Socius: Jurnal Penelitian Ilmu-Ilmu Sosial
Volume 4, Issue 2, September 2026, p 208-218
E-ISSN: 3025-6704
DOI: https://doi.org/10.5281/zenodo.22211310

Augmentasi Moderat untuk Peningkatan Kemampuan Deteksi Anomali Serangan pada Dataset UNSW-NB15

Moderate Data Augmentation for Improving Anomalous Attack Detection on the UNSW-NB15 Dataset

Sebastianus Hartantyo[1], Alva Hendi Muhammad[2]
[1,2]Program Studi S2 Informatika, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta, Indonesia

ARTICLE INFO

Article history:
Received 10 Auguts 2026
Revised 18 Auguts 2026
Accepted 25 August 2026
Available online 1 September 2026

Keywords

augmentasi data; class imbalance; intrusion detection system; SDAE-GAN; CTGAN; UNSW-NB15

ABSTRACT

Class imbalance remains a persistent challenge in intrusion detection because infrequent attacks may carry substantial security impact while being poorly represented during model training. This study evaluates a moderate augmentation strategy for UNSW-NB15 by selectively limiting synthetic data growth instead of forcing every class to match the majority class. Four extremely underrepresented classes—Analysis, Backdoor, Shellcode, and Worms—were augmented using the target T_k = min(10,000, 15 x n_k). Network records were first mapped into a latent representation using a Stacked Denoising Autoencoder (SDAE), followed by five generative scenarios: Hybrid SDAE-GAN, SDAE Feature Extraction, SDAE Dimensionality Reduction, SDAE-WGAN-GP, and SDAE-CTGAN. No-augmentation and SMOTE baselines were included for comparison. S5-CTGAN achieved the highest macro F1-score of 0.4135, a 5.9% improvement over the no-augmentation baseline (0.3906), with an MCC of 0.5897. The most visible class-level gains were obtained for Analysis (F1 from 0.018 to 0.105) and DoS (0.206 to 0.375). S2-Feature Extraction produced the closest synthetic distribution with a mean Wasserstein Distance of 0.0172. A one-way ANOVA confirmed highly significant differences among experimental setups (p = 4.37 x 10^-41; eta-squared = 0.957), supported by Friedman and Tukey HSD tests. The results indicate that generating more synthetic samples is not inherently beneficial: aggressive full balancing can cause over-amplification, whereas controlled latent-space augmentation provides more stable gains while preserving the role of genuine observations.

ABSTRAK

Ketidakseimbangan kelas menjadi persoalan yang sulit diabaikan pada sistem deteksi intrusi, karena serangan yang jarang muncul justru sering memiliki nilai keamanan yang tinggi. Penelitian ini menguji strategi augmentasi moderat pada UNSW-NB15 dengan membatasi penambahan sampel sintetis secara selektif, bukan menyamakan seluruh kelas dengan kelas mayoritas. Empat kelas dengan ketimpangan paling ekstrem, yaitu Analysis, Backdoor, Shellcode, dan Worms, menjadi sasaran augmentasi dengan target T_k = min(10.000, 15 x n_k). Data dipetakan terlebih dahulu ke ruang laten menggunakan Stacked Denoising Autoencoder (SDAE), kemudian dibangkitkan melalui lima skenario generatif: Hybrid SDAE-GAN, SDAE Feature Extraction, SDAE Dimensionality Reduction, SDAE-WGAN-GP, dan SDAE-CTGAN. Dua baseline, tanpa augmentasi dan SMOTE, digunakan sebagai pembanding. Hasil pengujian menunjukkan bahwa S5-CTGAN memberikan F1-Macro tertinggi sebesar 0,4135, meningkat 5,9% dibanding baseline tanpa augmentasi sebesar 0,3906, dengan MCC 0,5897. Perbaikan paling nyata terjadi pada kelas Analysis (F1 0,018 menjadi 0,105) dan DoS (0,206 menjadi 0,375). S2-Feature Extraction menghasilkan kemiripan distribusi sintetis terbaik dengan Wasserstein Distance 0,0172. ANOVA menunjukkan perbedaan yang sangat signifikan antarsetup (p = 4,37 x 10^-41; eta-kuadrat = 0,957), dan hasil tersebut konsisten dengan uji Friedman serta Tukey HSD. Temuan utama penelitian ini adalah bahwa jumlah data sintetis yang lebih besar tidak selalu lebih baik: full balancing memicu over-amplification, sedangkan augmentasi yang dibatasi dan dilakukan pada ruang laten memberikan perbaikan yang lebih stabil tanpa membuat data sintetis mendominasi data asli.

PENDAHULUAN

Sistem deteksi intrusi (Intrusion Detection System/IDS) berbasis pembelajaran mesin berkembang karena mampu mengenali pola serangan yang tidak selalu dapat dirumuskan sebagai signature statis. Namun, kualitas keputusan model sangat bergantung pada pola yang

Email: hartantyo@students.amikom.ac.id, alva@amikom.ac.id

tersedia selama pelatihan. Pada trafik jaringan, distribusi kelas hampir tidak pernah benar-benar seimbang: aktivitas normal dan beberapa tipe serangan mendominasi, sementara serangan lain hanya muncul dalam jumlah kecil. Ketimpangan tersebut membuat metrik akurasi mudah terlihat tinggi pada saat kemampuan mendeteksi kelas langka masih rendah. Literatur mengenai IDS modern juga menempatkan class imbalance sebagai salah satu faktor utama yang memengaruhi false negative pada kelas minoritas [1]-[4].

UNSW-NB15 menyediakan trafik normal dan sembilan kelompok serangan yang lebih modern dibanding benchmark klasik seperti KDD99. Dataset ini banyak digunakan untuk mengevaluasi anomaly-based NIDS karena memuat variasi fitur aliran, paket, waktu, dan konteks layanan jaringan [1], [2]. Di sisi lain, distribusinya sangat timpang. Dalam pembagian resmi yang digunakan pada penelitian ini, kelas Normal berjumlah 93.000 sampel, sedangkan Worms hanya 174 sampel. Rasio sekitar 1:534,5 tersebut membuat satu kesalahan pada kelas langka mempunyai konsekuensi statistik yang jauh lebih besar daripada pada kelas mayoritas.

Pendekatan yang paling langsung untuk menghadapi ketimpangan adalah oversampling. SMOTE dan berbagai turunannya menambah sampel minoritas melalui interpolasi di antara tetangga terdekat [5], [6]. Teknik tersebut praktis, tetapi pada distribusi trafik yang sangat miring dan memiliki banyak outlier, interpolasi dapat menghasilkan titik baru pada wilayah yang tidak cukup merepresentasikan struktur kelas. Pengembangan model generatif kemudian membuka alternatif lain: alih-alih sekadar menginterpolasi, model mempelajari distribusi data dan membangkitkan sampel baru [7]. Autoencoder, GAN, WGAN-GP, dan CTGAN telah dilaporkan dapat memperbaiki representasi kelas serangan pada beberapa dataset IDS [9]-[15].

Meski demikian, persoalan pentingnya bukan hanya bagaimana membangkitkan data sintetis, tetapi juga berapa banyak data sintetis yang layak ditambahkan. Eksperimen awal pada penelitian ini menunjukkan bahwa full balancing—menyamakan setiap kelas hingga 56.000 sampel pelatihan—justru menimbulkan over-amplification. Pada Worms, misalnya, 130 data asli dipaksa menjadi 56.000 sehingga hampir seluruh kelas hasil balancing berasal dari generator. Kondisi semacam ini berisiko membuat classifier lebih mengenali karakter sintetis daripada variasi nyata. Temuan sejenis menjadi relevan dengan kajian oversampling yang menunjukkan bahwa penambahan sampel tanpa pengendalian dapat memperbesar overlap, noise, dan over-generalization [6].

Berangkat dari masalah tersebut, penelitian ini memosisikan rasio augmentasi sebagai bagian utama dari desain IDS, bukan sekadar parameter tambahan. Strategi yang diuji disebut augmentasi moderat: hanya kelas dengan ketimpangan ekstrem yang ditambah dan jumlah akhirnya dibatasi oleh nilai minimum antara 10.000 sampel atau 15 kali jumlah data asli. Augmentasi dilakukan pada ruang laten SDAE agar classifier tidak perlu menerima hasil decode yang berpotensi membawa reconstruction error. Dengan rancangan tersebut, penelitian tidak berusaha membuat seluruh kelas sama besar, tetapi berupaya memberikan cukup contoh tambahan agar batas keputusan kelas langka menjadi lebih terbentuk tanpa menenggelamkan data nyata.

Kontribusi penelitian ini terdiri atas empat bagian. Pertama, merumuskan skema pembatasan augmentasi $T_k = \min(10.000, 15 \times n_k)$ yang diterapkan hanya pada Analysis, Backdoor, Shellcode, dan Worms. Kedua, membandingkan strategi tersebut pada lima skenario generatif berbasis SDAE serta dua baseline. Ketiga, mengevaluasi bukan hanya performa klasifikasi melalui F1-Macro dan Matthews Correlation Coefficient (MCC), tetapi juga kualitas distribusi sintetis menggunakan Wasserstein Distance dan t-SNE. Keempat, memvalidasi perbedaan performa melalui pengulangan eksperimen dan uji statistik parametrik maupun nonparametrik. Fokus artikel ini dengan demikian adalah menjawab pertanyaan yang lebih operasional: apakah

augmentasi yang dibatasi dapat meningkatkan kemampuan deteksi serangan pada kondisi imbalance ekstrem secara lebih sehat daripada augmentasi agresif?

METODE PENELITIAN

Dataset dan Distribusi Kelas

Penelitian menggunakan UNSW-NB15 dengan pembagian 175.341 data latih dan 82.332 data uji, sehingga totalnya 257.673 rekaman pada sepuluh kelas. Pemisahan data latih dan uji dipertahankan sebelum proses resampling untuk mencegah kebocoran data. Distribusi kelas menunjukkan empat kelompok dengan imbalance ratio tertinggi, yaitu Analysis, Backdoor, Shellcode, dan Worms. Keempatnya kemudian ditetapkan sebagai target augmentasi. Distribusi lengkap disajikan pada Tabel 1.

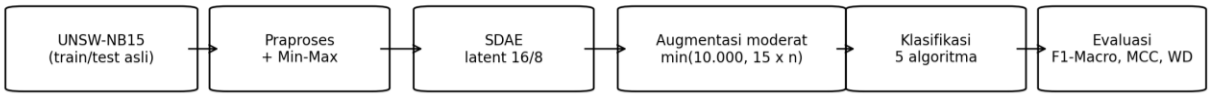
Tabel 1. Distribusi kelas UNSW-NB15 yang digunakan dalam eksperimen

Kelas	Latih	Uji	Total	IR vs Normal
Normal	56.000	37.000	93.000	1,0
Generic	40.000	18.871	58.871	1,6
Exploits	33.393	11.132	44.525	2,1
Fuzzers	18.184	6.062	24.246	3,8
DoS	12.264	4.089	16.353	5,7
Reconnaissance	10.491	3.496	13.987	6,6
Analysis	2.000	677	2.677	34,7
Backdoor	1.746	583	2.329	39,9
Shellcode	1.133	378	1.511	61,5
Worms	130	44	174	534,5

Praproses dan Representasi Fitur

Tahap praproses mencakup pemeriksaan nilai kosong dan duplikasi, transformasi fitur kategorikal, serta normalisasi fitur numerik menggunakan Min-Max Scaling pada rentang [0,1]. Tiga puluh satu fitur numerik inti digunakan sebagai masukan SDAE. Nilai ekstrem tidak dihapus secara otomatis. Pada data keamanan jaringan, outlier dapat merepresentasikan lonjakan volume paket, jitter, kehilangan paket, atau pola waktu yang justru merupakan ciri serangan; menghapusnya hanya karena berada di luar IQR berisiko menghilangkan informasi keamanan yang relevan.

SDAE digunakan untuk memperoleh representasi yang lebih ringkas dan tahan noise. Arsitektur utama bersifat simetris: Input(31)-128-64-32-Latent(16)-32-64-128-Output(31). Lapisan tersembunyi menggunakan aktivasi ReLU dan Batch Normalization, sedangkan latent layer menggunakan sigmoid agar konsisten dengan skala [0,1]. Model dioptimasi dengan Adam pada learning rate 10^-4. Sebanyak 41.009 sampel hasil stratified sampling digunakan untuk melatih SDAE sehingga kelas minoritas tetap hadir pada proses pembelajaran representasi. SDAE-16 menghasilkan reconstruction MSE 0,00097. Sebagai pembanding kompresi yang lebih agresif, SDAE-8 menghasilkan MSE 0,00136 dan digunakan khusus pada skenario reduksi dimensi.



Gambar 1. Alur eksperimen augmentasi moderat pada ruang laten

Strategi Augmentasi Moderat

Target augmentasi untuk setiap kelas minoritas k ditentukan berdasarkan jumlah data latih asli n_k. Alih-alih mengacu pada jumlah kelas mayoritas, penelitian menetapkan batas berikut:

$$T_k = \min(10.000, 15 \times n_k)$$

Jumlah data sintetis yang dibangkitkan adalah T_k - n_k. Dua pembatas bekerja sekaligus: plafon absolut 10.000 sampel mencegah kelas kecil menjadi terlalu besar, sedangkan plafon relatif 15 kali data asli mencegah kelas ultra-langka mengalami amplifikasi yang tidak proporsional. Dengan aturan ini, total set pelatihan setelah augmentasi menjadi 202.282 sampel, jauh lebih kecil daripada full balancing yang dapat melampaui 560.000 sampel. Rinciannya terlihat pada Tabel 2.

Tabel 2. Rincian augmentasi moderat pada empat kelas minoritas ekstrem

Kelas	Data asli	Sintetis	Total	Amplifikasi
Analysis	2.000	8.000	10.000	5,0 x
Backdoor	1.746	8.254	10.000	5,7 x
Shellcode	1.133	8.867	10.000	8,8 x
Worms	130	1.820	1.950	15,0 x (batas)

Strategi ini sengaja tidak mengaugmentasi seluruh kelas. Kelas seperti DoS dan Reconnaissance tetap dipertahankan sesuai distribusi aslinya karena jumlah sampelnya sudah cukup untuk membentuk representasi yang relatif lebih baik. Prinsip tersebut membedakan augmentasi moderat dari balancing konvensional: tujuan utamanya bukan kesetaraan jumlah kelas, melainkan menambah informasi pada titik distribusi yang paling kekurangan bukti pelatihan.

Skenario Eksperimen dan Model Generatif

Tujuh setup digunakan agar manfaat augmentasi moderat dapat dipisahkan dari pengaruh arsitektur generator. Conditional GAN mewakili model adversarial konvensional, WGAN-GP digunakan untuk stabilitas optimasi melalui gradient penalty [10], sedangkan CTGAN dipilih karena dirancang untuk distribusi tabular yang multimodal dan tidak seimbang [11]. Konfigurasi eksperimen dirangkum pada Tabel 3.

Tabel 3. Setup eksperimen

Kode	Setup	Komponen utama
B1	Baseline No-Augmentation	Data imbalanced asli
B2	Baseline SMOTE	Oversampling berbasis tetangga terdekat
S1	Hybrid SDAE-GAN	SDAE-16 + Conditional GAN
S2	SDAE Feature Extraction	SDAE-16 frozen + Conditional GAN
S3	SDAE Dimensionality Reduction	SDAE-8 + Conditional GAN
S4	SDAE-WGAN-GP	SDAE-16 + Wasserstein GAN-GP

Kode	Setup	Komponen utama
S5	SDAE-CTGAN	SDAE-16 + Conditional Tabular GAN

Klasifikasi dan Evaluasi

Data asli dan sintetis pada ruang laten digunakan untuk melatih lima classifier: Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbors (KNN), dan Naive Bayes (NB). Evaluasi utama menggunakan data uji asli yang tidak pernah diresampling. Random Forest kemudian digunakan sebagai classifier representatif pada analisis statistik karena memberikan performa paling konsisten pada eksperimen latent-space.

Metrik klasifikasi mencakup accuracy, macro precision, macro recall, F1-Macro, weighted F1, G-Mean, dan MCC. F1-Macro ditempatkan sebagai metrik utama karena memberi bobot yang sama kepada setiap kelas, sehingga lebih sensitif terhadap kegagalan pada kelas minoritas daripada accuracy. MCC digunakan sebagai ukuran korelasi prediksi-label yang tetap informatif pada distribusi tidak seimbang. Kualitas data sintetis dianalisis secara terpisah menggunakan Wasserstein Distance (WD); semakin kecil WD, semakin dekat distribusi sintetis terhadap data asli. Visualisasi t-SNE digunakan sebagai pemeriksaan kualitatif atas overlap dan struktur manifold.

Untuk mengurangi kemungkinan kesimpulan yang hanya berasal dari satu random seed, evaluasi Random Forest diulang sepuluh kali dengan seed berbeda. Normalitas F1-Macro diperiksa menggunakan Shapiro-Wilk, homogenitas varians dengan Levene, perbedaan antarkelompok dengan one-way ANOVA, dan besar efek dengan eta-squared. Uji Friedman digunakan sebagai validasi nonparametrik, sedangkan Tukey HSD mengidentifikasi setup yang berbeda signifikan terhadap baseline tanpa augmentasi. Taraf signifikansi ditetapkan pada 0,05. Seluruh eksperimen dijalankan pada Google Colaboratory menggunakan akselerator GPU NVIDIA T4.

HASIL DAN PEMBAHASAN

Mengapa Full Balancing Tidak Dipertahankan

Sebelum strategi moderat digunakan, penelitian menguji augmentasi pada feature-space dengan full balancing hingga 56.000 sampel pada setiap kelas. Hasilnya tidak mendukung asumsi bahwa semakin seimbang jumlah kelas maka semakin baik performa IDS. Dalam sepuluh iterasi eksperimen awal, baseline tanpa augmentasi masih memiliki rata-rata F1-Macro sekitar 0,4542, sedangkan skenario augmentasi agresif berada pada kisaran 0,42-0,43. Angka tersebut berasal dari eksperimen feature-space dan tidak dibandingkan langsung dengan nilai latent-space pada bagian berikutnya; fungsi utamanya adalah menunjukkan arah kegagalan ketika jumlah data sintetis dibuat terlalu dominan.

Kelas Worms menjadi contoh paling jelas. Data pelatihan asli hanya 130 rekaman, tetapi full balancing meningkatkannya menjadi 56.000. Artinya, sekitar 99,8% observasi pada kelas tersebut merupakan data sintetis. Pada proporsi seperti ini, generator tidak lagi berperan sebagai pelengkap variasi, melainkan hampir menggantikan distribusi asli. Masalah kedua terjadi ketika sampel laten harus didekode kembali ke feature-space. Walaupun reconstruction error SDAE kecil, proses encode-generate-decode tetap dapat menggeser detail yang bersifat diskriminatif. Kombinasi over-amplification dan reconstruction error inilah yang mendorong perubahan desain menuju latent-space dan pembatasan rasio augmentasi.

Temuan ini sejalan dengan gagasan bahwa oversampling perlu mempertimbangkan wilayah keputusan dan kualitas sampel, bukan hanya rasio akhir kelas [6]. Pada konteks IDS, pembatasan

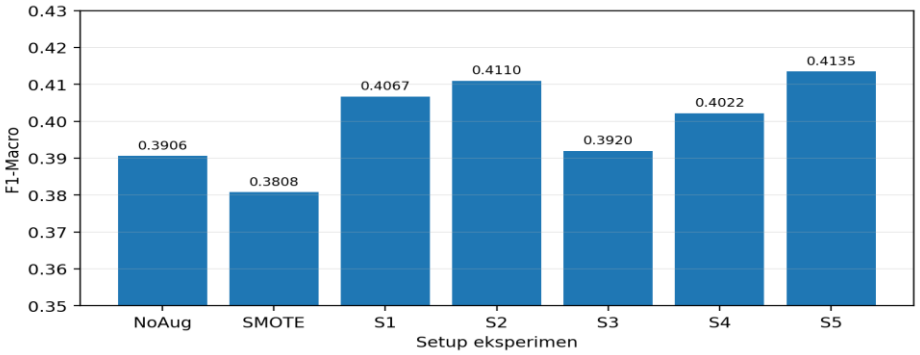
menjadi lebih penting karena data serangan langka sering kali mengandung pola yang heterogen. Menyalin atau membangkitkan terlalu banyak sampel dari basis data asli yang sangat kecil dapat memperkuat bias generator dan menghasilkan keyakinan palsu pada classifier.

Kinerja Klasifikasi pada Augmentasi Moderat

Setelah augmentasi dipindahkan ke ruang laten dan jumlah sampel sintetis dibatasi, empat dari lima skenario generatif (S1, S2, S4, dan S5) melampaui baseline tanpa augmentasi pada F1-Macro. S5-CTGAN memberikan F1-Macro tertinggi, yaitu 0,4135, dibanding baseline 0,3906. Peningkatan relatifnya 5,9%. Nilai MCC S5 sebesar 0,5897 juga sedikit lebih tinggi daripada baseline 0,5758. Walaupun kenaikan absolut F1-Macro tidak besar, konsistensinya penting karena masalah yang dievaluasi adalah klasifikasi sepuluh kelas dengan ketimpangan hingga 1:534,5.

Tabel 4. Kinerja Random Forest pada tujuh setup latent-space

Setup	Accuracy	F1-Macro	G-Mean	MCC
B1 NoAug	0,6712	0,3906	0,2476	0,5758
B2 SMOTE	0,6659	0,3808	0,3561	0,5723
S1 Hybrid	0,6838	0,4067	0,2734	0,5917
S2 FeatExt	0,6820	0,4110	0,3112	0,5897
S3 DimRed	0,6639	0,3920	0,3064	0,5659
S4 WGAN-GP	0,6822	0,4022	0,2099	0,5899
S5 CTGAN	0,6821	0,4135	0,3073	0,5897



Gambar 2. Perbandingan F1-Macro antarsetup pada ruang laten

S2-Feature Extraction berada sangat dekat dengan S5 pada F1-Macro (0,4110), sementara S1-Hybrid memperoleh MCC tertinggi sebesar 0,5917. Pola ini memperlihatkan bahwa tidak ada satu model yang dominan pada seluruh metrik. S5 lebih baik dalam keseimbangan F1 per kelas, sedangkan S1 sedikit lebih kuat pada korelasi keseluruhan prediksi. S3-Dimensional Reduction hanya mencapai F1-Macro 0,3920. Hasil tersebut mendukung dugaan bahwa bottleneck delapan dimensi terlalu sempit untuk mempertahankan seluruh informasi diskriminatif serangan.

Baseline SMOTE memperoleh F1-Macro 0,3808, lebih rendah daripada baseline tanpa augmentasi. Temuan ini tidak berarti SMOTE selalu buruk, melainkan menunjukkan bahwa interpolasi linier pada distribusi UNSW-NB15 yang skewed, heavy-tailed, dan heterogen belum cukup untuk kasus ini. Model generatif yang bekerja pada latent representation memiliki ruang untuk mempelajari pola nonlinier yang tidak dapat direpresentasikan hanya dari garis antara dua tetangga terdekat.

Dampak pada Tingkat Kelas

Performa agregat perlu dibaca bersama perubahan per kelas. Pada S5-CTGAN, kenaikan terbesar terjadi pada Analysis dan DoS. F1 Analysis meningkat dari 0,018 menjadi 0,105 (+0,087 atau sekitar 483% secara relatif), sedangkan DoS meningkat dari 0,206 menjadi 0,375 (+0,169 atau sekitar 82%). Fuzzers meningkat tipis. Sebaliknya, Shellcode dan Backdoor belum membaik, dan Worms hanya bergerak dari 0,077 menjadi 0,078. Hasil lengkap ditampilkan pada Tabel 5.

Tabel 5. Perbandingan F1 per kelas: baseline vs S5-CTGAN

Kelas	N uji	Baseline	S5-CTGAN	Delta F1
Worms	44	0,077	0,078	+0,001
Shellcode	378	0,411	0,393	-0,018
Backdoor	583	0,038	0,034	-0,004
Analysis	677	0,018	0,105	+0,087
Reconnaissance	3.496	0,546	0,544	-0,002
DoS	4.089	0,206	0,375	+0,169
Fuzzers	6.062	0,319	0,322	+0,003
Exploits	11.132	0,539	0,535	-0,004
Generic	18.871	0,975	0,975	0,000
Normal	37.000	0,776	0,772	-0,004

Peningkatan Analysis menunjukkan bahwa penambahan 8.000 sampel sintetis pada basis 2.000 data asli memberi tambahan variasi yang berguna untuk membentuk decision boundary. DoS tidak termasuk kelas yang diaugmentasi secara langsung, tetapi F1-nya ikut meningkat. Efek tidak langsung ini masuk akal pada klasifikasi multikelas: ketika representasi kelas tetangga menjadi lebih baik, batas antarserangan dapat berubah dan mengurangi sebagian kebingungan yang sebelumnya dialihkan ke DoS.

Kelas ultra-langka tetap menjadi batas utama metode. Pada data uji S5, recall Worms hanya 0,045 (2 dari 44 sampel benar) dan recall Backdoor 0,074. Sebagian besar Worms keliru sebagai Exploits, sedangkan Backdoor dan Analysis banyak tertukar dengan DoS. Menariknya, kesalahan tersebut lebih sering mengarah ke kelas serangan lain daripada Normal. Dari sudut pandang operasional, model masih menangkap sifat anomali pada sebagian trafik, namun belum cukup presisi untuk mengidentifikasi subkategori serangan. Hal ini penting agar hasil penelitian tidak diinterpretasikan seolah-olah augmentasi generatif telah menyelesaikan extreme imbalance sepenuhnya.

Kualitas Distribusi Data Sintetis

Kualitas data sintetis tidak dapat dinilai hanya dari kenaikan classifier. Data yang memberi skor klasifikasi tinggi tetapi menyimpang jauh dari distribusi asli dapat menimbulkan model yang rapuh. Karena itu, penelitian mengukur Wasserstein Distance antara distribusi data sintetis dan data nyata. S2-Feature Extraction menghasilkan WD mean terendah, yaitu 0,0172. Nilai S5-CTGAN sebesar 0,0193 dan S4-WGAN-GP 0,0195 masih berada pada rentang yang dekat, sedangkan S1-Hybrid 0,0197 dan S3-Dimensional Reduction 0,0214.

Visualisasi t-SNE memperlihatkan pola yang konsisten dengan hasil kuantitatif. Sampel sintetis pada skenario terbaik berada pada manifold yang sama dengan data asli dan membentuk overlap yang natural, bukan cluster terpisah. Pada sisi lain, overlap yang terlalu sempurna juga

tidak selalu diinginkan karena dapat menandakan memorization. Dalam hasil ini, sebaran sintetis tetap menunjukkan variasi di sekitar data nyata, sehingga lebih sesuai dengan fungsi augmentasi: menambah kemungkinan bentuk observasi tanpa melepaskan karakter kelas asal.

Perbedaan posisi S2 dan S5 memberikan pelajaran metodologis. S2 menghasilkan distribusi paling mirip secara statistik, tetapi S5 memberikan F1-Macro tertinggi. Artinya, kesamaan distribusi marginal bukan satu-satunya syarat keberhasilan classifier. CTGAN dirancang untuk menangani distribusi tabular yang multimodal serta kategori yang tidak seimbang [11]; kemampuan tersebut dapat menghasilkan variasi yang lebih berguna untuk decision boundary meskipun WD rata-ratanya sedikit lebih besar. Oleh sebab itu, evaluasi data sintetis sebaiknya menggabungkan fidelity dan utility, bukan memilih salah satu secara terpisah.

Validasi Statistik

Pengulangan sepuluh kali menunjukkan bahwa perbedaan antarsetup bukan sekadar variasi acak satu seed. Seluruh kelompok memenuhi normalitas Shapiro-Wilk dan uji Levene menghasilkan p = 0,577. One-way ANOVA memberikan F = 233,907 dengan p = 4,37 x 10^-41. Eta-squared 0,957 berada pada kategori efek besar. Uji Friedman juga signifikan (p = 1,48 x 10^-9), sehingga kesimpulan tidak bergantung pada asumsi parametrik saja.

Tabel 6. Ringkasan uji statistik inferensial

Uji	Statistik	Interpretasi
Shapiro-Wilk	W > 0,88; p > 0,05	Asumsi normalitas terpenuhi
Levene	0,795; p = 0,577	Varians homogen
One-way ANOVA	F = 233,907; p = 4,37 x 10^-41	Perbedaan antargroup signifikan
Effect size	eta-kuadrat = 0,957	Large effect
Friedman	chi-square = 52,500; p = 1,48 x 10^-9	Perbedaan signifikan

Uji lanjut Tukey HSD menunjukkan S1, S2, S4, dan S5 berbeda signifikan dan lebih baik daripada baseline No-Augmentation (p-adj < 0,05), sedangkan S3 tidak berbeda signifikan (p-adj = 0,2607). Dengan demikian, yang mendukung peningkatan bukan label 'GAN' semata. Representasi laten dan derajat kompresi ikut menentukan apakah data sintetis menjadi berguna. Reduksi ke delapan dimensi pada S3 tampaknya menghilangkan sebagian informasi yang diperlukan untuk membedakan pola serangan.

Implikasi terhadap Perancangan IDS

Hasil penelitian menggeser cara melihat balancing pada IDS. Distribusi yang secara numerik rapi tidak otomatis menghasilkan model yang lebih aman. Pada kelas yang sangat jarang, full balancing dapat menciptakan ilusi kelimpahan data: jumlah sampel meningkat, tetapi keragaman informasinya tetap dibatasi oleh sedikit observasi asli. Karena itu, rasio data sintetis terhadap data asli perlu diperlakukan sebagai hyperparameter yang memiliki risiko, bukan target administratif yang selalu harus dibuat 1:1.

Strategi moderat menawarkan kompromi yang lebih praktis. Pertama, biaya komputasi menurun karena dataset hasil augmentasi 202.282 sampel jauh di bawah skenario full balancing lebih dari 560.000 sampel. Kedua, data nyata tetap mempunyai porsi yang berarti dalam kelas minoritas. Ketiga, classifier bekerja langsung pada latent-space yang digunakan generator, sehingga tidak perlu menanggung distorsi tambahan dari decoding. Prinsip ini relevan untuk

implementasi IDS di lingkungan dengan sumber daya terbatas karena peningkatan tidak diperoleh dengan membesarkan dataset tanpa batas.

Walaupun S5-CTGAN menjadi konfigurasi terbaik pada penelitian ini, hasilnya tidak membenarkan pemilihan CTGAN secara otomatis pada seluruh kasus. Dataset lain dapat memiliki struktur berbeda, dan jenis serangan baru dapat mengubah distribusi. Penelitian terbaru juga menunjukkan bahwa GAN dapat membantu NIDS ketika sampel serangan terbatas, tetapi besarnya manfaat bergantung pada kualitas data hasil generasi dan skenario validasi [13]-[15]. Oleh karena itu, pendekatan yang lebih aman adalah menguji generator bersama batas amplifikasi dan metrik fidelity pada dataset yang menjadi sasaran implementasi.

Keterbatasan Penelitian

Penelitian memiliki beberapa keterbatasan. Pertama, seluruh eksperimen menggunakan satu benchmark utama sehingga generalisasi strategi min(10.000, 15 x n_k) belum dapat dianggap universal. Nilai batas tersebut merupakan keputusan empiris yang berhasil pada distribusi penelitian ini, bukan konstanta teoritis untuk semua IDS. Kedua, jumlah data asli pada Worms dan Backdoor tetap terlalu kecil bagi generator untuk mempelajari variasi kelas secara memadai. Augmentasi hanya dapat memperluas informasi yang tersedia; ia tidak dapat menciptakan pengetahuan yang sama sekali tidak pernah terlihat pada data asli.

Ketiga, evaluasi berfokus pada performa klasifikasi dan kualitas distribusi, belum pada latency inferensi, konsumsi memori, serta throughput trafik pada sistem real-time. Keempat, penelitian belum menguji robustness terhadap concept drift atau serangan yang belum pernah muncul. Pengembangan berikutnya dapat menggabungkan moderate augmentation dengan cost-sensitive learning, few-shot learning, ensemble adaptif, atau model generatif baru seperti diffusion model untuk data tabular. Evaluasi lintas dataset seperti CIC-IDS2017, CICIDS2018, TON_IoT, dan IoT-23 juga diperlukan agar batas amplifikasi dapat diuji pada karakter trafik yang berbeda.

SIMPULAN

Penelitian ini menunjukkan bahwa penanganan ketidakseimbangan kelas pada UNSW-NB15 tidak harus diarahkan pada distribusi yang sepenuhnya seimbang. Strategi augmentasi moderat yang membatasi target setiap kelas minoritas pada min(10.000, 15 x jumlah data asli), diterapkan pada ruang laten SDAE, memberikan hasil yang lebih stabil daripada pendekatan full balancing pada eksperimen awal. Dengan skema tersebut, jumlah data latih setelah augmentasi menjadi 202.282 sampel, sehingga data sintetis tidak mendominasi distribusi kelas secara berlebihan.

Pada evaluasi latent-space, S5-CTGAN memperoleh F1-Macro 0,4135 dan MCC 0,5897, sedangkan baseline tanpa augmentasi memperoleh F1-Macro 0,3906. Perbaikan paling terlihat pada Analysis dan DoS. S2-Feature Extraction memberikan fidelity distribusi terbaik dengan Wasserstein Distance 0,0172. ANOVA, Friedman, dan Tukey HSD mendukung adanya perbedaan performa yang signifikan, sementara S3 dengan latent dimension delapan tidak berbeda signifikan dari baseline. Temuan tersebut memperkuat bahwa keberhasilan augmentasi ditentukan bersama oleh kualitas representasi, arsitektur generator, dan rasio amplifikasi.

Pada saat yang sama, Worms dan Backdoor tetap sulit dideteksi. Hasil ini menegaskan batas yang perlu dijaga dalam penggunaan data sintetis: augmentasi dapat memperkaya representasi kelas langka, tetapi tidak menggantikan kebutuhan akan data asli yang beragam. Bagi pengembangan IDS, kontribusi utama penelitian bukan sekadar model generatif tertentu, melainkan prinsip bahwa augmentasi yang terkendali dan selektif lebih layak diprioritaskan dibanding penyeimbangan agresif yang mengejar kesamaan jumlah kelas.

PERNYATAAN DATA

UNSW-NB15 merupakan dataset publik yang dikembangkan oleh UNSW Canberra Cyber. Artikel ini menggunakan pembagian training dan testing yang telah tersedia pada dataset tersebut. Parameter utama strategi augmentasi, arsitektur SDAE, konfigurasi setup, dan metrik evaluasi dilaporkan agar eksperimen dapat direplikasi. Implementasi kode dapat disertakan sebagai materi tambahan pada saat artikel diajukan ke jurnal yang mensyaratkan open materials.

REFERENSI

[1] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in 2015 Military Communications and Information Systems Conference (MilCIS), 2015, pp. 1-6, doi: 10.1109/MilCIS.2015.7348942.

[2] N. Moustafa and J. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," Information Security Journal: A Global Perspective, vol. 25, no. 1-3, pp. 18-31, 2016, doi: 10.1080/19393555.2015.1125974.

[3] V. Shanmugam, R. Razavi-Far, and E. Hallaji, "Addressing Class Imbalance in Intrusion Detection: A Comprehensive Evaluation of Machine Learning Approaches," Electronics, vol. 14, no. 1, art. 69, 2025, doi: 10.3390/electronics14010069.

[4] C. Wheelus, E. Bou-Harb, and X. Zhu, "Tackling Class Imbalance in Cyber Security Datasets," in 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018, pp. 229-232, doi: 10.1109/IRI.2018.00041.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.

[6] P. Soltanzadeh and M. Hashemzadeh, "RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem," Information Sciences, vol. 542, pp. 92-111, 2021, doi: 10.1016/j.ins.2020.07.014.

[7] I. Goodfellow et al., "Generative Adversarial Nets," in Advances in Neural Information Processing Systems 27, 2014, pp. 2672-2680.

[8] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," Journal of Machine Learning Research, vol. 11, pp. 3371-3408, 2010.

[9] J. H. Lee and K. H. Park, "AE-CGAN Model based High Performance Network Intrusion Detection System," Applied Sciences, vol. 9, no. 20, art. 4221, 2019, doi: 10.3390/app9204221.

[10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved Training of Wasserstein GANs," in Advances in Neural Information Processing Systems 30, 2017.

[11] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular Data using Conditional GAN," in Advances in Neural Information Processing Systems 32, 2019.

[12] D. Li, D. Kotani, and Y. Okabe, "Improving Attack Detection Performance in NIDS Using GAN," in 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), 2020, pp. 817-825, doi: 10.1109/COMPSAC48688.2020.0-162.

[13] C. Park, J. Lee, Y. Kim, J.-G. Park, H. Kim, and D. Hong, "An Enhanced AI-Based Network Intrusion Detection System Using Generative Adversarial Networks," IEEE Internet of Things Journal, vol. 10, no. 3, pp. 2330-2345, 2023, doi: 10.1109/JIOT.2022.3211346.

[14] H. Ding, Y. Sun, N. Huang, Z. Shen, and X. Cui, "TMG-GAN: Generative Adversarial Networks-Based Imbalanced Learning for Network Intrusion Detection," IEEE Transactions on

Information Forensics and Security, vol. 19, pp. 1156-1167, 2024, doi: 10.1109/TIFS.2023.3331240.

[15] X. Zhao, K. W. Fok, and V. L. L. Thing, "Enhancing network intrusion detection performance using generative adversarial networks," Computers & Security, art. 104005, 2024, doi: 10.1016/j.cose.2024.104005.

[16] J. Cui et al., "A novel multi-module integrated intrusion detection system for high-dimensional imbalanced data," Applied Intelligence, vol. 53, pp. 272-288, 2023, doi: 10.1007/s10489-022-03361-2.

[17] S. Bourou, A. El Saer, T.-H. Velivassaki, A. Voulkidis, and T. Zahariadis, "A Review of Tabular Data Synthesis Using GANs on an IDS Dataset," Information, vol. 12, no. 9, art. 375, 2021, doi: 10.3390/info12090375.

[18] W. Tian, Y. Shen, N. Guo, J. Yuan, and Y. Yang, "VAE-WACGAN: An Improved Data Augmentation Method Based on VAEGAN for Intrusion Detection," Sensors, vol. 24, art. 6035, 2024, doi: 10.3390/s24186035.