

Analisis Perbandingan Data Sintesis Privacy-Preserving Menggunakan CTGAN dan Metode Differential Privacy untuk Pelatihan Model LSTM dalam Klasifikasi Deteksi Jatuh pada Lansia

Oliver Servatio Leon
Fakultas Informatika
Universitas Telkom
Bandung
oliverleon@students.telkomuniversity.ac.id

Aji Gautama Putrada
Fakultas Informatika
Universitas Telkom
Bandung
ajigps@telkomuniversity.ac.id

Sidik Prabowo
Fakultas Informatika
Universitas Telkom
Bandung
pawkowo@telkomuniversity.ac.id

Abstrak

Jatuh pada lansia menjadi salah satu penyebab masalah cedera fatal dan kematian di dunia. Pertumbuhan populasi lansia yang signifikan menyebabkan tingginya angka prevalensi jatuh dan mendorong kebutuhan solusi deteksi dini jatuh pada lansia. Pengembangan *Fall Detection Systems* (FDS) menjadi jawaban karena FDS dapat menawarkan respons yang cepat ketika kejadian jatuh terjadi. Namun, pengembangan FDS memiliki tantangan berupa keterbatasan data, ketidakseimbangan kelas, dan risiko pelanggaran privasi.

Penelitian ini bertujuan untuk membandingkan dua metode privacy-preserving, yaitu model CTGAN dan DP. Penelitian diawali dengan implementasi FDS berbasis *Internet of Things* (IoT) dan mengumpulkan data aktivitas lansia. Lalu, melakukan pengukuran utilitas statistik dengan Kolmogorov-Smirnov (KS) complement dan Wasserstein distance, pengukuran aspek privasi dengan Nearest Neighbor Distance Ratio (NNDR), dan pengukuran utilitas model dengan skenario *Train-on-Synthetic, Test-on-Real* (TSTR). Hasil menunjukkan bahwa CTGAN menjadi metode privacy-preserving yang lebih baik daripada DP, pada utilitas statistik, CTGAN mengungguli DP dengan nilai KS complement di 0.902 dan Wasserstein distance di 0.329, dibandingkan dengan DP di 0.861 dan 0.690. Pada segi privasi, CTGAN mengungguli DP di persentil ke-5 dengan nilai 0.63 dibandingkan dengan 0.57 pada DP. Terakhir, dari segi utilitas model, CTGAN mengungguli DP dengan nilai akurasi dan *F1-score* yang lebih tinggi dari DP.

Kata kunci: conditional tabular generative adversarial network, differential privacy, fall detection, long short-term memory, privacy-preserving

Abstract

Falls among the elderly are a leading cause of fatal injuries and deaths worldwide. The significant growth of the elderly population has led to a high prevalence of falls and driven the need for early fall detection solutions for the elderly. The development of Fall Detection Systems (FDS) offers a solution, as FDS can provide a rapid response when a fall occurs. However, the development of FDS faces challenges such as data limitations, class imbalance, and privacy violation risks.

This study aims to compare two privacy-preserving methods: the CTGAN model and DP. The research begins with the implementation of an Internet of Things (IoT)-based FDS and the collection of elderly activity data. Next, statistical utility was measured using the Kolmogorov-Smirnov (KS) complement and Wasserstein distance, privacy aspects were assessed using the Nearest Neighbor Distance Ratio (NNDR), and model utility was evaluated using the Train-on-Synthetic, Test-on-Real (TSTR) scenario. The results show that CTGAN is a better privacy-preserving method than DP; in terms of statistical utility, CTGAN outperforms DP with a KS complement value of 0.902 and a Wasserstein distance of 0.329, compared to DP's 0.861 and 0.690. In terms of privacy, CTGAN outperforms DP at the 5th percentile with a value of 0.63 compared to 0.57 for DP. Finally, in terms of model utility, CTGAN outperforms DP with higher accuracy and *F1-score* values than DP.

Keywords: conditional tabular generative adversarial network, differential privacy, fall detection, long short-term memory, privacy-preserving

1. Pendahuluan

Latar Belakang

Populasi lansia di seluruh dunia mengalami peningkatan secara signifikan, Populasi manusia dengan usia lebih dari 60 tahun diperkirakan mencapai 1.4 miliar jiwa di tahun 2030, meningkat 1 miliar jiwa sejak tahun 2020 [1]. Jatuh merupakan masalah kesehatan serius yang menjadi salah satu penyebab utama cedera bagi lansia, mengakibatkan lebih dari 684,000 kematian di dunia setiap tahun [2]. Pengembangan teknologi *Fall Detection Systems* (FDS) berbasis wearable sensors dan *deep learning* (DL) menjadi fokus penelitian dalam beberapa tahun terakhir sebagai respons terhadap masalah jatuh [3]. Model Long Short-Term Memory (LSTM) dapat menangkap pola temporal yang kompleks dari data sensor. Penelitian oleh Butt *et al.* [4] menerapkan model LSTM pada dataset SisFall dan mendapatkan akurasi 88% pada data latih. Meskipun sistem deteksi jatuh ini menunjukkan kinerja positif, terdapat tantangan dalam pengembangannya. Pengumpulan data aktivitas lansia memiliki kendala pada keterbatasan jumlah data, terutama pada aktivitas jatuh yang jarang terjadi [5]. Selain itu, data sensor wearable mengandung informasi sensitif yang menimbulkan tantangan pada aspek privasi dan etika [6].

Penelitian terdahulu menggunakan metode *Differential Privacy* (DP) untuk melindungi data dengan menambah noise statistik pada data asli. Namun, metode ini memiliki keterbatasan pada *trade-off* antara privasi dan utilitas. Penelitian oleh Ouadrhiri *et al.* [7] mengungkapkan bahwa penerapan privasi yang ketat sering kali dapat menurunkan akurasi data. Oleh karena itu, dibutuhkan metode yang dapat menyediakan data pelatihan yang representatif untuk model deteksi jatuh, tetapi juga mampu memastikan perlindungan privasi individu.

Model generatif Conditional Tabular Generative Adversarial Network (CTGAN) merupakan salah satu alternatif implementasi metode *privacy-preserving*. CTGAN dapat mengatasi ketidakseimbangan kelas dan mempertahankan distribusi statistik, sehingga data sintesis yang dibuat dapat merepresentasikan data asli [8]. Dalam konteks privasi, data sintesis CTGAN memungkinkan pelatihan model tanpa mengungkap informasi individu dari dataset asli [9]. Beberapa penelitian di bidang kesehatan menunjukkan bahwa CTGAN memungkinkan penambahan *dataset*

dan menunjukkan kenaikan performa yang signifikan pada model *machine learning*.

Topik dan Batasannya

Penelitian ini membahas tentang penyelesaian permasalahan privasi menggunakan CTGAN dengan mengevaluasi *trade-off* antara utilitas data dan privasi dari hasil gneerasi data sintesis CTGAN sebagai metode *privacy-preserving*. Penelitian ini juga akan membandingkan performa pelatihan LSTM yang dilatih dengan data sintesis CTGAN dengan data asli yang diberikan oleh *noise DP*.

Untuk menjaga ruang lingkup penelitian, pembahasan penelitian ini dibatasi oleh beberapa aspek. Pertama, *dataset* yang digunakan untuk pembuatan data sintesis merupakan *dataset* yang diperoleh dari hasil pengambilan data dengan sistem FDS berbasis IoT. Kedua, model generatif data sintesis terbatas pada CTGAN, dengan alternatif penambahan *noise DP* pada data asli. Ketiga, model deteksi jatuh terbatas pada arsitektur LSTM. Keempat, evaluasi utilitas model diukur berdasarkan akurasi dan *F1-score*. Dan terakhir, evaluasi privasi diukur menggunakan metrik *nearest-neighbor-distance-ratio* (NNDR).

Tujuan

Berdasarkan permasalahan yang diangkat berdasarkan topik penelitian, penelitian ini ditujukan untuk menjembatani celah penelitian dengan mengevaluasi *trade-off* antara utilitas data dan privasi dari penggunaan data sintesis CTGAN sebagai metode *privacy-preserving*. Penelitian ini juga membandingkan performa model pada data sintesis CTGAN dan *dataset* asli dengan *noise DP*, sehingga dapat memberikan kontribusi dalam pengembangan sistem deteksi jatuh yang efektif dan mampu memastikan perlindungan privasi individu. Tabel 1 menunjukkan keterkaitan antara tujuan, pengujian, dan kesimpulan dari penelitian ini.

Tabel 1. Keterkaitan antara tujuan, pengujian dan kesimpulan

No	Tujuan	Pengujian	Kesimpulan
1	Menganalisis hasil data sintesis dengan CTGAN yang dapat merepresentasikan pola dan distribusi statistik data	Mengukur dan Wasserstein distance	CTGAN dengan DP

	yang dapat menjaga privasi data asli		
2	Menganalisis perbandingan performa LSTM yang dilatih dengan data sintesis CTGAN dengan model yang dilatih dengan DP	Membandingkan akurasi dan $F1$ - $score$	CTGAN memiliki akurasi dan $F1$ - $score$ yang lebih tinggi dibandingkan dengan DP
3	Mengevaluasi <i>trade-off</i> antara utilitas dan privasi data pada penggunaan data sintesis CTGAN dan data asli yang dilindungi oleh noise DP melalui metrik evaluasi	Mengukur nilai privasi pada CTGAN dan DP dengan metrik NNDR	CTGAN memiliki nilai NNDR yang lebih baik pada persentil kelima dibandingkan dengan DP

Organisasi Tulisan

Jurnal ini dibagi menjadi 5 bagian utama, dengan bagian selanjutnya akan membahas topik berikut: Bagian 2 membahas terkait penelitian terkait yang relevan pada penelitian ini. Bagian 3 membahas tentang metode penelitian dan sistem yang dibangun. Bagian 4 membahas tentang hasil penelitian. Terakhir, Bagian 5 akan membahas terkait kesimpulan dan jawaban atas penelitian ini.

2. Studi Terkait

Penelitian terdahulu banyak mengembangkan teknologi FDS berbasis sensor akselerometer dan giroskop, serta *deep learning*. Penelitian oleh Butt *et al.* [4] membandingkan struktur model *deep learning* LSTM dengan model CNN berbasis *transfer learning* untuk klasifikasi jatuh dengan fokus utama evaluasi performansi model. Selain itu, penelitian oleh Zhang *et al* [10] menggunakan pendekatan CNN untuk mengatasi masalah ketidakseimbangan kelas pada *dataset* jatuh dengan menyesuaikan *threshold* prediksi berdasarkan rasio ketidakseimbangan kelas untuk meningkatkan sensitivitas terhadap kelas minoritas.

Penelitian terdahulu sudah menggunakan CTGAN untuk generasi data sintesis. Penelitian oleh Alabdulwahab *et al.* [11] menunjukkan efektivitas dari CTGAN untuk generasi data pada sensor IoT. Selain itu, penelitian oleh Parise *et al.* [12] juga menggunakan CTGAN untuk generasi data sensitif pasien yang bisa

mendukung pendekatan *privacy-preserving*. Pengujian Kolmogorov-Smirnov (KS) dan skema TSTR untuk menguji kualitas data sintesis dapat membuktikan kemampuan CTGAN untuk membuat data yang dapat menyerupai data asli secara statistik.

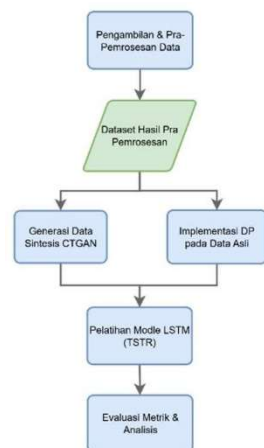
Beberapa penelitian telah menggunakan metrik NNDR sebelumnya, Park *et al* [13] menggunakan NNDR pada computer vision untuk membandingkan data hasil prediksi dengan data asli. Metrik ini sangat penting untuk membedakan inlier dan outlier pada machine learning. Penelitian oleh Lee *et al.* [14] menggunakan NNDR sebagai deskriptor MMPEG. Metode ini digunakan untuk seleksi fitur pada penelitian mereka. Dalam kasus ini, terdapat celah penelitian untuk membahas pentingnya pengukuran nilai privasi pada data sintesis CTGAN dengan NNDR untuk mencegah kebocoran privasi. Tabel 2 menunjukkan daftar kontribusi penelitian terkait, dengan relevansinya terhadap penelitian.

Tabel 2. Perbandingan kontribusi penelitian terkait

Referensi	Fall Detection	CTGAN	NNDR
Butt <i>et al.</i> [4]	✓	✗	✗
Zhang <i>et al.</i> [10]	✓	✓	✗
Alabdulwahab <i>et al.</i> [11]	✗	✓	✗
Parise <i>et al.</i> [12]	✗	✓	✗
Park <i>et al.</i> [13]	✗	✗	✓
Lee <i>et al.</i> [14]	✗	✗	✓
Metode yang Diusulkan	✓	✓	✓

3. Sistem yang Dibangun

Penelitian ini mencakup empat tahapan utama dalam pengerjaannya. Tahapan pertama dimulai dari melakukan pengambilan dan pra-pemrosesan data dengan sensor akselerometer dan giroskop untuk skenario kegiatan jatuh dan tidak jatuh, dilanjutkan dengan pembersihan data hingga data siap digunakan sebagai *input* model. Tahap kedua mencakup generasi data sintesis dengan CTGAN untuk pengujian *privacy-preserving*. Tahap ketiga adalah menerapkan DP pada data asli yang akan digunakan sebagai pembandingan metode CTGAN. Terakhir, melatih *dataset* dengan model LSTM dan dilanjutkan dengan evaluasi metrik kinerja model, metrik utilitas, dan metrik privasi. Visualisasi alur tahapan penelitian dapat dilihat pada gambar 1.



Gambar 1. Diagram Alur Tahapan Penelitian

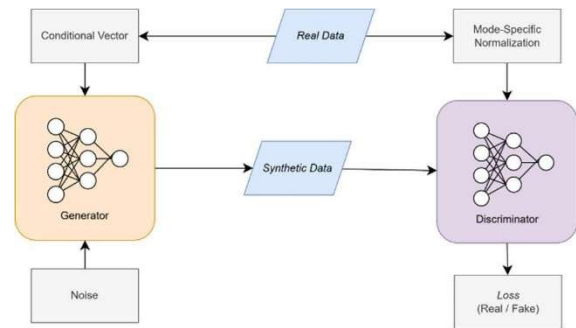
3.1 FDS IoT untuk Lansia

Data aktivitas sehari-hari dan data jatuh diambil menggunakan alat berbasis *Internet of Things* (IoT). Sistem ini menggunakan sensor giroskop & akselerometer MPU6050 yang terhubung dengan mikrokontroler ESP32 untuk merekam gerak tubuh pengguna. Data hasil rekaman sistem lalu dikirim dengan protokol *Message Queuing Telemetry Transport* (MQTT) melalui broker MQTT sebagai perantara antara komputer dan *end device* [15].

Data yang dikumpulkan berupa simulasi kejadian sehari-hari, seperti berjalan, duduk, dan berlari, serta kejadian jatuh, seperti jatuh ke depan, ke samping, dan ke belakang [16]. Data lalu disimpan ke dalam bentuk CSV dengan setiap baris dataset menghasilkan 6 kolom fitur sinyal percepatan dan kecepatan sudut pada tiga sumbu, kemudian diberi label “jatuh” dan “tidak jatuh” berdasarkan skenario kejadian yang dicatat. Selanjutnya, data digunakan untuk generasi data sintesis dan melatih model LSTM untuk evaluasi.

3.2 Generasi Data dengan CTGAN

Penelitian ini menggunakan arsitektur CTGAN sebagai metode generasi data. CTGAN merupakan model generatif berdasarkan *Generative Adversarial Network* (GAN) yang mengatasi ketidakseimbangan data pada distribusi *non-Gaussian* dan multimodal [17]. Arsitektur CTGAN terdiri dari 2 jaringan utama, generator dan discriminator yang masing-masing bekerja secara adversarial. Gambar 2 menampilkan ilustrasi arsitektur CTGAN yang digunakan pada penelitian.

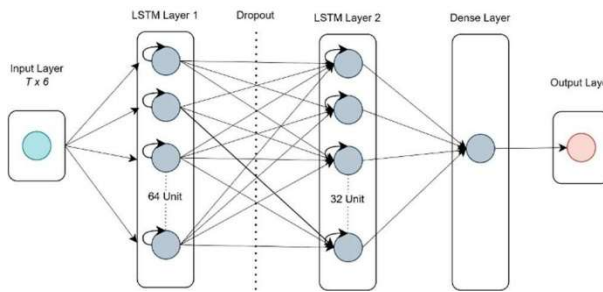


Gambar 2. Arsitektur CTGAN

Sebelum pelatihan, data multimodal seperti data sensor akselerometer dan giroskop dinormalisasi oleh CTGAN dengan metode *Mode-Specific Normalization* [18]. *Generator* menerima input vektor noise acak dan vektor kondisi kelas, lalu menghasilkan sampel data sintetik yang menyerupai distribusi fitur pada data asli. *Discriminator* mengambil sampel data asli dan sintetik, serta informasi kondisi sebagai masukan, kemudian *discriminator* belajar untuk membedakan keduanya dan menghasilkan nilai loss untuk memperbarui kedua jaringan secara adversarial. Penelitian ini juga menambahkan DP pada data asli sebagai *dataset* pembandingan untuk CTGAN. Pada tahap ini, *noise* DP ditambahkan berdasarkan fitur numerik yang dikalibrasi dengan parameter privasi ϵ [19].

3.3 Model LSTM

Model LSTM digunakan untuk menerima input data sekuensial akselerometer dan giroskop dari hasil pra-pemrosesan. Model dibangun menggunakan dua lapisan LSTM berurutan, dengan lapisan pertama sebagai ekstraksi fitur temporal dari setiap langkah waktu dan lapisan kedua digunakan untuk memproses fitur tersebut untuk menangkap pola temporal. Setiap lapisan LSTM diikuti oleh lapisan *dropout* untuk mengurangi *overfitting*, dan lapisan *fully connected* dengan satu neuron keluaran merepresentasikan probabilitas kejadian jatuh. Gambar 3 menampilkan ilustrasi arsitektur jaringan LSTM yang digunakan.



Gambar 3. Arsitektur LSTM

Proses pelatihan model LSTM mengikuti skenario *Train on Synthetic, Test on Real* [20]. Selain itu, pelatihan model dengan skenario *Train on Real, Test on Real* (TRTR) juga diaplikasikan sebagai metode pembandingan utama. Pada setiap skenario, *dataset* dibagi menjadi tiga bagian, yaitu *train*, *test*, dan *validation*. Data *train* digunakan untuk pelatihan model, *validation* digunakan untuk *early stopping* dan *hyperparameter tuning*, dan *test* digunakan untuk evaluasi performa pada hasil model generasi.

3.4 Formulasi Metrik Evaluasi

Penelitian ini menggunakan beberapa metrik evaluasi, seperti utilitas statistik, utilitas model, dan privasi untuk mengevaluasi perbandingan kedua metode *privacy-preserving*. Evaluasi ini bertujuan untuk menguji *trade-off* antara utilitas dan privasi dari kedua metode.

Penilaian kualitas data sintesis terhadap data asli dapat diukur dengan melihat kesesuaian distribusi fitur antara kedua data. Uji Kolmogorov-Smirnov (KS) komplement dan jarak Wasserstein (*Wasserstein distance*) dapat digunakan untuk menilai kesesuaian distribusi fitur tersebut.

Berdasarkan penelitian Fabuyi *et al.* [21] *KS complement* diukur berdasarkan nilai KS statistik untuk data kontinu. KS Statistik mengukur jarak maksimum antar fungsi distribusi kumulatif dari data asli (F_{real}) dan data sintesis (F_{synth}) pada setiap fitur kontinu.

$$KS = \sup |F_{real}(x) - F_{synthetic}(x)| \quad (1)$$

Nilai KS berada pada rentang $[0,1]$, dengan nilai yang kecil menandakan nilai yang semakin mirip distribusinya. *KS complement* adalah nilai komplement dari KS statistik dengan persamaan sebagai berikut:

$$KS_{comp} = 1 - KS \quad (2)$$

Semakin besar nilai *KS complement* menandakan semakin mirip distribusi data sintesis dengan data asli.

Nilai *Wasserstein distance* mengukur usaha yang diperlukan untuk mengubah distribusi dari satu sampel ke sampel lain [22]. Penelitian ini mengadopsi model 1-*Wasserstein distance* dengan persamaan sebagai berikut:

$$W_1(\mu, \nu) = \sup_{hLip(f) \leq 1} \int f d\mu - \int f d\nu \quad (3)$$

Di mana μ dan ν merepresentasikan probabilitas distribusi dari data sintetis dan data asli, dan f adalah fungsi dengan konstanta Lipschitz, dengan nilai *Wasserstein distance* yang kecil menandakan kemiripan distribusi yang besar.

Untuk menguji nilai privasi, penelitian ini menggunakan pendekatan metrik *Nearest-Neighbor Distance Ratio* (NNDR). NNDR mengukur jarak sampel pada fitur [23]. Nilai NNDR berada pada rentang $[0,1]$, dengan nilai yang semakin tinggi menandakan nilai privasi yang besar. Penelitian ini menggunakan nilai NNDR pada persentil ke-5 pada nilai distribusi NNDR untuk mengukur potensi kebocoran data pada skenario terburuk. Penelitian ini juga mengevaluasi utilitas model dengan *accuracy* dan *F1-score* pada model LSTM.

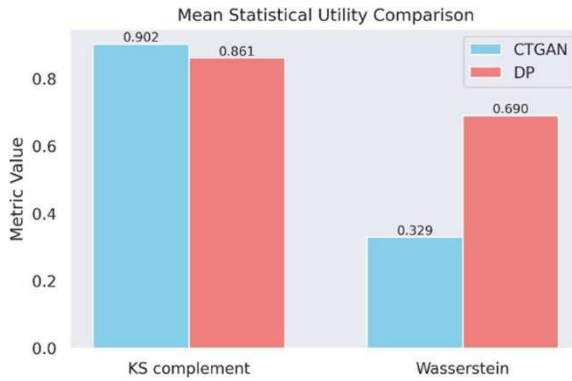
4. Evaluasi

4.1 Hasil Pengujian

Penelitian dilakukan dengan pengambilan data pada FDS menggunakan sistem IoT dengan sensor MPU6050. Pengambilan data meliputi pencatatan aktivitas sehari-hari seperti berjalan, duduk, dan berdiri, serta pencatatan kejadian jatuh. Data disimpan ke dalam bentuk csv dengan nama *datasensor_gabungan.csv*. Setiap baris data memiliki enam fitur yang berisi percepatan dan derajat kemiringan pada masing-masing tiga sudut x, y dan z. Setiap baris juga berisi 1 kolom label yang memiliki dua jenis label, 0 untuk “tidak jatuh” dan 1 untuk “jatuh”, dengan masing-masing memiliki 8,751 dan 3,759 data dan total 12,510 baris data. Setelah pengumpulan data, Tahapan selanjutnya adalah membagi data menjadi 3 bagian, 80% untuk training, 10% untuk testing, dan 10% untuk validation. Penelitian ini juga melakukan preproceasing dalam bentuk scaling menggunakan robust scaling untuk menghilangkan pencilaan pada data.

Penelitian ini menguji nilai utilitas statistik untuk melihat kemiripan nilai distribusi antara

kedua metode data sintesis dengan data asli. Untuk menguji nilai utilitas statistik, digunakan dua metrik, yaitu *KS complement* dan *Wasserstein distance*. Gambar menunjukkan rata-rata perbandingan nilai *KS complement* dan *Wasserstein distance* pada kedua metode data sintesis.



Gambar 4. Nilai mean KS complement dan Wasserstein

Berdasarkan gambar 5, terlihat bahwa data sintesis CTGAN unggul dibandingkan dengan DP dari segi utilitas statistik, hal ini dapat dilihat dari rata-rata nilai *KS complement* CTGAN di 0.902 yang lebih tinggi jika dibandingkan dengan DP di 0.861. Selain itu, nilai rata-rata *Wasserstein distance* pada CTGAN juga jauh lebih rendah di 0.329 dibandingkan dengan DP di 0.690 yang menandakan perubahan distribusi statistik yang cukup besar pada nilai DP dibanding dengan CTGAN.

Penelitian ini menggunakan perbandingan *F1-score* dan *accuracy* untuk melakukan uji utilitas model dengan pendekatan skenario TSTR. Selain itu, pelatihan dengan skenario *Train on Real, Test on Real* (TRTR) sebagai pembandingan baseline. Tabel 3 menunjukkan hasil dari skenario tersebut.

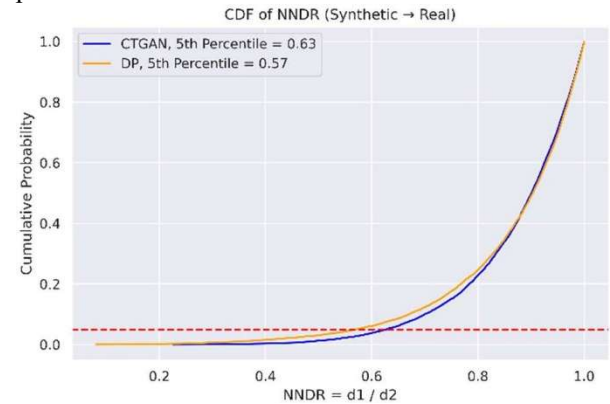
Tabel 3. Accuracy dan F1-score antara CTGAN, DP, dan baseline

Skenario	Accuracy	F1 (Macro)	F1 (Weighted)
TSTR DP	0.87	0.83	0.86
TSTR CTGAN	0.89	0.86	0.87
TRTR (baseline)	0.92	0.90	0.92

Tabel 3 menunjukkan hasil pelatihan model LSTM pada tiga skenario, TSTR dengan DP, TSTR CTGAN, dan TRTR sebagai *baseline*. Kedua skenario TSTR menunjukkan penurunan pada akurasi di 0.87 dan 0.89 untuk DP dan CTGAN, dibandingkan dengan 0.92 pada

baseline dengan data asli. Namun hasil akurasi pada CTGAN lebih mendekati dibanding dengan TSTR DP dari segi akurasi. Penurunan performa TSTR DP juga terlihat pada hasil *F1-score*, terutama di *macro F1-score*, di mana TSTR DP memperoleh hasil 0.83 dibandingkan dengan 0.86 pada TSTR CTGAN.

Penelitian ini menggunakan nilai NNDR di persentil ke-5 untuk menunjukkan hasil pada skenario terburuk (worst case). Gambar 5 menunjukkan grafik hasil nilai NNDR pada persentil ke-5.



Gambar 5. Nilai NNDR CTGAN dan DP pada persentil ke-5

Gambar 4.2 menunjukkan bahwa nilai NNDR pada persentil ke-5 di CTGAN memiliki nilai yang lebih tinggi di 0.63 dibandingkan dengan DP di 0.57. Hal ini menunjukkan bahwa di skenario terburuk, CTGAN tetap lebih unggul dalam segi privasi dibanding dengan DP.

4.2 Analisis Hasil Pengujian

Untuk menilai kualitas data sintesis CTGAN sebagai metode *privacy-preserving*, penelitian ini menggunakan tiga aspek penilaian, yaitu utilitas statistik data dengan uji *KS complement* dan *Wasserstein distance*, utilitas model dengan skenario TSTR, serta utilitas privasi dengan NNDR. Hasil menunjukkan bahwa penggunaan data sintesis CTGAN lebih baik dari segi privasi untuk FDS daripada DP.

Dari segi utilitas statistik, CTGAN menunjukkan hasil yang lebih baik dari DP, ditunjukkan dengan nilai *KS complement* yang lebih tinggi di 0.902 dibandingkan dengan 0.861 DP. Hal ini sejalan dengan penggunaan uji Kolmogorov–Smirnov sebagai metrik kesamaan distribusi berdasarkan penelitian oleh Fabuyi *et al* [21]. Kemiripan distribusi data juga dipertegas pada nilai *Wasserstein distance* yang menunjukkan perbedaan yang cukup signifikan

dengan nilai CTGAN di 0.329 dibanding DP di 0.690, sejalan dengan penelitian oleh He *et al.* [22] yang menyatakan bahwa semakin kecil nilai *Wasserstein distance*, distribusi data sintesis semakin mendekati distribusi data asli.

Pada uji utilitas model, kedua metode data sintesis mengalami penurunan kualitas jika dibandingkan dengan data asli. Namun, hasil TSTR kedua metode menunjukkan bahwa CTGAN memiliki trade-off yang lebih baik untuk utilitas dan privasi jika dibanding dengan DP. Hal ini dibuktikan dengan nilai accuracy yang lebih baik pada CTGAN di 89% dibandingkan dengan DP di 87%. Selain itu, penurunan kualitas pada DP terlihat pada nilai *F1-score*, khususnya pada *macro F1*, di mana DP memiliki skor 83% dibandingkan dengan CTGAN di 86%. Hal ini menandakan penurunan performa pada DP di bagian klasifikasi kelas minoritas (jatuh) yang signifikan dibandingkan dengan DP.

Pada aspek privasi, hasil nilai NNDR menunjukkan bahwa CTGAN juga mengungguli DP dengan nilai NNDR 0.63 pada persentil ke-5 (p5), dibandingkan dengan DP dengan nilai 0.57 pada p5. P5 digunakan mengukur 5% data dengan nilai NNDR terkecil. Hal ini sesuai dengan pendekatan pada penelitian oleh Liu *et al.* [23] yang menggunakan p5 sebagai estimasi robust terhadap risiko kebocoran. Hal ini menunjukkan bahwa, di skenario terburuk 5% nilai NNDR terkecil, CTGAN tetap lebih baik dibandingkan dengan DP.

5. Kesimpulan

Penelitian tugas akhir ini berhasil mencapai tujuan melalui perbandingan metode generasi CTGAN dan DP. Kualitas data sintesis CTGAN dapat merepresentasikan pola dan distribusi dari data FDS, yang ditunjukkan dengan nilai *KS complement* sebesar 0.902 dan *Wasserstein distance* sebesar 0.329. Meskipun CTGAN mengalami penurunan akurasi dan *F1-score* dibandingkan dengan data asli, performa model LSTM pada CTGAN tetap lebih unggul jika dibandingkan dengan DP dengan akurasi mencapai 89% dan *macro F1-score* mencapai 86%, dibandingkan dengan DP yang hanya mencapai akurasi 87% dengan penurunan signifikan pada *macro F1-score* di 83% akibat ketidakmampuan DP dalam klasifikasi kelas minoritas.

Evaluasi trade-off antara utilitas dan privasi CTGAN menunjukkan bahwa CTGAN dapat menjadi solusi yang lebih optimal. Dari segi privasi, CTGAN dapat mengungguli DP dengan nilai NNDR pada p5 yang lebih besar di 0.63 dibandingkan dengan 0.57 pada DP, yang berarti CTGAN lebih kuat dalam tingkat perlindungan, bahkan dalam skenario terburuk. Dapat disimpulkan bahwa CTGAN merupakan metode *privacy-preserving* yang lebih seimbang dan dapat diaplikasikan pada IoT untuk FDS pada lansia. Sebagai saran penelitian, penelitian selanjutnya dapat difokuskan pada penerapan TSTR pada *edge computing* dan penerapan pendekatan yang diusulkan pada *dataset* yang berbeda untuk meningkatkan kemampuan generalisasi hasil.

REFERENSI

- [1] R. Patel dan J. E. Gallagher, "Healthy ageing and oral health: priority, policy and public health," 1 Desember 2024, *Springer Nature*. doi: 10.1038/s41405-024-00262-z.
- [2] Y. Chen *dkk.*, "Global, regional, and national burden of falls among older adults: findings from the Global Burden of Disease Study 2021 and Projections to 2040," *npj Aging*, vol. 11, no. 1, Des 2025, doi: 10.1038/s41514-025-00275-4.
- [3] Z. Mohammad, A. R. Anwary, M. F. Mridha, M. S. H. Shovon, dan M. Vassallo, "An Enhanced Ensemble Deep Neural Network Approach for Elderly Fall Detection System Based on Wearable Sensors," *Sensors*, vol. 23, no. 10, Mei 2023, doi: 10.3390/s23104774.
- [4] A. Butt, S. Narejo, M. R. Anjum, M. U. Yonus, M. Memon, dan A. A. Samejo, "Fall Detection Using LSTM and Transfer Learning," *Wirel. Pers. Commun.*, vol. 126, no. 2, hlm. 1733–1750, Sep 2022, doi: 10.1007/s11277-022-09819-3.
- [5] N. Verma, S. Mundody, dan R. M. R. Guddeti, "An Efficient AI and IoT Enabled System for Human Activity Monitoring and Fall Detection," dalam *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICCCNT61001.2024.10725908.
- [6] A. I. K. Kalupahana, A. N. Balaji, X. Xiao,

- dan L. S. Peh, "SeRaNDiP - Leveraging Inherent Sensor Random Noise for Differential Privacy Preservation in Wearable Community Sensing Applications," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 2, Jun 2023, doi: 10.1145/3596252.
- [7] A. El Ouadrhiri dan A. Abdelhadi, "Differential Privacy for Deep and Federated Learning: A Survey," *IEEE Access*, vol. 10, hlm. 22359–22380, 2022, doi: 10.1109/ACCESS.2022.3151670.
- [8] A. Kiran dan S. S. Kumar, "A Methodology and an Empirical Analysis to Determine the Most Suitable Synthetic Data Generator," *IEEE Access*, vol. 12, hlm. 12209–12228, 2024, doi: 10.1109/ACCESS.2024.3354277.
- [9] E. Callahan dan L. MacNeill, "Synthetic Tabular Data Generation for Privacy-Preserving Machine Learning. Journal of Computer Science and Software Applications," *Journal of Computer Science and Software Applications*, vol. 5, no. 7, 2025.
- [10] J. Zhang, J. Li, dan W. Wang, "A class-imbalanced deep learning fall detection algorithm using wearable sensors," *Sensors*, vol. 21, no. 19, Okt 2021, doi: 10.3390/s21196511.
- [11] S. Alabdulwahab, Y. T. Kim, A. Seo, dan Y. Son, "Generating Synthetic Dataset for ML-Based IDS Using CTGAN and Feature Selection to Protect Smart IoT Environments," *Applied Sciences (Switzerland)*, vol. 13, no. 19, Okt 2023, doi: 10.3390/app131910951.
- [12] O. Parise, R. Kronenberger, G. Parise, C. de Asmundis, S. Gelsomino, dan M. La Meir, "CTGAN-driven synthetic data generation: A multidisciplinary, expert-guided approach (TIMA)," *Comput. Methods Programs Biomed.*, vol. 259, Feb 2025, doi: 10.1016/j.cmpb.2024.108523.
- [13] H. Park, H. Mitsumine, dan M. Fujii, "Automatic Adjustment of the Distance Ratio Threshold in Nearest Neighbor Distance Ratio Matching for Robust Camera Tracking," *IEICE Transactions*, vol. 94-D, hlm. 938–940, Apr 2011, doi: 10.1587/transinf.E94.D.938.
- [14] J.-W. Lee, J.-H. Jeong, J.-W. Kang, S.-I. Na, dan D.-S. Jeong, "Image Feature Point Selection Method Using Nearest Neighbor Distance Ratio Matching," *Journal of the Institute of Electronics and Information Engineers*, vol. 49, Apr 2012, doi: 10.5573/ieek.2012.49.12.124.
- [15] A. P. S. S. Kumar, dan V. K. R., "MQTT Implementations, Open Issues, and Challenges: A Detailed Comparison and Survey," *International Journal of Sensors, Wireless Communications and Control*, vol. 13, Apr 2022, doi: 10.2174/2210327913666221216152446.
- [16] Y. J. Park, S. Y. Jung, T. Y. Son, dan S. J. Kang, "Self-Organizing IoT Device-Based Smart Diagnosing Assistance System for Activities of Daily Living," *Sensors*, vol. 21, no. 3, 2021, doi: 10.3390/s21030785.
- [17] S. Watanuki dkk., "Applying a Method for Augmenting Data Mixed from Two Different Sources Using Deep Generative Neural Networks to Management Science," *Applied Sciences*, vol. 14, no. 1, 2024, doi: 10.3390/app14010378.
- [18] Z. Zhao, A. Kunar, R. Birke, H. der Scheer, dan L. Y. Chen, "CTAB-GAN+: enhancing tabular data synthesis," *Front. Big Data*, vol. Volume 6-2023, 2024, doi: 10.3389/fdata.2023.1296508.
- [19] M. Khavkin dan E. Toch, "Investigating the impact of differential privacy obfuscation on users' data disclosure decisions," *Decis. Support Syst.*, vol. 196, hlm. 114474, 2025, doi: <https://doi.org/10.1016/j.dss.2025.114474>.
- [20] M. Pereira, M. Kshirsagar, S. Mukherjee, R. Dodhia, J. L. Ferres, dan R. de Sousa, "Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data," *PLoS One*, vol. 19, no. 2 February, Feb 2024, doi: 10.1371/journal.pone.0297271.
- [21] J. Fabuyi, "Leveraging Synthetic Data as a Tool to Combat Bias in Artificial Intelligence (AI) Model Training," *Journal of Engineering Research and Reports*, vol. 26, hlm. 24–46, Apr 2024, doi: 10.9734/jerr/2024/v26i121337.
- [22] Y. He, R. Vershynin, dan Y. Zhu, "Algorithmically Effective Differentially Private Synthetic Data," dalam *Proceedings of Thirty Sixth Conference on Learning Theory*, G. Neu dan L. Rosasco, Ed., dalam *Proceedings of Machine Learning Research*, vol. 195. PMLR, Apr 2023, hlm. 3941–3968. [Daring]. Tersedia pada:

- <https://proceedings.mlr.press/v195/he23a.html>
- [23] Q. Liu, M. Khalil, J. Jovanovic, dan R. Shakya, "Scaling While Privacy Preserving: A Comprehensive Synthetic Tabular Data Generation and Evaluation in Learning Analytics," dalam *ACM International Conference Proceeding Series*, Association for Computing Machinery, Mar 2024, hlm. 620–631. doi: 10.1145/3636555.3636921.