

Pendekatan Machine Learning Dengan Menggunakan Algoritma Xgboost (*Extreme Gradient Boosting*) Untuk Peningkatan Kinerja Klasifikasi Serangan Syn
Rahmad Gunawan¹, Erik Suanda Handika¹, Edi Ismanto²

 Email: ¹goengoen78@umri.ac.id, ¹180401174@umri.ac.id, ²edi.ismanto@umri.ac.id

¹Teknik Informatika, Ilmu Komputer, Universitas Muhammadiyah Riau

²Pendidikan Informatika, FKIP, Universitas Muhammadiyah Riau

Diterima: 05 April 2020 | Direvisi: 05 Mei 2020 | Disetujui: 27 Mei 2020

 ©2020 Program Studi Teknik Informatika Fakultas Ilmu Komputer,
 Universitas Muhammadiyah Riau, Indonesia

Abstrak

Denial of Service (DoS) adalah salah satu serangan cyber populer yang ditargetkan pada situs web organisasi terkenal dan berpotensi memiliki biaya ekonomi dan waktu yang tinggi. Dalam makalah ini, beberapa metode pembelajaran mesin termasuk model ensemble dan pengklasifikasi deep learning berbasis autoencoder dibandingkan dan disetel menggunakan optimasi Bayesian. Kerangka autoencoder memungkinkan untuk mengekstrak fitur baru dengan memetakan input asli ke ruang baru. Metode tersebut dilatih dan diuji baik untuk klasifikasi biner dan multi-kelas pada kumpulan data Digiturk dan Labris, yang baru-baru ini diperkenalkan untuk mendeteksi berbagai jenis serangan DDoS. Semakin penting koneksi data melalui Internet membuat kebutuhan akan keamanan jaringan data semakin meningkat. Salah satu tools yang penting adalah Intrusion detection systems (IDS). Sistem Deteksi Intrusi (IDS) adalah proses pemantauan lalu lintas jaringan dalam sistem untuk mendeteksi pola dan aktivitas yang mencurigakan yang memungkinkan ada serangan dalam sistem itu. beberapa jenis serangan, yaitu Botnet, UDP, SYN, broadcast, sleep deprivation, dan serangan bertubi-tubi. klasifikasi pertama, hasilnya menunjukkan bahwa baik Precision (PR) dan Recall (RE) adalah 89% untuk Algoritma Random Forest. Akurasi rata-rata (AC) dari model yang kami usulkan adalah 89% yang luar biasa dan cukup baik. Pada klasifikasi kedua, hasilnya menunjukkan bahwa baik Precision (PR) dan Recall (RE) sekitar 90% untuk algoritma XGBoost. Akurasi rata-rata (AC) dari model yang kami sarankan adalah 90% pada dataset CICDDoS2019.

Kata kunci: *Machine learning, XGBoost, Serangan SYN, dataset CICDDoS2019*
Machine Learning Approach Using Xgboost (Extreme Gradient Boosting) Algorithm For Increasing Performance Syn Attack Classification
Abstract

Denial of Service (DoS) is one of the popular cyber-attacks targeted at websites of well-known organizations and has the potential to have high economic and time costs. In this paper, several machine learning methods including ensemble models and autoencoder-based deep learning classifiers are compared and tuned using Bayesian optimization. The autoencoder framework makes it possible to extract new features by mapping the original input to a new space. The method was trained and tested for both binary and multi-class classification on the recently introduced Digiturk and Labris datasets to detect various types of DDoS attacks. The more important the data connection via the Internet makes the need for data network security increases. One of the important tools is Intrusion detection systems (IDS). An intrusion Detection System (IDS) is the process of monitoring network traffic in a system to detect suspicious patterns and activities that may allow an attack in that system. There are several types of attacks, namely Botnet, UDP, SYN, broadcast, sleep deprivation, and barrage attacks. the first classification, the results show that both Precision (PR) and Recall (RE) are 89% for the Random Forest Algorithm. The average accuracy (AC) of our proposed model is an excellent 89% quite good. In the second classification, the results show that both Precision (PR) and Recall (RE) are around 90% for the XGBoost algorithm. The average accuracy (AC) of our recommended model is 90% on the CICDDoS2019 dataset.

Keywords: *Machine learning, XGBoost, SYN Attack, CICDDoS2019 dataset*

1. PENDAHULUAN

Pesatnya perkembangan internet telah membawa kemudahan dan evolusi bagi masyarakat. saat ini dikonsep untuk memanfaatkan kreativitas para ahli manusia dalam kolaborasi dengan mesin yang efisien, cerdas, dan akurat. Sayangnya, perkembangan pesat teknologi intrusi jaringan telah menyebabkan ketidakamanan dan merusak keandalan layanan Internet. Akibatnya, serangan keamanan jaringan[1]. Mendeteksi perilaku jahat penting untuk mencegah ancaman keamanan jaringan komputer. Denial of Service (DoS) adalah salah satu serangan cyber populer yang ditargetkan pada situs web organisasi terkenal dan berpotensi memiliki biaya ekonomi dan waktu yang tinggi. Dalam makalah ini, beberapa metode pembelajaran mesin termasuk model ensemble dan pengklasifikasi deep learning berbasis autoencoder dibandingkan dan disetel menggunakan optimasi Bayesian. Kerangka autoencoder memungkinkan untuk mengekstrak fitur baru dengan memetakan input asli ke ruang baru. Metode tersebut dilatih dan diuji baik untuk klasifikasi biner dan multi-kelas pada kumpulan data Digiturk dan Labris, yang baru-baru ini diperkenalkan untuk mendeteksi berbagai jenis serangan DDoS. Metode pembentukan terbaik ditemukan dalam bentuk ansambel meskipun pengklasifikasi pembelajaran mendalam mencapai tingkat akurasi yang sebanding [2].

Internet membuat kebutuhan akan keamanan jaringan data semakin meningkat. Salah satu tools yang penting adalah Intrusion detection systems (IDS). Sistem Deteksi Intrusi (IDS) adalah proses pemantauan lalu lintas jaringan dalam sistem untuk mendeteksi pola dan aktivitas yang mencurigakan yang memungkinkan adanya serangan dalam sistem itu. teknik pembelajaran mendalam (Khoei et al., 2021), untuk mendeteksi serangan dalam jaringan. Kedua studi tersebut terutama berfokus pada pendeteksian beberapa jenis serangan, termasuk Botnet, UDP, SYN, broadcast, sleep deprivation, dan serangan bertubi-tubi. Salah satu jenis serangan yaitu, SYN Serangan SYN adalah salah satu serangan DDoS yang paling dieksploitasi terkait dengan kerentanan dalam fase jabat tangan tiga arah TCP dari protokol TCP yang digunakan oleh hampir semua komunikasi jaringan termasuk Internet. Dalam serangan ini, penyerang mengirimkan sejumlah besar paket SYN berulang kali hingga mesin target menjadi tidak responsif terhadap pengguna yang sah[3].

Pada penelitian ini, banyak peneliti menggunakan Machine Learning dalam mendeteksi serangan pada IDS. Pada penelitian (Ismail et al., 2022) klasifikasi pertama, hasilnya menunjukkan bahwa baik Precision (PR) dan Recall (RE) adalah 89% untuk Algoritma Random Forest. Akurasi rata-rata (AC) dari model yang kami usulkan adalah 89% yang luar biasa dan cukup baik. Pada klasifikasi kedua, hasilnya menunjukkan bahwa baik Precision (PR) dan Recall (RE) sekitar 90% untuk algoritma XGBoost. Akurasi rata-rata (AC) dari model yang kami sarankan adalah 90%. Dengan membandingkan pekerjaan kami dengan penelitian yang ada, akurasi penentuan cacat adalah meningkat secara signifikan yaitu masing-masing sekitar 85% dan 79%. Ada juga peneliti sedang menjelaskan identifikasi berbagai jenis serangan DDoS, pada penelitian (Yungaicela-Naula, Vargas-Rosales and Perez-Diaz, 2021) model LSTM dan GRU memberikan kinerja terbaik dalam mendeteksi serangan SYN dan UDP, dengan tingkat deteksi rata-rata di atas 90% Khususnya, model GRU mencapai tingkat deteksi rata-rata. Beberapa metode dikembangkan untuk melakukan pendeteksian pada serangan SYN Flood . Penelitian ini dapat meningkatkan keamanan server dilihat dari capturing traffic yang mengarah kepada server dan monitoring terhadap proses resource CPU saat terjadi serangan mampu menurunkan usage cpu 46% hingga turun menjadi 4,6 %. Dengan mempunyai firewall packet filtering menjaga penggunaan sumber daya CPU pada server tetap rendah dan memfilter koneksi yang tidak sah [4].

Untuk penelitian ini penelitian memilih dataset CICDDoS2019 digunakan untuk tujuan merancang model untuk memantau perangkat jaringan dan menghasilkan peringatan untuk berbagai insiden atau serangan. Dataset CICIDS memiliki 14 jenis serangan jinak dan terbaru yang menggambarkan data dunia nyata. Kumpulan data memiliki volume yang sangat besar, sekitar 11,5 GB data yang berisi lebih dari 2,2 Juta instans dari 83 fitur aliran jaringan yang berbeda [3]. Namun Menurut [5]. Basis data CICDDoS 2019 terdiri dari 88 fitur yang diekstraksi menggunakan CICFlowMeter. Dalam pekerjaan ini, kami mengurangi, Model pelatihan dengan nilai yang hilang dapat mempengaruhi evaluasi model ini. Beberapa teknik telah diusulkan untuk memecahkan masalah ini dalam beberapa tahun terakhir, termasuk imputasi dek panas dan dingin, imputasi rata-rata, imputasi ekstrapolasi dan interpolasi, dan imputasi berbasis regresi. Dalam penelitian ini, kami menerapkan imputasi rata-rata, di mana nilai yang hilang untuk suatu fitur diganti dengan rata-rata nilai fitur tersebut untuk semua fitur. Data Preprocessing Pada tugas akhir ini, data preprocessing terdiri dari beberapa langkah, yaitu class rebalancing dan sample size reduction, menghilangkan fitur, missing value imputasi, normalisasi data input, dan encoding data berlabel. Berikut ini, kita akan membahas langkah-langkah tersebut secara lebih rinci. dimensi fitur menjadi 37 fitur yang relevan. Fitur yang dihapus termasuk Flow ID, Source IP, Destination IP, dan Timestamp

Berdasarkan penjabaran diatas, peneliti menerapkan algoritma XGBoost dengan menggunakan dataset CICIDS2019. Algoritma XGBoost adalah mesin yang dapat diskalakan sistem pembelajarannya untuk meningkatkan pohon keputusan. Adapun keunggulan dari model pembelajaran XGBoost memiliki kecepatan, skalabilitas, efisiensi, dan kesederhanaan yang sangat cepat.

2. METODE PENELITIAN

Pada tahapan ini peneliti membuat alur diagram untuk mempersentasikan proses secara sistimatis dan logis jalannya penelitian yang dilakukan.

2.1 Identifikasi dataset

Dataset yang digunakan adalah pubic dataset CICDDoS2019 [6] yang diambil dari The Canadian Institute for Cybersecurity .html,

dataset ini berisi serangan DDoS dan yang jinak, yang menyerupai serangan real di dunia nyata (PCAP). Dan juga mencakup hasil analisis lalu lintas jaringan menggunakan CICFlowMeter-V3 dengan arus berlabel berdasarkan stempel waktu, IP sumber dan tujuan, port sumber dan tujuan, protokol dan serangan (file CSV). Data yang digunakan dalam penelitian ini yaitu serangan DDoS yang berbasis exploitasi dengan menganalisis paket serangan pada SYN dengan extension file CSV.

No	FeatureName	Description	No	Feature Name	Description
1	Unnamed: 0		45	Bwd Packets/s	Number of backward packets per second
2	Flow ID	Flow Identity	46	Min Packet Length	Minimum length of a flow
3	Source IP	Source IP Address	47	Max Packet Length	The maximum length of a flow
4	Source Port	Source Port	48	Packet Length Mean	Mean length of a flow
5	Destination IP	Destination IP Address	49	Packet Length Std	Standard deviation length of a flow
6	Destination Port	Destination Port	50	Packet Length Variance	Minimum inter-arrival time of packet
7	Protokol	Protokol	51	FIN Flag Count	Number of packets with FIN
8	Timestamp	Date format that is distributed on Unix-based servers	52	SYN Flag Count	Number of packets with SYN
9	Flow Duration	Flow Duration	53	RST Flag Count	Number of packets with RST
10	Total Fwd Packets	Total Forward Packets	54	PSH Flag Count	Number of packets with PSH
11	Total Backward Packets	Total Backward Packets	55	ACK Flag Count	Number of packets with ACK
12	Total Length of Fwd Packets	Total Length Forward Packets	56	URG Flag Count	Number of packets with URG
13	Total Length of Bwd Packets	Total Length Backward Packets	57	CWE Flag Count	Number of packets with CWE
14	Fwd Packet Length Max	Forward Packet Length Max	58	ECE Flag Count	Number of packets with ECE
15	Fwd Packet Length Min	Forward Packet Length Min	59	Down/Up Ratio	Download and upload ratio
16	Fwd Packet Length Mean	Forward Packet Length Mean	60	Average Packet Size	The average size of the packet
17	Fwd Packet Length Std	Forward Packet Length Standard	61	Avg Fwd Segment Size	Average size observed in the forward direction
18	Bwd Packet Length Max	Backward Packet Length Max	62	Avg Bwd Segment Size	Average size observed in the backward direction
19	Bwd Packet Length Min	Backward Packet Length Min	63	Fwd Header Length.1	Forward Header Length

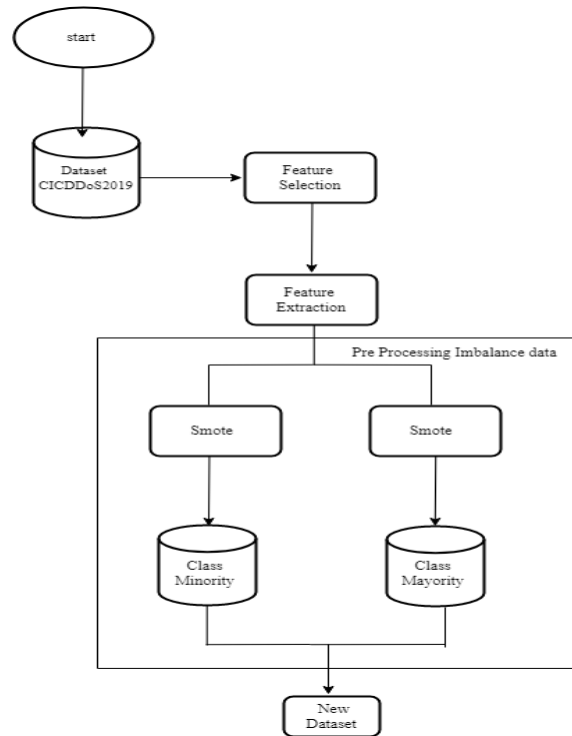
No	FeatureName	Description	No	Feature Name	Description
20	Bwd Packet Length Mean	Backward Packet Length Mean	64	Fwd Avg Bytes/Bulk	The average number of bytes bulk rate in the forward direction
21	Bwd Packet Length Std	Backward Packet Length Standard	65	Fwd Avg Packets/Bulk	The average number of packets bulk rate in the forward direction
22	Flow Bytes/s	flow byte rate that is number of packets transferred per second	66	Fwd AvgBulk Rate	Average number of bulk rate in the forward direction
23	Flow Packets/s	Flow packets rate that is number of packetstransferred persecond	67	Bwd Avg Bytes/Bulk	Average number of bytes bulk rate in the backward direction
24	Flow IATMean	Mean time between two flows	68	Bwd Avg Packets/Bulk	Average number of packets bulk rate in the backward direction
25	Flow IATStd	Standard time between two flows	69	Bwd AvgBulk Rate	Average number of bulk rate in the backward direction
26	Flow IATMax	Max time between two flows	70	Subflow Fwd Packets	The average number of packets in a sub flow in the forward direction
27	Flow IATMin	Min time between two flows	71	Subflow Fwd Bytes	The average number of bytes in a sub flow in the forward direction
28	Fwd IATTotal	Total time between two packets sent in the forward direction	72	Subflow Bwd Packets	The average number of packets in a sub flow in the backward direction
29	Fwd IATMean	Mean time between two packets sent in the forward direction	73	Subflow Bwd Bytes	The average number of bytes in a sub flow in the backward direction
30	Fwd IATStd	Standard deviation time between two packets sent in the Forward direction	74	Init Win bytes forward	Number of bytes sent in initial window in the forward direction
31	Fwd IAT Max	Maximum time between two packets sent in the forward direction	75	Init Win bytes backward	Number of bytes sent in initial window in the backward direction
32	Fwd IAT Min	Minimum time between two packets sent in the forward direction	76	Act data pkt fwd	Number of packets with at least 1 byte of TCP data payload in the forward direction
33	Bwd IAT Total	Maximum time between two packets sent in the forward direction	77	Min seg size forward	Minimum segmentsize observed in the forward direction

No	FeatureName	Description	No	Feature Name	Description
34	Bwd IAT Mean	Mean time Between two packets sent in The backward direction	78	Active Mean	Mean time a flow was active before becoming idle
35	Bwd IATStd	Standard deviation time between two packets sent in the backward direction	79	Active Std	Standard deviation time a flow was active before becoming idle
36	Bwd IAT Max	Maximum time between two packets sent in the backward direction	80	Active Max	Maximum time a flow was active before becoming idle
37	Bwd IAT Min	Minimum time between two packets sent in the backward direction	81	Active Min	Minimum time a flow was active before becoming idle
38	Fwd PSH Flags	Number of times the PSH flag was set in packets travelling in the forward direction (0 for UDP)	82	Idle Mean	Mean time a flow was idle before becoming active
39	Bwd PSH Flags	Number of times the PSH flag was set in packets travelling in the backward direction (0 for UDP)	83	Idle Std	Standard deviation time a flow was idle before becoming active
40	Fwd URG Flags	Number of times the URG flag was set in packets travelling in the forward direction (0 for UDP)	84	Idle Max	Maximum time a flow was idle before becoming active
41	Bwd URG Flags	Number of times the URG flag was set in packets travelling in the backward direction (0 for UDP)	85	Idle Min	Minimum time a flow was idle before becoming active
42	Fwd Header Length	Total bytes used for headers in the forward Direction	86	SimillarHTTP	HTTP Simillarity
43	Bwd Header Length	Total bytesUsed for headers in the Backward direction	87	Inbound	Inbound Traffic

No	FeatureName	Description	No	Feature Name	Description
44	Fwd Packets/s	Number of forward packets persecond	88	Label	Label Attack

Gambar 1. Fitur-fitur pada dataset

2.2. Menerapkan Metode Smote untuk Mengatasi Masalah Imbalance Class dan Metode Imputasi untuk Mengatasi Masalah Missing Value.



Gambar 2. Skema mengatasi Imbalaced Data

2.3. Data Preprocessing dan Data Cleaning

Adapun tahapan preprocessing sebagai berikut:

1. Data Cleaning

Proses menghilangkan noise dari data yang tidak konsisten atau tidak relevan. Pembersihan data ini akan mempengaruhi performansi teknik/ metode data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

2. Feature Selection

Feature selection adalah adalah suatu proses subset *feature* yang paling relevan dari *feature* asli untuk digunakan dalam konstruksi model dari dataset. Sehingga waktu yang digunakan mengeksekusi dari pengklasifikasi yang memproses data berkurang, dan dapat meningkatkan akurasi juga karena features yang tidak relevan dapat memperburuk data mempengaruhi akurasi klasifikasi secara negatif. Dengan *feature selection* dapat meningkatkan pemahaman dan biaya penanganan data menjadilebih kecil [7].

3. Feature extraction

Setelah proses preprocessing, perlu ekstraksi fitur untuk mendapatkan hasil klasifikasi yang lebih maksimal. Ekstraksi fitur berfungsi untuk mengurangi noise dengan menghapus feature yang tidak relevan, sehingga dapat meningkatkan akurasi

4. Perhitungan Smote

```
def SMOTE (k=2, m=50%, r=2): # defaults
    dimana mayoritas > m do
        hapus setiap item mayoritas # random
    sementara Minoriti < m do
        tambahkan sesuatu seperti(setiap item minoritas)
    def sesuatu seperti (X0):
        relevan = set kosong
        k1 = 0
        dimana (k1++ < 20 dan ukuran (ditemukan) < k) {
            semua = k1 tetangga terdekat
            relevan += item di “semua” kelas XO}
        Z = ada yang di temukan
        Y = interpolasi (X0, Z)
        kembalikan Y
    def minkowski_jarak(a,b,r):
        kembalikan ( $\sum_i \text{abs}(a_i - b_i)^r$ )1/r
```

5. Perhitungan Imputasi

```
Require: input dataset DS having NA values
Ensure: output dataset DS_WONA without NA values
1: COLS_WONA ← Select columns in DS not having NA values
2: COLS_WNA ← Select columns in DS having NA values
3: DS_WONA ← DS (COLS_WONA)
4: for i=0 to len(COLS_WNA) do
5: TRAIN ← Select all nonNA rows of COLS_WNA[i] from DS
6: TEST ← Select all rows from DS having NA value in COLS_WNA[i]
7: X_Train ← TRAIN(COLS_WONA)
8: Y_Train ← TRAIN(COLS_WNA[i])
9: model.fit (X_Train, Y_Train)
10: X_Test ← TEST(COLS_WONA)
11: Y_Test ← model.predict (X_Test)
12: DS_WONA[COLS_WNA[i]] ← Combine (Y_Train, Y_Test)
13: end for
```

Gambar 3. Psaucode Imputasi Dataset

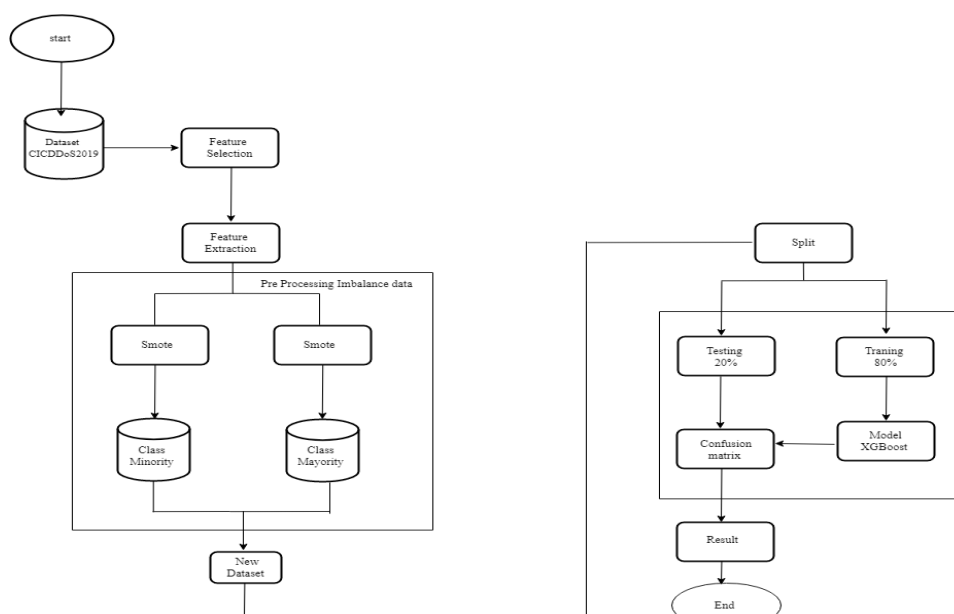
6. New Dataset

New Dataset adalah dataset baru yang di hasil kan dari proses fitur *selection* dan fitur *extraction*.

7. Hasil

Pada tahap proses penyelesaian masalah *imbalance* data menghasilkan sebuah dataset baru yang siap di terapkan kedalam model selanjutnya.

2.4 Menerapkan Teknik Machine Learning Dengan Metode Xgboost Untuk Performance Akurasi Yang Lebih Baik Dalam



Gambar 4. Skema Penerapan Metode Algoritma XGboost

1. Proprocessing

Adapun tahapan preprocessing sebagai berikut :

a. New Dataset

Setelah dilakukan proses penyelesaian permasalahan data yang tidak seimbang yang terdapat pada dataset. Menghasilkan dataset yang baru. Dataset ini merupakan dataset yang sudah mempunyai data yang seimbang dan sudah siap digunakan pada proses penerapan model yang akan di gunakan.

b. Feature Splitting

Setelah di lakukan proses penyeimbangan data maka di peroleh dataset yang baru. Yang mana dataset yang baru ini nanti nya akan di split atau di bagi menjadi dua yaitu: data training (80%) dan data testing (20%) yang nantinya di gunakan dalam pengklasifikasian

c. Model XGBoost

Data dibagi menjadi dua, pada bagian data traning (80%) selanjutnya dilakukan pengujian menggunakan Model XGBoost untuk mengetahui tingkat akurasi pada data.

d. Confusion Matrix

Data pada bagian testing (20%) di terapkan ke confusion matrix. Yang mana fungsi confusion matrix ini sendiri untuk mengukur performa dalam permasalahan klasifikasi biner.

Confusion matrix adalah salah satu teknik yang dapat digunakan dalam mengukur performa untuk menemukan kebenaran dan keakuratan suatu model. Dengan memberikan informasi membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang sebenarnya. Berikut ini adalah istilah-istilah yang berkaitan dengan *Confusion Matrix* dapat dilihat pada tabel 1 berikut ini (Li et al., 2021) :

Actual	Predicted	
	<i>Attack</i>	Normal
<i>Attack</i>	TP	FN
Normal	FP	TN

Tabel 1. Tabel confusion matrix

- 1) *True Positive* (TP): adalah kondisi dimana kasus yang dikerjakan, datanya benar (data positif) dan diprediksi dengan benar.
- 2) *False Negative* (FN): adalah suatu kondisi dimana kelas data sebenarnya adalah benar (positif) dan diprediksi salah atau dengan kata lain model diprediksi sebagai data negatif sedangkan datanya positif.
- 3) *False Positive* (FP): adalah kondisi di mana kelas sebenarnya dari titik data salah (adalah data negatif) dan diprediksi sebagai data positif (benar).
- 4) *True Negative* (TN): adalah kondisi data yang sebenarnya salah (data negatif) dan benar diprediksi sebagai data negatif (salah).

Dalam proses evaluasi performa dan hasil klasifikasi yang diperoleh, digunakan model metrik kinerja yang dijelaskan sebagai berikut [1]:

1. *Accuracy*: mengukur jumlah klasifikasi yang benar

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

2. *Precision*: mengukur jumlah klasifikasi yang benar yang dihukum oleh jumlah klasifikasi yang salah.

$$\frac{TP}{TP + FP} \quad (2.2)$$

3. *Recall*: mengukur jumlah klasifikasi yang benar yang dihukum oleh jumlah entri yang tidak terjawab.

$$\frac{TP}{TP + FN} \quad (2.3)$$

4. *F-score*: mengukur rata-rata harmonik presisi dan daya ingat, yang berfungsi sebagai *derived*. pengukuran efektivitas.

$$\frac{FP}{FP + TN} \quad (2.4)$$

e. Hasil

Setelah di lakukannya penerapan dengan menggunakan model XGBoost dan untuk melakukan pengklasifikasian serangan SYN dengan menggunakan confusion matrik. Yang mana di lakukannya metode pengujian yang digunakan untuk menghitung tingkat akurasi dengan membandingkan dengan hasil pengklasifikasian dari model yang di gunakan.

3. HASIL DAN PEMBAHASAN

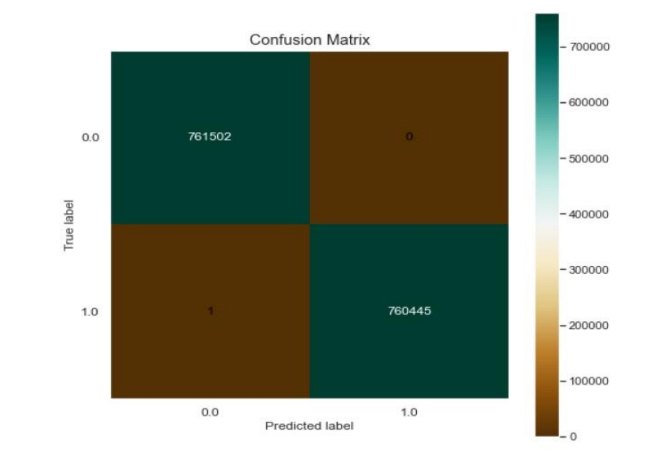
Pada tahapan ini merupakan sebuah tahapan mencari referensi penelitian terkait study kasus klasifikasi serangan padadataset CICDDoS2019 yang mana menggunakan model XGBoost. Sumber referensi yang digunakan yaitu dari beberapa jurnal dari penelitian terdahulu. Dari referensi tersebut penulis mempelajari teori teori dasar, dan juga mengambil kutipan yang nilai memiliki kaitan dengan penelitian yang dilakukan.

3.1. Identifikasi Persiapan Dataset

Penulis Menggunakan dataset CICDDoS2019. Datset memiliki jumlah kolom 88 fitur. Kolom dataset yang digunakan pada penelitian ini yaitu *category* dan *attack* yang berisi huruf dan angka. Pada kolom *attack* berisi angka 1 untuk *attack* dan 0 untuk *normal*. Jumlah data keseluruhan pada dataset terdiri dari 4320541 *attack* dan *normal*. Kemudian dalam analisis data ini dilakukan eksplorasi data yang menguraikan dan menjelaskan secara visual,serta menjelaskan informasi dari banyak data. Proses ini ditujukan agar dapat memahami struktur data yang akan diolah. Berikut adalah langkah-langkah dalam analisis data pada CICDDoS2019. Berikut hasil tampilan dari umlah dataset CICDoS2019.

3.2 Perhitungan dengan confusion matrix

Tahapan ini mengevaluasi hasil performa pada tiap algortima yang digunakan pada penelitian ini dengan melihat nilai prediksi pada algortima menggunakan confusion matrix Gambar 6 dan 7, melihat hasil score accuracy, precision, recall, dan f-score dari hasil confusion matrix kemudian mengevaluasi fitur yang berkontribusi pada tiap model prediksi dengan feature important.



Gambar 6. Grafik confusion matrix

```

Confusion matrix :
[[760445    1]
 [    0 761502]]
Outcome values :
760445 1 0 761502
Classification :

```

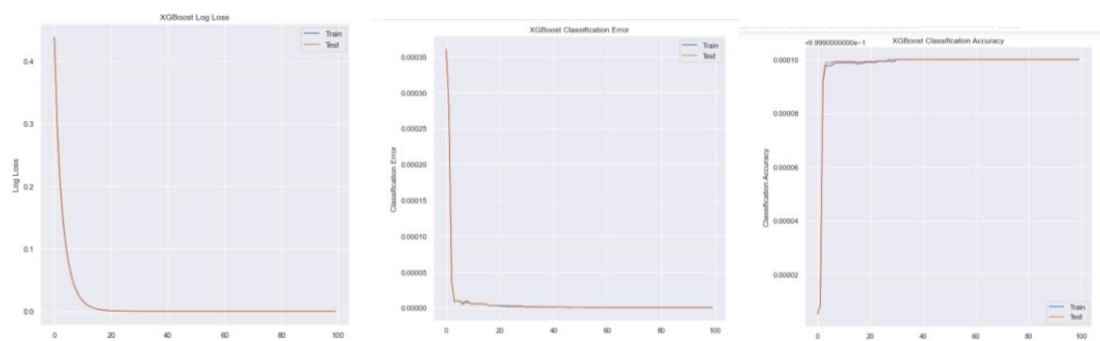
	precision	recall	f1-score	support
1	1.00	1.00	1.00	760446
0	1.00	1.00	1.00	761502
accuracy			1.00	1521948
macro avg	1.00	1.00	1.00	1521948
weighted avg	1.00	1.00	1.00	1521948

Gambar 7. Hasil perhitungan confusion matrix

Dari hasil di atas peneliti menggunakan algoritma XGBoost mendapatkan hasil akurasi precision 100%, recall 100%, f1-score 100% pada perhitungan confusion matrix.

3.3 Pembuktian Hasil dengan menggunakan algoritma XGBoost

Pada tahapan pembuktian ini dilakukan pengecekan dengan menggunakan grafik XGBoost *log loss* dan grafik klasifikasi *error* berguna untuk melihat tingkat *loss* dan *error* pada pengujian *performance* akurasi



Gambar 8. Grafik hasil dari XGBoost log loss dan Grafik Klasifikasi error

4. KESIMPULAN

Pada kumpulan dataset CICDDoS2019, hasil dari penelitian yang telah dilakukan serta menggunakan dataset sebanyak

4320541 record data dapat disimpulkan bahwa Hasil klasifikasi menggunakan algoritma XGBoost menghasilkan tingkat akurasi sebesar 99%. Dari dataset CICDDOS2019 pada serangan SYN yang menggunakan model XGBoost dengan pembagian 80% untuk data training dan 20% untuk data testing. Selanjutnya Mengatasi imbalance class menggunakan teknik SMOTE dan mengatasi missing value menggunakan teknik Imputasi dan mendapatkan jumlah kelas yang seimbang. Setelah dilakukan pengujian menggunakan confusion matrix dengan hasil yang didapat untuk recall, presisi dan f1-score adalah 100%, 100%, 100%.

DAFTAR PUSTAKA

- [1] Y. Li, W. Xu, W. Li, A. Li, and Z. Liu, "Research on hybrid intrusion detection method based on the ADASYN and ID3 algorithms," *Math. Biosci. Eng.*, vol. 19, no. 2, pp. 2030–2042, 2021, doi: 10.3934/MBE.2022095.
- [2] Y. Gormez, Z. Aydin, R. Karademir, and V. C. Gungor, "A deep learning approach with Bayesian optimization and ensemble classifiers for detecting denial of service attacks," *Int. J. Commun. Syst.*, vol. 33, no. 11, 2020, doi: 10.1002/dac.4401.
- [3] M. D. F. N. T. using M. L. Nawaz, M. A. Paracha, A. Majid, and H. Durad, "Attack Detection From Network Traffic using Machine Learning," *VFAST Trans. Softw. Eng.*, vol. 8, no. 1, pp. 1–7, 2020.
- [4] S. Sahren, "Implementasi Teknologi Firewall Sebagai Keamanan Server Dari Syn Flood Attack," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 7, no. 2, pp. 159–164, 2021, doi: 10.33330/jurteksi.v7i2.933.
- [5] C. Hu, "Ensemble Feature Learning-Based Event Classification for Cyber-Physical Security of the Smart Grid," no. September, 2019, [Online]. Available: <https://spectrum.library.concordia.ca/985779/>
- [6] Canadian Institute for Cybersecurity, "DATASET CICDDOS2019," 2019, [Online]. Available: <https://www.unb.ca/cic/datasets/ddos-2019.html>
- [7] C. Ding, H. Han, Q. Li, X. Yang, and T. Liu, "IT3SE-PX: Identification of Bacterial Type III Secreted Effectors Using PSSM Profiles and XGBoost Feature Selection," *Comput. Math. Methods Med.*, vol. 2021, 2021, doi: 10.1155/2021/6690299.