

A GAN-Based Approach for Identifying Fake Accounts on Twitter

**M Syafrizal Zain^{*}, Better Swengky, Bradika Almandin Wisesa,
Vivin Mahat Putri, Reddy Ibrachim**

Politeknik Manufaktur Negeri Bangka Belitung

Penulis Korespondensi: msyafrizalz@polman-babel.ac.id,
better@polman-babel.ac.id, bradika@polman-babel.ac.id, vivin@polman-babel.ac.id,
reddyibrachim@student.polman-babel.ac.id

Article Information:

Received: December 18, 2024, Accepted: January 10, 2024, Published: January 25, 2025

Abstract

The multiple security threats on the network make the need for robust security measures a major concern. The increasing presence of fake accounts and malicious actors on online platforms poses significant challenges, requiring sophisticated detection techniques to maintain network integrity. To address these issues, we propose a novel method for detecting fake accounts by leveraging Generative Adversarial Networks (GANs). By analyzing data extracted from platform APIs, our approach leverages the unique characteristics of GANs to improve the accuracy and efficiency of the detection process. In this study, we develop a GANs-based model specifically designed to detect fake accounts. The model is built through several key stages: first, we collect a comprehensive dataset, then perform data processing and preprocessing to make it suitable for machine learning applications. Next, the model is trained using various hyperparameters to optimize accuracy, thus learning the underlying patterns associated with fake accounts. After the training stage, the model is tested on previously unseen data to evaluate its generalization and performance in real-world scenarios. Experimental results show that our model achieves a threshold value of 0.0054779826. This value plays a crucial role in determining the accuracy of the detection system. The smaller the threshold value, the higher the model accuracy, as it shows a lower error rate in distinguishing between real and fake accounts. The ability of GANs-based models to adaptively learn from data during the training process contributes to high precision in detecting anomalies as well as minimizing false positives.

Keyword: Deteksi akun palsu, Generative Adversarial Networks (GANs), Machine learning, Threshold value, Deteksi anomali

1. Pendahuluan

Jaringan Sosial Online (OSN) merupakan salah satu platform aplikasi paling populer di perangkat mobile. Saat ini, OSN telah menjadi elemen penting dalam kehidupan sehari-hari, berfungsi sebagai platform utama untuk menghubungkan orang-orang di seluruh dunia dalam berbagi berbagai jenis data, termasuk foto, video, dan pesan. Namun, keberadaan akun palsu menjadi ancaman serius terhadap keamanan OSN. Berbagai penelitian telah mengusulkan metode untuk meningkatkan perlindungan OSN dengan berbagai pendekatan [1]. Sebagai contoh, sebuah studi memperkenalkan model saluran virtual untuk meningkatkan keamanan OSN [2].

Secara umum, setiap perangkat memiliki mekanisme keamanan untuk membuka kunci dan mengaksesnya, seperti PIN, kata sandi, atau pola kunci. Sayangnya, karena tidak adanya keamanan tambahan untuk memeriksa perilaku dan aktivitas pengguna setelah perangkat atau aplikasi terbuka, model konvensional masih memiliki kelemahan dalam melindungi data pengguna. Kelemahan ini memungkinkan pengguna tidak sah membobol kata sandi atau PIN yang sederhana pada ponsel maupun perangkat wearable [3].

Seiring dengan pesatnya perkembangan teknologi internet, jumlah pengguna OSN terus meningkat secara signifikan, sementara data jaringan mengalami perubahan yang sangat cepat. Berbagai artikel telah membahas metode deteksi anonim untuk meningkatkan privasi di OSN. Namun, banyak pendekatan yang masih mengabaikan perbedaan kebutuhan perlindungan privasi antar pengguna [4].

Salah satu tantangan utama dalam OSN publik, seperti Twitter, adalah kemunculan bot digital yang berperan sebagai aktor manusia-robotik. Bot ini sebenarnya merupakan program komputer yang mampu melakukan berbagai tugas dengan cepat dan efisien, yang sulit dilakukan oleh manusia secara manual. Statistik menunjukkan bahwa pada kuartal pertama tahun 2018, Twitter memiliki rata-rata 336 juta pengguna aktif bulanan, dengan sekitar 5% di antaranya merupakan akun palsu atau spam. Selain itu, sekitar 9%-15% dari total tweet yang dipublikasikan berasal dari akun palsu [5].

Dalam beberapa tahun terakhir, platform Twitter mendapatkan kritik tajam, terutama terkait tekanan politik dan dugaan keterlibatannya dalam pemilu Amerika Serikat. Sebagai respons, Twitter mengambil tindakan dengan menghapus lebih dari 70 juta akun palsu pada tahun 2019 [6].

Besarnya jumlah pengguna OSN ini menyebabkan berbagai permasalahan terkait keamanan data dan privasi. Bot yang dirancang untuk aktivitas berbahaya telah berkembang menjadi ancaman serius di internet. Penyedia layanan OSN telah menerapkan berbagai metode untuk melawan bot jahat, seperti Twitter dan Facebook yang menyediakan fitur "laporkan sebagai spam" untuk membantu mengidentifikasi bot spam. Bot ini sering digunakan sebagai alat untuk menyebarkan informasi berbahaya, memanipulasi opini publik, menyesatkan pemahaman konsumen, serta melakukan aktivitas penipuan. Secara umum, bot dapat diklasifikasikan ke dalam beberapa kategori, yaitu Spam Bots, Social Bots, Like Bots, Influential Bots, dan Fake Follower Bots [6].

Di antara berbagai tweet yang ada di Twitter, sebagian besar berasal dari bot. Twitter bots memiliki dampak signifikan terhadap opini publik dalam berbagai topik dan dapat dimanfaatkan untuk menyebarkan ide atau mempromosikan bisnis karena Twitter memiliki

jangkauan luas dan biaya implementasi yang sangat rendah. Tujuan utama dari Twitter bots dapat dirangkum sebagai berikut:

1. Menyebarkan misinformasi dan berita palsu.
2. Mencemarkan nama baik seseorang.
3. Mencuri kredensial pengguna dan menyebarkan komunikasi berbahaya.
4. Menyesatkan pengguna ke situs palsu.
5. Mempengaruhi pola pikir individu atau kelompok, serta membentuk opini publik secara manipulatif [7].

Dalam penelitian ini, kami mengusulkan penggunaan GAN untuk mendeteksi intrusi pada jaringan dengan menganalisis dan mengekstraksi fitur dalam dataset yang berguna untuk mendeteksi intrusi serta mengklasifikasikannya ke dalam kategori normal dan anomali. Dalam mendeteksi intrusi jaringan menggunakan GAN, kami berkontribusi dalam beberapa aspek utama sebagai berikut:

1. Mengusulkan teknik baru untuk mendeteksi intrusi jaringan dengan melatih dataset dan membangun model yang andal. Kami memanfaatkan berbagai fitur data intrusi jaringan yang tersedia di internet untuk membangun model ini.
2. Mengembangkan model yang mampu membedakan antara jaringan yang normal dan abnormal. Model ini merupakan solusi untuk menggantikan metode konvensional yang masih digunakan dalam deteksi intrusi jaringan.
3. Melakukan pengujian terhadap model yang diusulkan untuk mencapai tingkat akurasi yang tinggi dalam mendeteksi intrusi secara cepat dan akurat berdasarkan fitur yang dianalisis. Kami mengoptimalkan parameter untuk memperoleh hasil akurasi terbaik.

Struktur Penelitian: Jurnal ini disusun sebagai berikut: Bagian II membahas lebih lanjut hasil penelitian sebelumnya. Bagian III menjelaskan permasalahan dalam penelitian ini. Bagian IV menguraikan desain eksperimen, termasuk algoritma pembelajaran fitur, dataset, dan proses pra-pemrosesan. Bagian V menyajikan hasil penelitian beserta analisis mendalamnya. Terakhir, bagian VI merangkum permasalahan yang belum terselesaikan dalam penelitian ini.

2. Tinjauan Literatur

Identifikasi akun palsu di jaringan sosial telah menarik perhatian yang signifikan dalam beberapa tahun terakhir, dan berbagai strategi telah diusulkan untuk mengatasi permasalahan ini. Penelitian sebelumnya mengelompokkan solusi dalam tiga pendekatan utama, yaitu Crowdsourcing, Graph-based, dan Feature-based detection. Dalam studi ini, kami berfokus pada deteksi berbasis fitur (feature-based detection). Baru-baru ini, algoritma machine learning digunakan untuk membedakan akun spam dari akun non-spam di Twitter. Fitur diekstraksi dan dipilih dari akun atau tweet, lalu diterapkan dengan teknik machine learning yang supervised, semi-supervised, dan unsupervised untuk mengidentifikasi atau mengklasifikasikan jenis akun palsu [6].

Berbagai penelitian telah mengusulkan pendekatan untuk menyelesaikan masalah Keamanan Jaringan dan Informasi. Salah satu penelitian memperkenalkan model DL-IDS menggunakan machine learning. Model ini mengekstraksi fitur spasial dan temporal dari data lalu lintas jaringan untuk meningkatkan deteksi intrusi. DL-IDS menggunakan metode optimasi bobot

kategorikal untuk meningkatkan ketahanan dan mengurangi ketidakseimbangan sampel dalam berbagai jenis serangan [7].

Metode steganografi juga digunakan untuk menyembunyikan informasi penting dalam gambar. Jika seseorang menyalin gambar tersebut dan mencoba membuat profil palsu menggunakan data curian, sistem akan secara otomatis mendeteksi penipuan tersebut. Selain itu, kombinasi entropi interval waktu dan kemiripan tweet digunakan untuk mendeteksi bot spam. Timestamp digunakan untuk mengukur entropi interval waktu setiap pengguna, sedangkan perbandingan uni-gram digunakan untuk mengukur kesamaan tweet [8].

Sistem pencocokan URL blacklist berbasis Support Vector Machine (SVM) juga telah dikembangkan. Sistem ini terdiri dari tiga komponen utama: pemetaan dan perakitan, pra-penyaringan, dan klasifikasi. Jika suatu URL cocok dengan daftar hitam, maka sistem akan menandainya sebagai spam. Jika tidak, SVM akan melakukan analisis lebih lanjut terhadap URL tersebut [10]. Deteksi keamanan siber real-time pada akun Twitter dilakukan berdasarkan tiga set fitur dan tiga metode machine learning, yaitu decision tree, random forest, dan SVM [11].

Model DeepScan membagi aktivitas pengguna ke dalam beberapa interval waktu berkelanjutan. DeepScan memanfaatkan teknologi deep learning dengan fitur time-series, yang lebih deskriptif dibandingkan fitur standar [11]. Sebuah metrik fleksibel juga diusulkan untuk mengukur perubahan aktivitas pengguna dan merancang fitur baru guna menganalisis pola evolusi pengguna. Teknik machine learning supervised dan unsupervised kemudian digunakan untuk membedakan akun spam [12].

Sistem deteksi spammer berbasis Bagging Extreme Learning Machine (ELM) juga telah dikembangkan. Beberapa ELM dikombinasikan dalam pendekatan bagging untuk meningkatkan akurasi deteksi spam di media sosial [13]. Teknik semi-supervised untuk deteksi spam di Twitter menggunakan ensemble-based framework yang terdiri dari empat klasifikator [14]. Kombinasi dari empat teknik machine learning, yaitu Bernoulli Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, dan Decision Tree, telah diterapkan. Mereka menggunakan voting classifier, sebuah metode ensemble learning, untuk mengukur akurasi kombinasi klasifikator yang berbeda [11]. Selain itu, metode logika fuzzy dan evaluasi multilayer perceptron berbasis neural network telah digunakan untuk mendeteksi spam. Logika fuzzy terbukti dapat menangani dataset besar secara efisien dan mampu mendeteksi spam dalam hitungan detik [13].

Analisis jaringan sosial semakin populer saat ini. Namun, hanya sedikit penelitian yang menganalisis deteksi akun palsu menggunakan pendekatan semantik dan rekayasa ontologi. Sebuah metode berbasis ontologi diusulkan untuk mengidentifikasi konten mencurigakan di Twitter, terutama dalam situasi di mana tweet terkait dengan ontologi dari tema tertentu. Metode ini membandingkan teks tweet dengan ontologi topik yang relevan untuk mendeteksi kemiripan dan memverifikasi keasliannya [12].

Algoritma baru untuk deteksi spam, yang disebut Social Event Detection, berbasis ontologi juga telah dikembangkan. Pendekatan ini terdiri dari beberapa langkah, dimulai dari pembuatan ontologi, ekstraksi atribut, korelasi kata dengan konteks kelas saat ini, hingga identifikasi apakah suatu konten merupakan spam atau tidak [10]. Dua jenis filter ontologi telah diterapkan: filter ontologi global dan filter ontologi yang disesuaikan dengan pengguna. Filter ontologi

pengguna dibuat berdasarkan latar belakang pengguna tertentu, sedangkan filter global digunakan untuk seluruh sistem [13].

Metode deteksi phishing berbasis ontologi juga diusulkan sebagai sistem semantik untuk identifikasi phishing. Pendekatan ini mengevaluasi karakteristik leksikal dan menghilangkan ambiguitas akibat variasi karakteristik superfisial. Solusi ini menambahkan semantik pada strategi bag-of-words dan part-of-speech, yang terbukti sangat akurat [12]. Sebuah pendekatan berbasis logika definisi ontologi digunakan untuk menggambarkan skenario potensial phishing dalam teks. Pendekatan ini dimulai dengan membuat taksonomi umum dari proses phishing email, kemudian mengembangkan konsep DL-based TBox dan ABox untuk mendeteksi phishing secara lebih efektif [13].

3. Metode Penelitian

3.1 Definisi Masalah

Studi ini berfokus pada Intrusion Detection System (IDS) yang berbasis pada fitur dataset menggunakan Generative Adversarial Network (GAN). Dalam dataset, terdapat sekumpulan fitur x ($x_1, x_2, x_3, \dots, x_n$). GAN membangun model generatif melalui pendekatan adversarial. GAN terdiri dari dua model utama, yaitu generator (G) dan discriminator (D).

Model G bertugas memperkirakan distribusi data $p(g)$ berdasarkan ruang data asli x . Dengan mempertimbangkan variabel gangguan input $p(z)$, tujuan utama G adalah menghasilkan sampel baru $G(z)$ yang memiliki distribusi yang sama dengan x .

Sementara itu, discriminator (D) berfungsi mengembalikan probabilitas $D(x)$ yang menunjukkan apakah suatu sampel x berasal dari dataset asli atau merupakan data yang dihasilkan oleh G [30]. Nilai dari fungsi $V(D, G)$ didefinisikan sebagai berikut:

$$\min_G \max_D V(D, G) = E_{X \sim P_{data}(x)} [\log D(X)] + E_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (1)$$

Table 1. Mathematic notation of the GAN

Notation	Description
$V(D, G)$	Value function
D	Discriminator
G	Generator
E	Denotes the expectation
P	Represent the number of samples x
$P_{data}(x)$	Input data
$p_z(z)$	Noise variables
x	Input vector
z	Noise vector
$G(z)$	The false data generated by generator
F	Accuracy function
$F(D(x))$	Output result of discriminator model

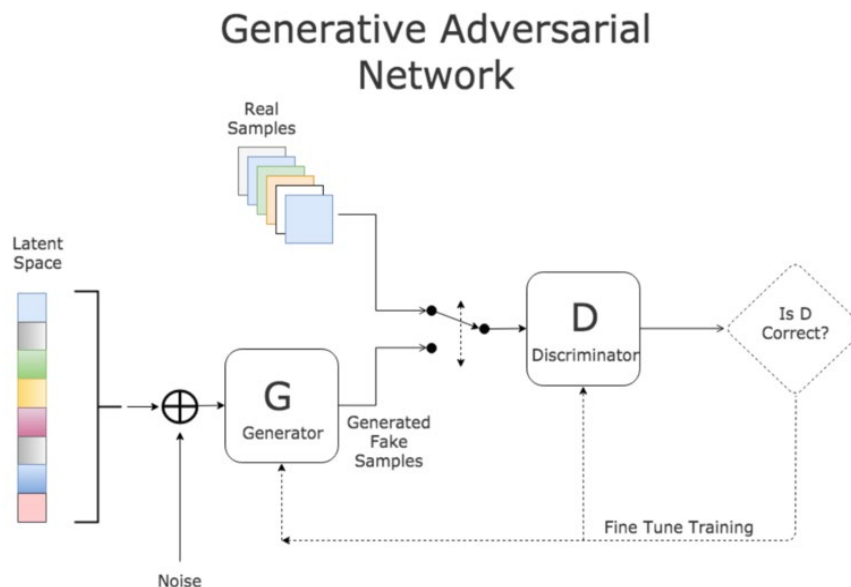
Karena tujuan utama model GAN adalah mengklasifikasikan sampel asli atau palsu melalui D , maka $V(D, G)$ menjadi pilihan terbaik sebagai fungsi objektif dalam masalah klasifikasi.

Model G memetakan noise (z) dari latent space ke dalam data input yang akan diberikan kepada D . Istilah $[\log (1 - D(G(z)))]$ menunjukkan bahwa jika suatu sampel berasal dari data asli, maka D akan memaksimalkan keluarannya. Sebaliknya, jika D menentukan bahwa suatu sampel berasal dari G , maka D akan meminimalkan keluarannya.

Pada saat yang sama, G berusaha untuk memaksimalkan keluaran D dengan mengirimkan sampel palsu hasil generasinya ke D , sehingga D mengalami kebingungan dalam membedakan antara sampel asli dan palsu.

3.2 Metode yang diusulkan

Eksperimen dalam penelitian ini menggunakan GAN. GAN adalah model generatif yang diperkenalkan oleh Goodfellow *et al.* pada tahun 2014 [16][17]. GAN terdiri dari dua jaringan utama, yaitu discriminator dan generator. Generator bertujuan untuk menciptakan sampel baru (*fake sample*) berdasarkan data asli. Sebaliknya, discriminator bertugas membedakan apakah suatu sampel berasal dari data asli atau dihasilkan oleh generator yang telah dilatih [18]. Gambar 1 menunjukkan arsitektur dari GAN.



Gambar 1 Architecture of GAN.

Berdasarkan fungsi loss pada model discriminator dan generator, algoritma back-propagation (BP) dapat digunakan untuk memperbarui parameter dalam GAN. Selain itu, parameter GAN dapat dihitung secara efektif melalui algoritma gradient BP. Oleh karena itu, metode berbasis GAN digunakan untuk menangani masalah intrusion detection. Pertama, generator (G) memilih contoh z dari random noise $p_z(z)$, sehingga $G(z)$ merepresentasikan data palsu yang dihasilkan oleh generator. Sementara itu, discriminator (D) mengenali contoh x dari input $Pdata(x)$ dan melakukan evaluasi terhadapnya:

$$E_{x \in Pdata(x)} \log (F(D(x))) \quad (2)$$

$F(D(x))$ adalah output dari model discriminator, yang merupakan nilai real dalam rentang 0 – 1, digunakan untuk menentukan probabilitas keakuratan data. Memaksimalkan fungsi ini berarti D dapat memprediksi nilai normal dengan akurat, sehingga $F(D(x)) = 1$ jika $x \in$

$Pdata(x)$. Selanjutnya, kita memverifikasi data yang dihasilkan oleh generator melalui persamaan:

$$E_{Z \in P_Z(z)} \log (1 - F(D(G(z)))) (3)$$

Jika fungsi ini dimaksimalkan, maka $F(D(G(z))) \approx 0$, yang berarti G gagal menghasilkan data palsu yang meyakinkan. Tujuan utama dari generator adalah menciptakan data yang mampu mengecoh discriminator. Selama proses pelatihan jaringan, fungsi objektif dari discriminator dapat didefinisikan sebagai:

$$\max_D E_{Z \in P_Z(z)} [\log (1 - D(G(z)))] + E_{x \in P_{data}(x)} [\log(D(x))] (4)$$

Fungsi ini menunjukkan bahwa tujuan discriminator adalah memaksimalkan jumlah dari dua ekspresi di atas. Oleh karena itu, fungsi nilai (value function) GAN dapat didefinisikan sebagai:

$$V(D, G) = E_{x \in P_{data}(x)} [\log ((F(D(x))))] + E_{Z \in P_Z(z)} [\log ((1 - F(D(G(z)))))] (5)$$

Ketika G tetap konstan, maka D^*G yang optimal dapat dinyatakan sebagai $D^*G = \arg \max_D V(G, D)$, yang merepresentasikan discriminator terbaik. Sebaliknya, ketika D tetap konstan, G harus memenuhi $G^*D = \arg \min_G V(G, D^*)$ untuk memperoleh generator (G) yang optimal. Setelah sejumlah iterasi pelatihan, kita dapat memperoleh discriminator optimal D^*G , yang memenuhi $D^*G = \arg \max_D V(G^*D, D)$ [19].

4. Rancangan Percobaan

4.1 Gagasan Utama

Makalah ini bertujuan untuk mengembangkan model deteksi intrusi jaringan dengan memanfaatkan fitur dalam dataset menggunakan algoritma GAN. Dalam pendekatan ini, GAN digunakan untuk menghasilkan sampel data anomali berdasarkan data standar, yang kemudian digunakan untuk mengoptimalkan detektor serta meningkatkan tingkat deteksi terhadap serangan lawan [16][17]. Metode ini dirancang untuk mendeteksi gangguan dan mengklasifikasikannya sesuai dengan label yang tersedia. Meskipun GAN awalnya banyak digunakan dalam bidang pengolahan gambar, belakangan ini teknologi tersebut telah diperluas ke berbagai domain, termasuk keamanan siber. Penelitian dalam bidang keamanan semakin banyak mengeksplorasi potensi GAN sebagai model generatif baru. Kemampuannya dalam mendeteksi ancaman yang tidak terlihat menjadikannya sebagai mekanisme pertahanan yang lebih tangguh terhadap serangan di masa depan. Dengan tingkat akurasi yang tinggi, GAN sangat cocok untuk menangani berbagai tantangan dalam deteksi ancaman siber. Oleh karena itu, dalam penelitian ini, kami menggunakan GAN untuk membangun model deteksi. Selain itu, karena GAN memiliki potensi besar dalam meniru distribusi data, model ini memungkinkan sistem IDS (Intrusion Detection System) untuk memanfaatkan lebih banyak data yang tersedia, sehingga meningkatkan efektivitas deteksinya.

4.2 Dataset

Dalam eksperimen ini, kami mengumpulkan dataset yang berisi informasi mengenai jaringan normal dan abnormal. Dataset ini kemudian dibagi menjadi dua bagian untuk membangun model, yaitu:

1. Training dataset – digunakan untuk membangun dan melatih model.
2. Testing dataset – digunakan untuk mengevaluasi kinerja dan akurasi model.

Dalam penelitian ini, kami membagi dataset dengan proporsi 70% untuk pelatihan dan 30% untuk pengujian. Tabel 2 menunjukkan rincian distribusi dataset yang digunakan dalam studi ini, seperti berikut:

Table 2. Details distribution of the dataset

Dataset	Sample
Data training (%)	546
Data testing (%)	234
Total	780

4.3 Data Pre-processing

Pada tahap pre-processing, kami melakukan pemrosesan dan penyaringan data dengan mengevaluasi tipe data dari setiap variabel dalam dataset serta memeriksa keberadaan data yang kosong (NA/missing values). Jika ditemukan data NA atau kosong, maka akan dihapus untuk meningkatkan efisiensi dan mengurangi penggunaan penyimpanan yang berlebihan [19]. Namun, jika terdapat data kosong yang tidak bisa dihapus, maka akan dilakukan penanganan nilai yang hilang (missing value handling).

Selanjutnya, proses pre-processing meliputi beberapa tahapan berikut:

- Vektorisasi (vectorization)
- Normalisasi (normalization)
- Penanganan missing values
- Ekstraksi fitur (feature extraction)

Pada tahap ini, dataset terlebih dahulu dibagi menjadi data training dan data testing. Model akan dilatih menggunakan data training dan divalidasi untuk memastikan performanya. Setelah model dinyatakan valid, maka model akhir akan diuji menggunakan data testing untuk mengevaluasi akurasi.

4.4 Metode Deteksi

Dalam eksperimen ini, kami mengumpulkan dataset yang berisi fitur jaringan normal dan anomali. Setelah dataset dikumpulkan, kami membaginya menjadi dua bagian sebelum digunakan dalam proses pelatihan (training) dan pengujian (testing). Untuk memudahkan proses pelatihan, dataset dikonversi ke dalam nilai vektor dengan memilih fitur-fitur yang dianggap relevan.

Langkah selanjutnya adalah pre-processing data, yang mencakup pemrosesan dan penyaringan dengan memeriksa tipe data dari setiap variabel dalam dataset serta mendeteksi keberadaan data kosong (NA/missing values). Jika terdapat data NA, maka data tersebut harus dihapus agar dataset yang digunakan lebih optimal untuk proses pelatihan dan pengujian, serta untuk menghindari penyimpanan data yang berlebihan. Setelah tahap pre-processing selesai, fitur-fitur yang telah diperoleh akan digunakan untuk melatih model deteksi.

Dalam penelitian ini, kami menggunakan GAN untuk membangun model Intrusion Detection System (IDS). Model ini bekerja dengan cara generator menghasilkan sampel baru yang menyerupai data pelatihan untuk membingungkan discriminator dalam membedakannya. Kami

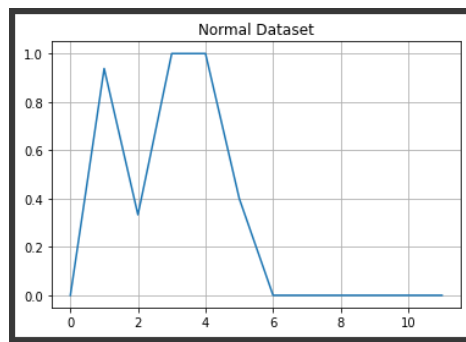
menggunakan dataset pelatihan asli untuk melatih model dan dataset pengujian untuk mengukur performanya.

Selama proses pengujian, kami menyediakan berbagai dataset yang berbeda untuk mendapatkan model yang optimal. Kami juga memodifikasi beberapa parameter guna memperoleh akurasi terbaik. Akhirnya, model yang diusulkan ini diharapkan mampu menghasilkan deteksi gangguan jaringan (network intrusion detection) yang lebih efektif.

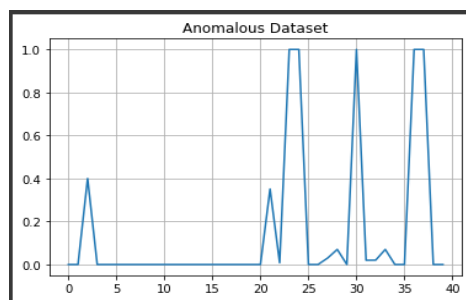
5. Hasil dan Pembahasan

5.1 Pengujian Deteksi

Dalam penelitian ini, kami mengumpulkan dataset yang berisi data normal dan anomali. Gambar 2 menunjukkan grafik dataset deteksi jaringan normal, sedangkan Gambar 3 menampilkan grafik dataset deteksi jaringan anomali. Grafik deteksi jaringan normal dan anomali ditampilkan dalam rentang 12, yang ditunjukkan oleh garis plot biru.

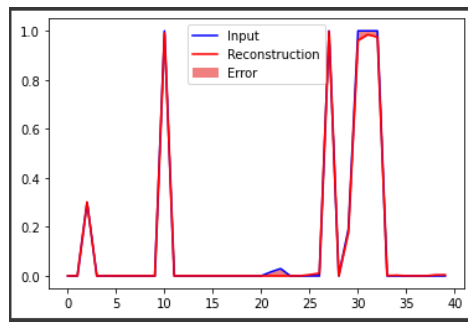


Gambar 2 Normal Data

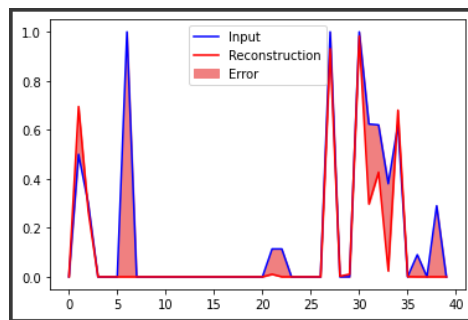


Gambar 3 Anomaly Data

Suatu jaringan diklasifikasikan sebagai anomali jika kesalahan rekonstruksi lebih besar dari satu standar deviasi dari sampel pelatihan normal. Dalam proses ini, model yang kami usulkan menghitung pola jaringan normal berdasarkan dataset pelatihan. Model ini melakukan rekonstruksi data menggunakan GAN Autoencoder dan menghitung kesalahan rekonstruksi pada tahap pengujian. Gambar 4 menunjukkan hasil pengujian deteksi normal, sedangkan Gambar 5 menampilkan hasil pengujian deteksi anomali sebagai berikut:



Gambar 4 Normal Detetion Test



Gambar 5 Anomaly Detection Test

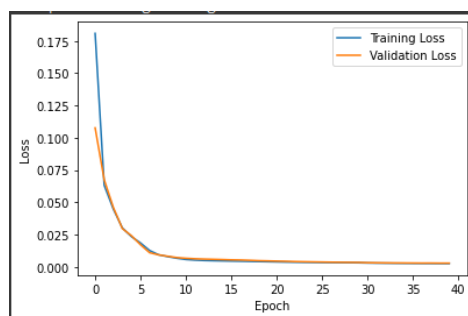
Gambar 3 menunjukkan hasil pengujian pada data normal, yang memiliki tingkat kesalahan rendah. Sementara itu, Gambar 4 menampilkan hasil pengujian pada data anomali, yang menunjukkan tingkat kesalahan lebih tinggi.

Dalam proses pengujian:

- Garis biru menunjukkan data input.
- Garis merah menggambarkan data hasil rekonstruksi.
- Garis merah tebal di antara garis biru dan merah merepresentasikan data kesalahan rekonstruksi.

5.2 Metrik Evaluasi

Dalam penelitian ini, sampel data dimasukkan ke dalam model untuk proses pelatihan dan pengujian guna membangun model deteksi. Gambar 6 menampilkan training loss dan validation loss, yang digunakan untuk mengevaluasi kinerja model.



Gambar 6 Training Loss and Validation Loss

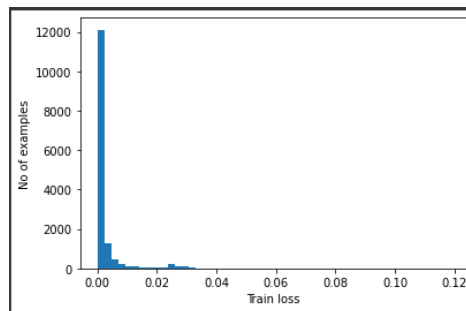
Dalam eksperimen ini, kami menyesuaikan berbagai hyperparameter untuk mendapatkan kinerja jaringan terbaik. Selama proses pelatihan dan pengujian, kami menetapkan:

- Epoch = 3.000
- Batch size = 128

Pada grafik:

- Garis biru menunjukkan training loss
- Garis merah menunjukkan validation loss

Kami menetapkan nilai hyperparameter yang tinggi untuk menghasilkan model pelatihan yang optimal. Semakin tinggi jumlah epoch, semakin rendah nilai training loss selama proses pelatihan. Training loss yang rendah menandakan bahwa kinerja model semakin baik. Model ini mendeteksi anomali dengan menghitung apakah reconstruction loss lebih besar dari ambang batas yang telah ditentukan. Model pertama-tama menghitung rata-rata error dari sampel normal dalam training set, kemudian mengklasifikasikan sampel baru sebagai anomali jika nilai reconstruction error lebih tinggi dari satu standar deviasi. Gambar 7 menunjukkan plot reconstruction error pada jaringan normal dalam training set.



Gambar 7 Reconstruction Error

Gambar 7 menunjukkan plot reconstruction error pada jaringan normal dari training set. Kami menganalisis hasil deteksi dengan menghitung threshold untuk menguji model deteksi. Penentuan nilai threshold dilakukan dengan menjumlahkan nilai rata-rata (mean) dan simpangan baku (standard deviation) dari training loss. Dari pendekatan ini, kami memperoleh threshold sebesar 0.0054779826.

- Semakin kecil nilai threshold, semakin akurat model deteksi dan semakin rendah tingkat error.
- Dengan nilai threshold yang optimal, model kami dapat memberikan hasil deteksi yang lebih akurat dibandingkan dengan metode deteksi tradisional sebelumnya dalam menangani anomali.

6. Kesimpulan

Akun palsu merupakan salah satu tantangan keamanan kritis yang dihadapi dalam sistem jaringan. Seiring dengan meningkatnya ketergantungan pada Internet dan perangkat digital, penyebaran ancaman keamanan juga semakin luas. Metode deteksi intrusi tradisional tidak lagi cukup untuk mendeteksi serangan dengan pola yang tidak terduga. Oleh karena itu, tantangan dalam menjaga keamanan sistem dan perangkat dari celah keamanan serta serangan siber juga

semakin besar.

Dalam penelitian ini, kami mengembangkan model deteksi intrusi berbasis fitur yang diekstrak dari data API. Model ini mampu menganalisis dan mendeteksi ancaman dengan cepat, sehingga dapat mencegah atau menangani intrusi sebelum merusak sistem dan data. Berdasarkan hasil eksperimen, model deteksi yang diusulkan menunjukkan nilai threshold sebesar 0.0054779826, di mana nilai threshold yang lebih kecil menunjukkan model yang lebih akurat dengan tingkat kesalahan yang lebih rendah.

Dengan demikian, model ini dapat menjadi solusi potensial dalam mendeteksi intrusi dan menjaga keamanan jaringan. Untuk meningkatkan efektivitas deteksi, penelitian selanjutnya dapat mengimplementasikan algoritma yang lebih mutakhir guna meningkatkan performa model ini. Seiring dengan kemajuan teknologi pembelajaran mesin, pendekatan pembelajaran dinamis dapat dikembangkan untuk lebih mengoptimalkan model deteksi intrusi di masa depan.

Referensi

- [1] Mohammadrezaei M, Shiri ME, Rahmani AM. Identifying fake accounts on social networks based on graph analysis and classification algorithms". *Hindawi Secur Commun Netw* 2018
- [2] Wanda, Putra and Huang Jin Jie." DeepProfile: Finding fake profile in online social network using dynamic CNN". *Journal of Information Security and Applications*.2020.
- [3] Li M, Xiang Y, Zhang B, Wei F, Song Q. "A novel organizing scheme of single topic user group based on trust chain model in social network. *Int J Commun Syst* 2018;31:3387.
- [4] Anglano C, Canonico M, Guazzone M. "Analysis of telegram messenger on android smartphones." *Digital investigation, Elsevier*, 23; 2017. p. 31–49.
- [5] C. T, E. Dwoskin, Twitter Is Sweeping Out Fake Accounts like Never before, Putting User Growth at Risk, 2018. <https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/>
- [6] A. Wani, S. Jabin, A. Nehaluddin, A Sneak into the Devil's Colony-fake Profiles in Online Social Networks, *arXiv.org ePrint Archive* vol. 5, 2018, pp. 26-39.
- [7] E.V.D. Walt, J. Eloff, Using machine learning to detect fake identities: bots vs humans, *IEEE Access* 6 (2018) 6540e6549, <https://doi.org/10.1109/ACCESS.2018.2796018>.
- [8] E. Ahmadizadeh, E. Aghasian, H.P. Taheri, R.F. Nejad, An automated model to detect fake profiles and botnets in online social networks using steganography technique, *IOSR J. Comput. Eng.* 17 (2015) 65e71, <https://doi.org/10.9790/0661-17146571>.
- [9] R.S. Perdana, T.H. Muliawati, R. Alexandro, Bot spammer detection in Twitter using tweet similarity and time interval entropy, *Jurnal Ilmu Komputer dan Informasi* 8 (2015) 19e25, <https://doi.org/10.21609/jiki.v8i1.280>.
- [10] S.M. Asadullah, S. Viraktamath, Classification of twitter spam based on profile and message model using Svm, *Int. J. Eng. Res. Technol.* 4 (2017) 2862e2865.
- [11] Ç.B. Aslan, R.B. Saglam, S. Li, Automatic detection of cyber security related accounts on online social networks: twitter as an example, in: *Proceedings of the 9th International*

- Conference on Social Media and Society, 2018, pp. 236e240, <https://doi.org/10.1145/3217804.3217919>.
- [12] G. Qingyuan, Y. Chen, X. He, Z. Zhuang, T. Wang, H. Huang, X. Wang, X. Fu, DeepScan: Exploiting deep learning for malicious account detection in location-based social networks, *IEEE Commun. Mag.* 56 (2018) 21e27, <https://doi.org/10.1109/MCOM.2018.1700575>.
- [14] Q. Fu, B. Feng, D. Guo, Q. Li, Combating the evolving spammers in online social networks, *Comput. Secur.* 72 (2018) 60e73, <https://doi.org/10.1016/j.cose.2017.08.014>.
- [15] S. Rathore, A.K. Sangaiah, J.H. Park, A novel framework for internet of knowledge protection in social networking services, *J. Comput. Sci.* 26 (2018) 55e65, <https://doi.org/10.1016/j.jocs.2017.12.010>.
- [16] Jie, Huang. "A Survey of Network and Information Security. *International Journal of Informatics and Computation*". [S.l.], v. 1, n. 2, p. 15-22, dec. 2019.
- [17] Dr. Dongseong Kim, Xiran Zhang. *Network Intrusion Detection Using Generative Adversarial Networks*.
- [18] Purser et al." *Using Generative Adversarial Networks For Intrusion Detection In Cyber-Physical Systems*".2020.
- [19] Nie, Laisen et al. "Intrusion Detection for Secure Social Internet of Things Based on Collaborative Edge Computing: A Generative Adversarial Network-Based Approach".*IEEE*.2021.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>)

