval yang digunakan. Proses *binning* pada penilitian ini dilakukan dengan menggunakan *tools* Clementine 12.0.

F. Parameter Penilaian

Pada bagian ini akan dibahas mengenai parameter penilaian yang digunakan pada hasil penelitian yang telah dilakukan, beberapa parameter tersebut adalah akurasi, *detection rate*, *false positif rate*, *running time*. Penjelasan untuk masing masing parameter dijelaskan pada bagian berikut :

1) Akurasi (*Accuracy)*

Akurasi merupakan parameter yang menilai ketepatan hasil percobaan yang dilakukan terhadap data aktual yang digunakan. Akursi dapat diperoleh dengan membandingkan hasil deteksi yang dilakukan oleh sistem yang disusun dan data test aktual yang digunakan. Pada kasus IDS ini akurasi diperoleh dengan membandingkan jumlah data yang berhasil dengan benar diklasifikasikan kedalam 5 buah kelas Normal, DoS, Probe, R2l, dan U2R. Akurasi dapat dihitung dengan menggunakan Persamaan 8 [*10*].

$$Akurasi\ (\%) = \frac{TP + FN}{TP + FP + TN + FN} * 100 \tag{8}$$

TP dan FN merupakan nilai yang menyatakan besarnya data yang benar diklasifikasikan dengan menggunakan 5 kelas terhadap terhadap jumlah data *testing* yang digunakan secara keseluruhan, sehingga akurasi dengan nilai tinggi menyatakan bahwa metode dan skenario yang digunakan tersebut baik dalam pengelompokkan jenis serangan.

2) False positif rate

False poditif rate merupakan banyaknya data yang memiliki kelas serangan yang berhasil diklasifikasikan dengan benar berdasarkan pada metode yang diterapkan.

3) Detection rate

Detection rate merupakan parameter penilaian yang menyatakan besarnya klasifikasi benar kedalam 2 kelas yang mampu dihasilkan oleh metode tersebut. *Detection rate* tidak mempertimbangkan kesalahan deteksi terhadap 4 jenis serangan yang terdapat dalam *dataset* namun hanya membedakan kondisi normal dan kondisi anomali. Perhitungan *nilai detection* rate dapat dilakukan dengan menggunakan persamaan 9 [*10*].

$$detection.rate\ \ (\%) = \frac{TP}{TP + FP} * 100 \tag{9}$$

Nilai TP merupakan nilai *true_positif* yang menyatakan jumlah data serangan yang benar terdeteksi sebagai serangan sedangkan FP merupakan jumlah data normal yang diklasifikasikan sebagai serangan. Parameter ini menyatakan banyaknya serangan yang terdeteksi dan benar merupakan sebuah serangan.

4) Running time

Running time merupakan jumlah waktu yang dibutuhkan untuk proses learning dan *testing* dilakukan sehingga diperoleh beberapa parameter yang telah dijelaskan pada bagian sebelumnya, satuan dari parameter penilaian *running time* dinyatakan sebagai detik (s)

III. HASIL PERCOBAAN

Percobaan pada penelitian ini dilakukan dengan menggunakan NSL-KDD99 yang telah mengalami proses pembersihan data sebelumnya dari *missing value* dan *outliers*. Data yang digunakan pada penelitian ini adalah sebesar 20% yang merupakan bagian data yang asli dan mungkin tidak mencakup semua pola serangan yang terdiri dari 22 jenis anomali. 22 jenis anomali tersebut akan dikelompokan kedalam 5 buah kelas berdasarkan pada penelitian [2], sehingga akan terbentuk kelas dengan nilai DOS, R2L, U2R, Probe, dan Normal. Hasil ekseperimen akan menggambarkan kemampuan dari setiap skenario dalam mengklasifikasikan suatu data kedalam 5 kelas tersebut beserta dengan hasil analisisnya. Berikut merupakan hasil eksperimen untuk setiap skenario yang dilakukan:

A. *Min-Max* normalisasi tanpa proses diskritisasi

Pada skenario ini data yang digunakan dalam percobaan telah melewati tahapan normalisasi dengan menggunakan metode Min-max yang telah dibahas pada bagian sebelumnya, setelah melewati proses normalisasi kemudian data tersebut melewati proses binning sebagai tahap diskritisasi nilai kontinu dalam dataset, jenis serangan pada dataset ini dibagi kedalam 5 jenis. Berdasarkan pada hasil ujicoba yang dilakukan diperoleh hasil tingkat akurasi dengan menggunakan dataset ini adalah 60% dengan confusion matrik disajikan pada Tabel II.

TABEL II
CONFUSION MATRIK UNTUK DATA TANPA DISKRITISASI

CM	1	2	3	4	5
1	2601	16	0	0	72
2	1818	37	0	1	28
3	2	0	0	0	0
4	5	0	0	0	0
5	66	0	0	0	392

Pada *Confusion matrix* yang dihasilkan terlihat bahwa kemampuan deteksi dari algoritma *naive bayes* dengan menggunakan dataset tanpa melewati proses dikritisasi hanya berkisar pada 60% sehingga kemampuan dari teknik ini masih sangat jauh dari harapan. Hal ini terjadi karena atribut dengan nilai kontinu akan memiliki probabilitas kemunculan yang sangat kecil dalam dataset sehingga data tersebut tidak dapat diklasifikasikan dengan baik.

B. *Min-Max* normalisasi dengan diskritisasi 3 Interval

Pada skenario kedua, data yang digunakan telah melewati proses normalisasi yang sama dengan menggunakan metode max-min, namun setelah proses tersebut data kemudian mengalami proses diskritisasi dengan menggunakan metode *mean*/standar deviasi kedalam 3 bagian dengan nilai (-1,0,1). Berdasarkan pada hasil uji coba diperoleh *confusion matrix* sebagai berikut:

TABEL III
CONFUSION MATRIK UNTUK DISKRITISASI 3 INTERVAL

CM	1	2	3	4	5
1	2384	159	0	0	146
2	142	1687	0	1	54
3	2	0	0	0	0
4	2	0	0	0	0
5	52	29	0	0	377

Berdasarkan pada confusion matrik pada Tabel III diperoleh tingkat akurasi deteksi dari metode ini dengan menggunakan dataset yang telah melewati proses *binning* dengan menggunakan *mean*/standar deviasi 3 interval adalah sebesar 88% meningkat dari hasil deteksi sebelumnya tanpa melewati proses *binning*.

C. *Min-Max* normalisasi dengan diskritisasi 5 inteval

Pada skenario ketiga dataset yang digunakan melewati proses yang sama dengan dataset pada skenario 1 dan skenario 2 namun pada skenario ini teknik diskritisasi yang dilakukan adalah mengubah nilai kontinu kedalam 5 buah interval dengan nilai (-2,-1,0,1,2). Kemudian berdasarkan pada hasil uji coba diperoleh *confusion matrix* sebagai berikut:

TABEL IV
CONFUSION MATRIK UNTUK DISKRITISASI 5 INTERVAL

CM	1	2	3	4	5
1	2385	164	0	0	141
2	149	1709	0	0	25
3	2	0	0	0	0
4	0	6	0	0	0
5	42	22	0	0	393

Berdasarkan pada confusion matrix yang dihasilkan pada skenario 3 disajikan pada Tabel 4, diperoleh tingkat akurasi deteksi serangan sebesar 89,6%. Tingkat akurasi ini sedikit lebih baik jika dibandingkan dengan skenario yang dilakukan sebelumnya.

D. Perbandingan nilai deteksi anomali setiap skenario

Pada bagian ini akan dibahas mengenai hasil perbandingan terhadap kemampuan deteksi anomali dari setiap skenario yang telah dilakukan. Deteksi anomali yang dimaksudkan adalah sistem yang disusun mampu menentukan apakah pola tersebut masuk kedalam pola normal atau serangan tanpa meperhitungan jenis serangan yang dideteksi sehingga kelas dari perbandingan ini akan terbagi kedalam 2 kategori yaitu normal dan anomali. Hasil dari perbandingan disajikan dalam Tabel V berikut.

TABEL V
CONFUSION MATRIK PERBANDINGAN HASIL PERCOBAAN

CM	Akurasi	*Detection rate*	*False positif rate*	*Running time*
Tanpa binning	59,6 %	26,6%	1744	31,17 s
Equal width binning [3]	88,20%	94,7%	138	-
3-Interval	89,10 %	93,69%	148	31,5 s
5-Interval	88,36 %	92,93%	166	36,24 s

Pada Tabel V terlihat bahwa nilai akurasi dengan menggunakan skema diskritisasi 3-Interval mampu memberikan akurasi deteksi terbaik sebesar 89,10 % sedangkan untuk tanpa proses diskritisasi algoritma *naive bayes* hanya mampu memberikan akurasi sebesar 59,6%. Sedangkan untuk *detection rate* tertinggi diperoleh pada skenario dengan menggunakan skema *equal width binning* yang mencapai 94,7% sedangkan 2 teknik lainnya memiliki selisih kurang dari 1% yang dapat dipengaruhi oleh proses pemecahan data menjadi *training* dan *testing* yang dilakukan secara acak,sedangkan untuk tanpa binning *naive bayes* hanya mampu memperoleh nilai sebesar 26,6%. Parameter *running time* berdasarkan hasil yang diperoleh tidak begitu memberikan hasil yang jelas dikarenakan proses *running time* hanya menghitung waktu yang dibutuhkan untuk melakukan *learning* dan *testing* dengan menggunakan algoritma *naive bayes*, tetapi tidak mengikutsertakan waktu yang dibutuhkan untuk melakukan proses *pre-processing* termasuk normalisasi, *feature selection*, diskritisasi.

IV. KESIMPULAN

Metode yang diusulkan pada penelitian ini memberikan hasil yang tidak terlalu berbeda dengan penggunaan skenario *equal width binning* namun memiliki kemampuan deteksi yang meningkat secara signifikan jika dibandingkan dengan proses klasifikasi tanpa menggunakan proses *binning*. Proses *binning* (diskritisasi) menjadikan probabilitas dari algoritma naive bayes yang digunakan dapat lebih dihandalakan untuk menentukan kelas dari suatu data, namun proses diskritisasi ini juga menghilang beberapa informasi penting yang ada dalam *dataset* karena teknik ini tidak mempertimbangkan kelas dari suatu data sebelum melewati proses diskritisasi.

DAFTAR PUSTAKA

[1] K. Scarfone and P. Mell, Guide to Intrusion Detection and Prevention Systems (IDPS), Computer Security Resource Center (National Institute of Standards and Technology) 800–94, 2007.

[2] H. A. R. P Amudha, "Performance Analysis of Data Mining Approaches in Intrusion Detection," *Network Security,* 2011.

[3] Datta H.Deshmukh , Tushar Ghorpade and Puja Padiya , "Improving Classification Using Preprocessing and Machine Learning Algorithms on NSL-KDD Dataset," in *International Conference on Communication, Information & Computing Technology (ICCICT)*, Mumbai, 2015.

[4] Santosh Kumar Sahu, Sauravranjan Sarangi and Sanjaya Kumar Jena, "A Detail Analysis on Intrusion Detection Datasets," in *International Advance Computing Conference (IACC)*, 2014.

[5] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in *Symposium on Computational Intelligence*, 2009.

[6] I. S. C. o. eXcellence.. [Online]. Available: http://nsl.cs.unb.ca/NSL-KDD/.

[7] D. A. M. S. Revathi, "Network Intrusion Detection Using Hybrid Simplified Swarm Optimization and Random Forest Algorithm on NSL-KDD Dataset," *IJECS Volume 3,* pp. 3873-3876, 2 Feb, 2014.

[8] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach (2nd ed.), 2nd Edition ed., Prentice Hall..

[9] N. S.-M. A. A.-B. V. Bolón-Canedo, "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset," *Expert Systems with Applications ,* no. 38, p. 5947–5957, 2011.

[10] S.-W. K. b. C.-F. T. Wei-Chao Lin a, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," *Knowledge-Based Systems,* vol. 78, pp. 13-21, 23 April 2015.