HOAQ: JURNAL TEKNOLOGI INFORMASI, Volume 17 Nomor 1, Mei 2026, p102-111
p-ISSN: 2337-5280, e-ISSN: 26207427
DOI: https://doi.org/10.52972/hoaq.vol17no1

EVALUASI EFEKTIVITAS TEKNIK PRIVACY-PRESERVING: K-ANONYMITY, L-DIVERSITY, T-CLOSENESS PADA DATA SENSITIF

Ronal[1], Desy Ebigael Tambunan[2], Yuliana[3]

[1] Program Studi Rekayasa Instrumentasi dan Automasi, Institut Teknologi Sumatera
[2] Program Studi Sistem Informasi, STMIK LIKMI
[3] Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera
Jl. Terusan Ryacudu, Jatiagung, Lampung Selatan, Indonesia
Email: [1]ronal@ia.itera.ac.id, [2]2025210018.student@likmi.ac.id, [3]yuliana@sd.itera.ac.id

ABSTRAK

Perlindungan privasi menjadi aspek krusial dalam pengumpulan, pengolahan, dan publikasi data sensitif, namun potensi risiko kebocoran informasi dapat menimbulkan konsekuensi hukum maupun kerugian reputasi. Untuk menjaga keseimbangan antara kegunaan data dan privasi individu, teknik anonimisasi menjadi pendekatan utama, termasuk penerapan k-anonymity dan evaluasi menggunakan *l-diversity* dan *t-closeness*. Penelitian ini bertujuan untuk mengevaluasi efektivitas teknik-teknik tersebut dalam mengurangi risiko pengungkapan identitas dan atribut sensitif pada *dataset* kesehatan. Studi kasus menggunakan 55500 dataset medis dengan *quasi-identifier Age*, Gender, dan *Blood Type*, serta atribut sensitif *Medical Condition*. *Dataset* dianonimkan menggunakan k-anonymity melalui proses generalisasi dan supresi untuk membentuk *equivalence class* dengan ukuran minimum $k \geq 5$. Selanjutnya, dataset dievaluasi menggunakan *l-diversity* untuk mengukur keberagaman atribut sensitif dalam setiap kelompok, serta t-*closeness* untuk menilai kesamaan distribusi atribut sensitif terhadap distribusi global menggunakan Earth Mover's Distance (EMD). Hasil pengujian menunjukkan bahwa seluruh *equivalence class* telah memenuhi $k \geq 5$ dengan *suppression rate* sebesar 1,15%. Evaluasi l-diversity menunjukkan tidak terdapat *equivalence class* dengan $l < 2$, sehingga risiko *attribute disclosure* dapat diminimalkan. Pengujian t-*closeness* menggunakan Earth Mover's Distance (EMD) menunjukkan mayoritas kelas memiliki EMD $\leq 0{,}15$ dan hanya satu kelas dengan nilai sedikit di atas ambang batas $t = 0{,}2$. Dari sisi utilitas data, nilai Normalized Generalized Information Loss (NGIL) sebesar 0,079 (7,9%) dan AECS sebesar 6,28 menunjukkan tingkat kehilangan informasi yang rendah tanpa terjadi over-generalization. Secara keseluruhan, kombinasi metode yang diterapkan berhasil mencapai keseimbangan antara perlindungan privasi dan *data utility*.

Kata kunci: *Data Privacy, Anonymization, k-Anonymity, l-Diversity, t-Closeness*

ABSTRACT

Privacy protection has become a crucial aspect in the collection, processing, and publication of sensitive data, as potential risks of information leakage may lead to legal consequences and reputational damage. To maintain a balance between data utility and individual privacy, anonymization techniques serve as a primary approach, including the implementation of k-anonymity and its evaluation using l-diversity and t-closeness. This study aims to evaluate the effectiveness of these techniques in reducing the risk of identity and attribute disclosure in a healthcare dataset. The case study utilizes a 55500 *dataset medis containing the quasi-identifiers Age, Gender, and Blood Type, as well as the sensitive attribute Medical Condition. The dataset was anonymized using k-anonymity through generalization and suppression to form equivalence classes with a minimum size of $k \geq 5$. Subsequently, the dataset was evaluated using l-diversity to measure the diversity of sensitive attributes within each group, and t-closeness to assess the similarity between the distribution of sensitive attributes in each group and the global distribution using Earth Mover's Distance (EMD). The results indicate that all equivalence classes satisfy $k \geq 5$ with a suppression rate of 1.15%. The l-diversity evaluation shows that no equivalence class has $l < 2$, thereby minimizing the risk of attribute disclosure. The t-closeness assessment reveals that the majority of classes have EMD ≤ 0.15, with only one class slightly exceeding the threshold of $t = 0.2$. In terms of data utility, the Normalized Generalized Information Loss (NGIL) value of 0.079 (7.9%) and an AECS of 6.28 indicate a low level of information loss without over-generalization. Overall, the combination of methods successfully achieves a balance between privacy protection and data utility, ensuring that the dataset remains suitable for further analysis and secondary data publication.*

Keywords: Data Privacy, Anonymization, k-Anonymity, l-diversity, t-Closeness.

1. PENDAHULUAN

Pengumpulan informasi digital oleh individu maupun lembaga telah menciptakan peluang besar untuk pengambilan keputusan berbasis pengetahuan. Data digital yang dikumpulkan dari media sosial, lembaga keuangan, layanan medis, *platform* pemasaran, dan berbagai aplikasi digital lainnya seringkali mengandung informasi sensitif yang jika bocor dapat mengancam privasi seseorang, menimbulkan konsekuensi hukum atau menyebabkan kerugian reputasi. Rumah sakit berlisensi di California diwajibkan untuk menyerahkan data demografis tertentu pada setiap pasien yang keluar dari fasilitas mereka [1]. Salah satu layanan penyewaan film daring yang populer yaitu Netflix mempublikasikan satu set data yang berisi penilaian film dari 500.000 pelanggannya dalam upaya meningkatkan akurasi rekomendasi film berdasarkan preferensi personal [2]. Pada tahun 2026, AOL (American *Online*) mempublikasikan log kueri pencarian pengguna, namun data publikasi log kueri pencarian segera ditarik dikarenakan re-identifikasi seorang pencari [2].

Data yang berisi informasi spesifik per individu dalam bentuk asli sering kali memuat informasi sensitif dan mempublikasikan data tersebut secara langsung akan melanggar privasi individu [3]. Sehingga muncul kebutuhan untuk melindungi informasi secara efektif yang memungkinkan data tetap bernilai untuk tujuan analisis dan penelitian sekaligus mencegah identifikasi individu. Kebutuhan dalam mengembangkan metode dan alat untuk mengumpulkan, mengolah maupun mempublikasikan data dalam lingkungan yang lebih berisiko merupakan hal yang penting sehingga data yang dikumpulkan, diolah dan dipublikasikan tetap berguna secara praktis sekaligus menjaga privasi individu. Salah satu pendekatan utama untuk menjaga privasi individu saat data dipublikasikan adalah dengan menerapkan anonimitas [4].

Anonimitas merupakan proses menyembunyikan identitas seseorang di dalam sebuah kumpulan data sehingga individu tersebut tidak dapat dikenali baik secara langsung melalui atribut identitas maupun secara tidak langsung melalui kombinasi atribut *quasi-identifier* seperti usia, jenis kelamin atau kode pos. Tanpa penerapan anonimitas, seorang penyerang dapat melakukan re-identification attack, yaitu upaya menghubungkan data anonim dengan identitas asli menggunakan sumber informasi eksternal seperti media sosial, catatan publik, atau basis data lain yang tersedia. Untuk mewujudkan anonimitas, sejumlah teknik telah dikembangkan seperti *suppression* dan *generalization*, hingga dilakukan evaluasi menggunakan metode yang lebih terstruktur seperti *k-anonymity*, *l-diversity*, dan *t-closeness*. Perlindungan privasi dalam data dapat dipahami melalui beberapa model. K-anonymity diterapkan sebagai teknik anonimisasi untuk memastikan bahwa setiap individu tidak dapat dibedakan dari setidaknya $k-1$ individu lain dalam dataset. Selanjutnya, tingkat perlindungan terhadap atribut sensitif dievaluasi menggunakan model *l-diversity*, yang menilai keberagaman nilai atribut sensitif dalam setiap kelompok sehingga penyerang tidak dapat menebak informasi sensitif meskipun kelompoknya telah teridentifikasi [5]. Selain itu, t-*closeness* digunakan sebagai ukuran evaluasi statistik untuk memastikan bahwa distribusi atribut sensitif dalam suatu grup tidak menyimpang secara signifikan dari distribusi data secara keseluruhan, sehingga mengurangi risiko serangan berbasis inferensi statistik [6]. Dengan demikian, anonimitas tidak hanya dipahami sebagai proses teknis pengaburan data, tetapi sebagai bagian dari strategi privasi yang komprehensif untuk menjaga keseimbangan antara kegunaan data dan perlindungan hak privasi individu.

Penelitian terdahulu telah banyak membahas anonimisasi data sebagai mekanisme perlindungan privasi melalui model *k-anonymity*, *l-diversity*, dan *t-closeness*. Konsep *k-anonymity* diperkenalkan untuk mencegah *identity disclosure* dengan menyamarkan atribut *quasi-identifier*, dan terbukti efektif dalam mengurangi risiko identifikasi langsung pada berbagai domain, termasuk data kesehatan [3], [4]. Namun, sejumlah studi menunjukkan bahwa pemenuhan *k-anonymity* saja belum cukup untuk melindungi atribut sensitif, karena nilai atribut sensitif dalam satu *equivalence class* dapat bersifat homogen sehingga rentan terhadap *attribute disclosure*. Untuk mengatasi keterbatasan tersebut, *l-diversity* dikembangkan dengan menekankan keberagaman nilai atribut sensitif dalam setiap kelompok, yang mampu mengurangi risiko penebakan atribut sensitif meskipun kelompok telah teridentifikasi.

Meskipun demikian, penelitian lanjutan menunjukkan bahwa *l-diversity* masih memiliki kelemahan ketika distribusi atribut sensitif dalam *equivalence class* menyimpang secara signifikan dari distribusi global dataset, sehingga membuka peluang serangan berbasis inferensi statistik [5]. Model *t-closeness* kemudian diperkenalkan untuk membatasi penyimpangan distribusi tersebut dan memberikan perlindungan yang lebih kuat terhadap serangan inferensi [6]. Namun, sebagian besar penelitian terdahulu masih mengevaluasi setiap model secara terpisah atau berfokus pada pemenuhan kriteria formal anonimitas tanpa mengkaji efektivitas pendekatan anonimitas berlapis serta dampaknya terhadap kegunaan data (*data utility*), khususnya dalam konteks kombinasi atribut *quasi-identifier* dan ketersediaan data eksternal. Selain itu penelitian mengenai anonimisasi berlapis masih terbatas dan tanpa analisis mendalam terhadap interaksi antar parameter maupun dampaknya terhadap *data utility*. Belum banyak penelitian yang mengidentifikasi *equivalence class* yang tetap berisiko meskipun telah memenuhi *k-anonymity* dan *l-diversity*, serta menjelaskan secara statistik mengapa penyimpangan distribusi masih terjadi.

Penelitian ini difokuskan pada evaluasi anonimisasi data secara berlapis melalui penerapan *k-anonymity*, *l-diversity*, dan *t-closeness* pada *dataset* yang mengandung atribut sensitif. Dengan pendekatan ini, penelitian diarahkan untuk mengidentifikasi sejauh mana keberagaman atribut sensitif tetap terjaga setelah proses *k-anonymity*, serta distribusi atribut sensitif dalam setiap *equivalence class* telah mendekati distribusi global sebagaimana dipersyaratkan oleh *t-closeness*. Sejalan dengan fokus tersebut, tujuan penelitian ini adalah menerapkan dan mengevaluasi anonimisasi data menggunakan model *k-anonymity*, *l-diversity*, dan *t-closeness.* Evaluasi dilakukan untuk mengukur efektivitas *l-diversity* dalam mencegah homogenisasi atribut sensitif serta

mengkaji keterbatasannya dalam menjamin keseragaman distribusi statistik. Selain itu, penelitian ini bertujuan mengidentifikasi *equivalence class* yang masih memiliki penyimpangan distribusi atribut sensitif terhadap distribusi global, sehingga dapat dirumuskan strategi lanjutan berupa generalisasi tambahan atau supresi selektif guna memenuhi ambang batas *t-closeness* yang ditetapkan.

Penelitian ini memberikan kontribusi ilmiah dengan mengembangkan kerangka evaluasi anonimisasi berlapis yang secara sistematis menganalisis interaksi antara k-*anonymity*, l-*diversity*, dan t-*closeness* dalam satu skema terpadu. Berbeda dari penelitian sebelumnya yang cenderung menguji masing-masing model secara terpisah, studi ini menunjukkan secara empiris bahwa pemenuhan k-*anonymity* dan l-*diversity* tidak secara inheren menjamin terpenuhinya t-*closeness*, khususnya ketika distribusi atribut sensitif dalam equivalence class mengalami penyimpangan terhadap distribusi global. Lebih lanjut, penelitian ini mengidentifikasi pola distribusional yang menyebabkan risiko inferensi statistik tetap muncul setelah penerapan anonimisasi berlapis, serta menganalisis implikasinya terhadap keseimbangan antara perlindungan privasi dan kegunaan data.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional untuk mengevaluasi efektivitas teknik anonimisasi data dalam melindungi privasi individu. Tahapan penelitian diawali dengan tinjauan pustaka untuk mengidentifikasi konsep, model, dan kelemahan metode anonimisasi data yang telah dikembangkan sebelumnya. Selanjutnya, dilakukan analisis dan penerapan teknik anonimisasi menggunakan model k-*anonymity*, l-*diversity*, dan t-*closeness* pada dataset yang mengandung atribut *quasi-identifier* dan atribut sensitif. Evaluasi anonimisasi difokuskan pada pengukuran keberagaman atribut sensitif serta kesamaan distribusi statistik antara setiap *equivalence class* dan distribusi global data. Pengukuran *t-closeness* dilakukan menggunakan Earth Mover's Distance (EMD) dengan ambang batas tertentu. Hasil anonimisasi kemudian dianalisis untuk menilai tingkat perlindungan privasi dan dampaknya terhadap kegunaan data, sehingga dapat diperoleh gambaran keseimbangan antara keamanan privasi dan utilitas data.

K-*Anonimity*

K-*Anonymity* merupakan salah satu teknik perlindungan privasi yang paling banyak digunakan dalam *privacy-preserving data publishing (PPDP)*. Suatu dataset dikatakan memenuhi k-anonymity apabila setiap kombinasi atribut *quasi-identifier (QI)* muncul pada sedikitnya k *record*, sehingga setiap individu tidak dapat dibedakan dari minimal *k−1* individu lainnya dalam *equivalence class*. Teknik ini dicapai melalui proses generalisasi dan *suppression* untuk mengurangi granularitas informasi sehingga risiko *re-identification* [7]. Persamaan Quasi Identifier dapat dilihat pada persamaan 1.

$$QI = \{A_i, \ldots, A_j\} \subseteq \{A_1, \ldots, A_n\} \quad (1)$$

Dan terdapat persamaan 2 berikut

$$\exists p_i \in U \quad (2)$$

Dimana

$$f_g(f_c(p_i)[QT]) = p_i \quad (3)$$

Pada persamaan dapat terlihat setelah data mengalami transformasi melalui fungsi anonimisasi fc dan fungsi generalisasi fg , identitas asli masih dapat dikembalikan berdasarkan atribut yang termasuk dalam QI sehingga atribut tersebut dianggap sebagai *Quasi Identifier*. Sebuah tabel T memenuhi k-anonymity jika setiap *record* t tidak bisa dibedakan dari minimal k-1 record lainnya berdasarkan seluruh atribut dalam *quasi identifier* QT, *k-anonymity* dapat dilihat pada persamaan 4.

$$\forall t \in T, \exists\{t_{i1}, t_{i2}, \ldots t_{ik-1}\} \subseteq T \quad (4)$$

Teknik k-*anonymity* dianalisis menggunakan konsep generalisasi data dimana dilakukan penyembunyian informasi asli tetapi tetap mempertahankan tren dan pola. Teknik penyembunyian pesan biasanya menggunakan teknik data *masking* seperti generalisasi dan *suppression* [6]. Ilustrasi pada dataset menggunakan teknik anonimisasi dapat dilihat pada Tabel 1.

Tabel 1. Dataset asli

	Identifying Variable		*Quasi Identifier*		*Sensitive Attribute*	*Quasi Identifier*
ID	*Name*	*Age*	*Gender*	*Blood Type*	*Medical Condition*	*Date of Admission*
1	Bobby Jackson	30	Male	B-	Cancer	2024-01-31
2	Leslie Terry	62	Male	A+	Obesity	2019-08-20
3	Danny Smith	76	Female	A-	Obesity	2022-09-22
4	Andrew Watts	28	Female	O+	Diabetes	2020-11-18

	Identifying Variable	*Quasi Identifier*			*Sensitive Attribute*	*Quasi Identifier*
ID	*Name*	*Age*	*Gender*	*Blood Type*	*Medical Condition*	*Date of Admission*
5	Adrienne Bell	43	Female	AB+	Cancer	2022-09-19
6	Emily Johnson	36	Male	A+	Asthma	2023-12-20
7	Edward Edwards	21	Female	AB-	Diabetes	2020-11-03
8	Christina Martinez	20	Female	A+	Cancer	2021-12-28
9	Jasmine Aguilar	82	Male	AB+	Asthma	2020-07-01

Data asli berisi data sensitif pasien yang akan diproses menggunakan teknik anonymization, atribut *quasi-identifier* seperti *Age*, *Gender*, *Blood Type*, dan *Date of Admission* mengalami transformasi melalui teknik generalisasi dan suppression. Generalisasi bertujuan menaikkan level abstraksi suatu atribut, sehingga detail yang terlalu spesifik dihilangkan dan beberapa individu menjadi tidak lagi unik. Teknik ini sangat umum digunakan pada privasi data publikasi medis [8]. Sementara itu suppression digunakan untuk menyembunyikan nilai tertentu, misalnya dengan mengganti kata atau huruf "*" atau mengosongkan nilai ketika suatu kombinasi atribut masih terlalu spesifik dan dapat memicu re-identifikasi [9].

Penerapan gabungan generalisasi dan suppression memungkinkan dataset mencapai k-anonymity, di mana setiap baris data memiliki minimal $k - 1$ baris lain yang identik dalam atribut QI. Dalam penelitian ini, dataset diolah agar mencapai $k \geq 5$, sehingga tidak ada individu yang memiliki kombinasi atribut unik yang dapat dicocokkan dengan data eksternal (*linkage attack*). Teknik ini digunakan pada publikasi data kesehatan karena risiko identifikasi ulang pada data medis sangat tinggi [7]. Meskipun demikian, peningkatan generalisasi dan suppression menurunkan utility data, sebab semakin banyak detail yang hilang sehingga analisis tertentu dapat menjadi kurang akurat [10]. Tabel 2 merupakan tabel hasil dari teknik anonimisasi yang dilakukan pada Tabel 1.

Tabel 2. Dataset Anonimisasi

ID	*Age*	*Gender*	*Blood Type*	*Medical Condition*	*Date*
1	30–39	Male	B	Cancer	2024
2	60–69	Male	A	Obesity	2019
3	70–79	Female	A	Obesity	2022
4	20–29	Female	O	Diabetes	2020
5	40–49	Female	AB	Cancer	2022
6	30–39	Male	A	Asthma	2023
7	20–29	Female	AB	Diabetes	2020
8	20–29	Female	A	Cancer	2021
9	80–89	Male	AB	Asthma	2020

Pada Tabel 2 teknik *anonymization* yang digunakan yaitu teknik generalisasi dan suppression untuk menghilangkan nilai-nilai unik pada *quasi-identifiers*, menunjukkan pemenuhan standar *k-anonymity* di semua entri. Teknik ini benar-benar membentuk kelompok yang tidak dapat dibedakan satu sama lain berdasarkan atribut publik, sehingga risiko identifikasi ulang berkurang secara signifikan. Hal ini sejalan dengan praktik umum dalam literatur privasi data bahwa kombinasi generalisasi dan suppression merupakan cara yang efektif untuk menghambat serangan identifikasi seperti *linkage attacks* [11]. Pendekatan ini menjadi tantangan bahwa anonymization harus dirancang sedemikian rupa agar tidak hanya memenuhi syarat formal k-anonymity tetapi juga mempertimbangkan kebutuhan praktis analisis data [12].

L-Diversity

L-diversity adalah *privacy model* untuk mengevaluasi data yang menerapkan *k-anonymity* dengan memastikan bahwa setiap kelompok records yang tidak bisa dibedakan oleh *quasi-identifier* memiliki minimal 1 nilai yang berbeda untuk atribut sensitif sehingga meminimalkan risiko inferensi terhadap nilai sensitif. Model ini diperlukan karena meskipun data telah dianonimisasi dengan *k-anonymity*, atribut sensitif masih dapat diprediksi jika nilainya homogen dalam kelompok tersebut atau jika penyerang memiliki pengetahuan tentang data tersebut. Untuk penerapan *l-diversity*, *dataset* medis yang terdiri dari 55500 *record* dengan contoh 9 *record* data sebagaimana ditunjukkan pada Tabel 2. Dataset tersebut memuat atribut Age, Gender, Blood Type, Medical *Condition* dan *Date* [13][15]. Setelah dilakukan proses generalisasi terhadap atribut *quasi-identifier*, *record-record* yang memiliki nilai *Age* dan *Gender* yang sama dikelompokkan ke dalam satu *equivalence class*. Berdasarkan prinsip *l-diversity* yaitu suatu *equivalence class* data memenuhi *l-diversity* dimana terdapat kelas mengandung paling sedikit l nilai atribut sensitif yang terwakili dengan baik.

Sebagai contoh equivalence class dengan Age = 30–39 dan Gender = Male, yang ditunjukkan pada Tabel 2 terdiri dari dua *record* dengan *nilai Medical Condition* yang berbeda, yaitu Cancer dan Asthma. Dengan demikian, *equivalence class* ini memenuhi 2-*diversity*, karena memiliki dua nilai sensitif yang berbeda. Namun, beberapa

equivalence class lainnya hanya terdiri dari satu *record* seperti Age = 70–79, Gender = *Female*, yang hanya memiliki satu nilai sensitif, yaitu Obesity. *Equivalence class* tersebut tidak memenuhi persyaratan *l-diversity*. Apabila penyerang mengetahui informasi *quasi-identifier* seseorang, maka nilai atribut sensitifnya dapat disimpulkan secara pasti. Situasi ini dikenal sebagai homogeneity attack, yang tidak dapat dicegah hanya dengan menerapkan *k-anonymity*. Sehingga meskipun generalisasi atribut *quasi-identifier* dapat memenuhi kriteria k-anonymity perlindungan terhadap kebocoran atribut sensitif belum tentu tercapai. Penerapan *l-diversity* memperkuat perlindungan privasi dengan memastikan adanya keragaman nilai pada atribut sensitif di setiap *equivalence class*, sehingga mengurangi risiko inferensi terhadap informasi sensitif individu [14].

T-*Closeness*

T-*closeness* adalah model anonimisasi untuk mencegah *attribute disclosure* atau pengungkapan atribut sensitif. Prinsip t-*closeness* yaitu untuk setiap *equivalence class* yaitu kelompok baris yang tidak dapat dibedakan berdasarkan *quasi-identifiers*, distribusi nilai atribut sensitif di dalam kelompok harus "dekat" dengan distribusi atribut sensitif pada seluruh dataset. Selisih distribusi tersebut dibatasi oleh parameter t [15]. Dengan demikian, mengetahui bahwa seseorang ada dalam suatu kelompok tidak memberi pengamat informasi probabilistik yang jauh berbeda tentang nilai sensitifnya dibandingkan jika pengamat hanya mengetahui distribusi populasi.

Earth Mover's Distance (EMD) adalah metrik untuk mengukur "kedekatan" distribusi. EMD dapat mengukur biaya minimum yang diperlukan untuk mentransformasikan satu distribusi probabilitas menjadi distribusi lain dengan mempertimbangkan jarak antar nilai dalam domain atribut sensitif. EMD dapat menangkap perbedaan struktural dan semantik antar distribusi khususnya pada atribut sensitif yang bersifat terurut atau memiliki makna numerik. Secara formal EMD ditunjukan pada persamaan 5. Pada persamaan 5 dapat dilihat P merupakan distribusi atribut sensitif global dan Q merupakan distribusi atribut sensitif pada suatu *equivalence class* [15].

$$EMD(P,Q) = \min \sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} \, . \, d_{ij} \qquad (5)$$

Dalam model t-*closeness*, suatu *equivalence class* dinyatakan memenuhi kriteria privasi apabila jarak antara distribusi atribut sensitif pada kelas tersebut dan distribusi atribut sensitif pada seluruh dataset yang diukur menggunakan EMD tidak melebihi ambang batas t. Pseudocode metode *t-closeness* yang digunakan dalam penelitian ini disajikan pada Algoritma 1.

```
# Input: D:Dataset, Q:Quasi-Identifier, S:Sensitive Attribut, k:k-anonymity parameter,
t:t-closeness threshold
# Output: D_prime : Anonymized dataset
# Compute global distribution of sensitive attribute
P_global = distribution(D[SA])
# Initialization
D_prime = []
UnassignedRecords = D.copy()
while len(UnassignedRecords) > 0:
    # Create an equivalence class using clustering on QI
    EC = select_subset(UnassignedRecords, QI)
    # Enforce k-anonymity
    if len(EC) < k:
        EC = merge_with_nearest_equivalence_class(EC)
        continue
    # Compute sensitive attribute distribution in EC
    P_EC = distribution(EC[SA])
    # Compute distance between distributions
    dist = distance_metric(P_EC, P_global)
    # Check t-closeness condition
    if dist <= t:
        # Apply generalization on QI attributes
        EC = generalize_QI(EC, QI)
        # Add EC to anonymized dataset
        D_prime.extend(EC)
        # Remove EC records from unassigned records
        UnassignedRecords = remove_records(UnassignedRecords, EC)
    else:
        # Increase generalization level on QI
        increase_generalization(QI)
return D_prime
```

Data Utility

Evaluasi *data utility* dilakukan untuk mengukur tingkat kehilangan informasi akibat proses generalisasi dan supresi yang diterapkan dalam pemenuhan k-*anonymity*, *l-diversity*, dan *t-closeness*. Pengukuran dilakukan menggunakan metrik Normalized Generalized Information Loss. NGIL dihitung sebagai proporsi *record* yang mengalami supresi terhadap total *record*, dirumuskan dengan persamaan 6.

$$NGIL = \frac{1}{N \times m}\sum_{i=1}^{N}\sum_{j=1}^{m} Loss_{ij} \quad (6)$$

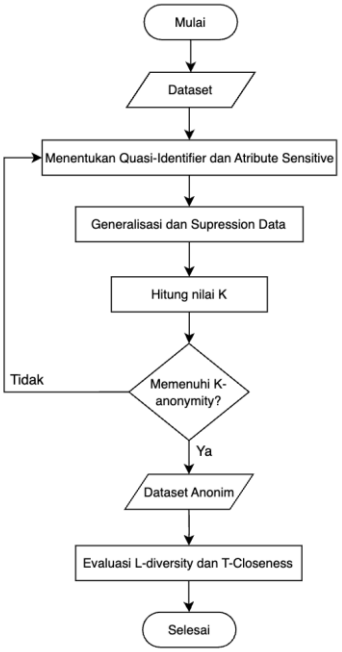
Dengan N adalah jumlah total *record*, m adalah jumlah atribut *quasi-identifier*, dan Lossij adalah kehilangan informasi pada atribut ke-j dari *record* ke-i. Normalisasi terhadap jumlah record dan atribut memungkinkan perbandingan antar skenario anonimisasi yang berbeda [2]. Untuk atribut numerik, kehilangan informasi dihitung berdasarkan proporsi rentang generalisasi dalam *equivalence class* terhadap rentang global atribut dengan persamaan 7.

$$Loss_{ij} = \frac{\max(EC\,(i),\, A_j) - \min(EC(i),\, A_j)}{\max(A_j) - \min(A_j)} \quad (7)$$

Normalized Generalized Information Loss (NGIL) mengukur rata-rata kehilangan informasi per atribut *quasi-identifier* setelah proses generalisasi dan supresi. Nilai ini dinormalisasi terhadap jumlah *record* dan jumlah atribut sehingga memungkinkan perbandingan antar skenario anonimisasi yang berbeda. Semakin kecil nilai NGIL, semakin tinggi tingkat preservasi informasi [16].

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan *Healthcare Dataset* yang diperoleh dari *platform* Kaggle. *Dataset* terdiri dari 55500 *record*, di mana setiap *record* merepresentasikan data pasien kesehatan sintetis. *Dataset* mencakup berbagai atribut, antara lain informasi demografis pasien, kondisi medis, serta detail administrasi perawatan. *Dataset* ini bersifat sepenuhnya sintetis, tidak mengandung data pasien nyata, dan disediakan untuk keperluan edukasi serta non-komersial. Dalam penelitian dilakukan tahap identifikasi atribut *quasi-identifier* (QI). Atribut yang dipilih sebagai QI pada penelitian ini adalah *Age*, *Gender*, dan *Blood Type*. Pemilihan atribut tersebut didasarkan pada kemampuan atribut untuk mengidentifikasi individu secara tidak langsung apabila dikombinasikan, meskipun tidak mengandung identitas eksplisit seperti nama atau nomor identitas. Tahap berikutnya adalah pembentukan *equivalence class*. *Dataset* dikelompokkan berdasarkan kombinasi nilai dari atribut *quasi-identifier* yang telah ditentukan. Setiap kelompok yang terbentuk merepresentasikan satu *equivalence class*, yaitu sekumpulan *record* yang memiliki nilai QI yang sama. Jumlah *record* pada setiap *equivalence class* kemudian dihitung untuk mengetahui ukuran masing-masing kelompok. Nilai k-*anonymity* ditentukan sebagai jumlah minimum *record* dari seluruh *equivalence class* yang terbentuk. Nilai ini merepresentasikan tingkat anonimitas dataset di mana semakin kecil nilai k, semakin tinggi risiko re-identifikasi data. Terakhir melakukan evaluasi terhadap nilai ambang k. Nilai k ditetapkan sebesar 5 untuk mencapai keseimbangan antara perlindungan privasi dan kegunaan data, di mana setiap individu tidak dapat dibedakan dari setidaknya empat individu lainnya, sementara tingkat generalisasi yang diterapkan masih memungkinkan analisis data yang bermakna. Secara keseluruhan, tahapan evaluasi memberikan gambaran awal mengenai tingkat risiko privasi pada dataset sensitif dan menjadi dasar penting dalam penerapan teknik *privacy-preserving* lanjutan. *Flowchart* Anonimisasi data hingga evaluasi dapat dilihat pada Gambar 1 *flowchart* Anonimisasi data.



Gambar 2. *Flowchart* Anonimisasi

Evaluasi *K-Anonymity*

Evaluasi *k-anonymity* pada penelitian ini dilakukan mengikuti alur yang ditunjukkan pada Gambar 1. Proses diawali dengan identifikasi atribut *quasi-identifier* dan atribut sensitif, di mana atribut *quasi-identifier* digunakan untuk membentuk *equivalence class*, sedangkan atribut sensitif dipertahankan tanpa generalisasi. Selanjutnya generalisasi dan supresi diterapkan pada atribut *quasi-identifier*. Evaluasi *k-anonymity* dilakukan dengan menghitung ukuran setiap *equivalence class* setelah proses generalisasi dan supresi pada atribut *quasi-identifier*. Nilai k ditentukan sebagai ukuran minimum *equivalence class*. Tabel 4 menunjukkan distribusi ukuran *equivalence class* pada setiap tahap pemrosesan data. Hasil evaluasi menunjukkan bahwa nilai k minimum meningkat pada data awal menjadi k = 5 setelah penerapan generalisasi dan supresi terhadap 55500 *record*, sehingga seluruh data yang dipublikasikan telah memenuhi *k-anonymity* dengan k ≥ 5. Peningkatan ini terjadi karena supresi menghilangkan kombinasi atribut yang sangat unik dan tidak dapat digabungkan lebih lanjut tanpa mengubah struktur generalisasi secara signifikan. Hasil menegaskan bahwa kombinasi generalisasi dan supresi diperlukan untuk mencapai ambang batas k-*anonymity* yang ditentukan, serta menunjukkan bahwa pengendalian keunikan distribusi *quasi-identifier* merupakan faktor utama dalam menurunkan risiko identifikasi langsung sebelum analisis perlindungan atribut sensitif dilakukan melalui *l-diversity* dan *t-closeness*.

Tabel 4. Hasil *K-Anonimity*

Tahap Pemrosesan	Jumlah *Record*	*Record* dalam *Equivalence Class* k < 5	Nilai k Minimum
Data awal (tanpa anonimisasi)	55.500	3.912	1
Setelah generalisasi	55.500	1.284	3
Setelah supresi	54.860	0	5

Evaluasi L-*Diversity*

Evaluasi *l-diversity* dilakukan menggunakan *dataset* hasil penerapan k-*anonymity* untuk menilai perlindungan terhadap risiko *attribute disclosure*. Nilai l ditentukan sebagai nilai minimum keberagaman atribut sensitif dari seluruh kelompok ekuivalensi. Hasil pengujian menunjukkan bahwa seluruh *equivalence class* memiliki nilai l ≥ 2 dan tidak ditemukan kelompok yang melanggar kriteria *l-diversity*. Hasil evaluasi dapat dilihat pada Tabel 5. Kondisi ini mengindikasikan bahwa setiap kelompok ekuivalensi tidak didominasi oleh satu nilai atribut sensitif, sehingga penyerang tidak dapat secara langsung menyimpulkan informasi sensitif individu meskipun kelompoknya telah teridentifikasi. Tidak ditemukannya pelanggaran *l-diversity* menunjukkan bahwa proses generalisasi dan supresi pada tahap k-*anonymity* tidak hanya meningkatkan anonimitas identitas, tetapi juga berhasil menjaga keberagaman atribut sensitif dalam setiap kelompok. Dengan demikian, risiko *attribute disclosure* dapat diminimalkan dan dataset memiliki tingkat perlindungan privasi yang lebih kuat dibandingkan dengan penerapan k-anonymity saja. Hal ini terjadi karena proses generalisasi dan supresi pada tahap *k-anonymity* tidak hanya memperbesar ukuran *equivalence class*, tetapi juga secara tidak langsung mendorong penggabungan record yang memiliki variasi atribut sensitif berbeda sehingga meningkatkan keberagaman dalam setiap kelompok. Namun demikian, distribusi nilai *l* yang didominasi oleh kelompok dengan *l*=2 menunjukkan bahwa sebagian besar *equivalence class* masih berada pada tingkat keberagaman minimum yang dipersyaratkan. Hal ini berarti bahwa meskipun kriteria formal *l-diversity* telah terpenuhi, komposisi atribut sensitif dalam beberapa kelompok masih berpotensi menunjukkan distribusi yang tidak seimbang.

Tabel 5. Hasil Evaluasi *l-diversity*

Rentang *l*	Jumlah *Equivalence Class*
l= 2	4.216
l = 3	2.891
l ≥ 4	1.635
l < 2	0

Evaluasi T-*Closeness*

Evaluasi t-*closeness* dilakukan untuk menilai kesamaan distribusi atribut sensitif antara setiap *equivalence class* dan distribusi global *dataset*. Pengujian ini bertujuan untuk mengidentifikasi potensi *attribute disclosure* yang masih dapat terjadi meskipun *dataset* telah memenuhi *k-anonymity* dan *l-diversity*. Pada penelitian ini perhitungan *t-closeness* diimplementasikan menggunakan Earth Mover's Distance (EMD) untuk mengukur jarak distribusi probabilitas atribut sensitif pada setiap *equivalence class* terhadap distribusi global, hasil pengujian t-closeness dapat dilihat pada Tabel 6. Implementasi perhitungan ini direalisasikan melalui *source code* 4, dengan menentukan nilai t sebagai nilai maksimum EMD dari seluruh kelompok ekuivalensi. Ambang batas t ditetapkan

sebesar t = 0.2, yang merepresentasikan tingkat toleransi perbedaan distribusi antara kelompok lokal dan distribusi global. Hasil pengujian menunjukkan bahwa nilai t-*closeness* maksimum (EMD) yang diperoleh adalah 0.20117. Nilai tersebut sedikit melebihi ambang batas yang telah ditentukan, sehingga secara formal *dataset* belum sepenuhnya memenuhi kriteria *t-closeness*. Kondisi ini muncul karena *t-closeness* membatasi kesesuaian distribusi probabilitas atribut sensitif antara setiap *equivalence class* dan distribusi global, sehingga mampu mendeteksi ketimpangan yang tidak terjangkau oleh *k-anonymity* maupun *l-diversity*. Ketimpangan proporsi secara langsung meningkatkan nilai Earth Mover's Distance hingga melampaui ambang batas toleransi yang ditetapkan.

Tabel 6. Hasil Evaluasi T-*Closeness*

Rentang EMD	Jumlah *Equivalence Class*
EMD ≤ 0.10	6.832
0.10 < EMD ≤ 0.15	1.674
0.15 < EMD ≤ 0.20	235
EMD > 0.20	1

Pengujian *Data Utility*

Evaluasi tidak hanya berfokus pada tingkat perlindungan privasi, tetapi juga pada sejauh mana data tetap memiliki kegunaan analitis (*data utility*). *Data utility* mengacu pada kemampuan *dataset* yang telah dianonimkan untuk tetap menghasilkan informasi yang valid, akurat, dan representatif terhadap kondisi asli. Oleh karena itu, metrik seperti *Normalized Information Loss* (NGIL) menjadi penting karena memberikan gambaran kuantitatif mengenai kompromi antara perlindungan privasi dan kualitas data yang dipertahankan. Hasil Pengujian NIGL ditunjukan pada tabel 7. Berdasarkan Tabel 7, total *record* awal sebesar 55.500 menunjukkan jumlah data sebelum anonimisasi yang menjadi dasar perbandingan dalam evaluasi *data utility*, sedangkan total *record final* sebesar 54.860 mengindikasikan bahwa sebagian besar data tetap dipertahankan setelah proses generalisasi dan supresi. *Suppression rate* sebesar 1,15% tergolong rendah, sehingga penghapusan data tidak memberikan dampak signifikan terhadap kualitas analisis. Jumlah *quasi-identifier* (QI) sebanyak tiga atribut menunjukkan kompleksitas yang masih terkendali dalam proses anonimisasi, sehingga kebutuhan generalisasi tidak berlebihan. Hal ini tercermin dari jumlah *equivalence class* (EC) yang masih mencapai 8.742, yang menandakan bahwa variasi struktur data tetap terjaga dan tidak terjadi penggabungan kelas secara ekstrem. Nilai AECS sebesar 6,28 yang sedikit di atas batas minimum k = 5 menunjukkan bahwa rata-rata ukuran kelompok tidak terlalu besar, sehingga tidak terjadi *over-generalization*. Secara keseluruhan, nilai NGIL sebesar 0,079 menunjukkan bahwa tingkat kehilangan informasi akibat proses anonimisasi berada pada kategori rendah, yaitu 7,9% dari total informasi struktural yang terkandung dalam atribut *quasi-identifier*. Angka ini mengindikasikan bahwa mayoritas informasi asli *dataset*, yakni sekitar 92,1%, masih dapat dipertahankan setelah penerapan k-*anonymity* dengan k = 5. Kontribusi terbesar terhadap nilai NGIL dalam kasus ini berasal dari proses generalisasi. Hal ini terlihat dari *suppression rate* yang hanya sebesar 1,15%, sehingga penghapusan *record* tidak memberikan dampak signifikan terhadap *total information loss*. Dataset telah memenuhi k-*anonymity* tanpa menyebabkan distorsi signifikan terhadap distribusi dan karakteristik asli data. Dengan tingkat *information loss* yang rendah dan struktur data yang masih terjaga, proses anonimisasi dapat dikategorikan efisien serta mampu mempertahankan *data utility* untuk kebutuhan analisis maupun publikasi data sekunder.

Tabel 7. Hasil Evaluasi *Data Utility*

Komponen	Nilai
Total Record Awal	55.500
Total Record Final	54.860
Suppression Rate	1,15%
Jumlah QI	3
Jumlah EC	8.742
AECS	6,28
NGIL	0,079 (7,9%)

Pembahasan

Berdasarkan hasil pengujian pada Tabel 4 hingga Tabel 7, penerapan k-*anonymity* dilakukan secara bertahap dan berhasil meningkatkan tingkat perlindungan privasi secara signifikan. Pada data awal terdapat 3.912 *record* dalam *equivalence class* dengan k < 5 dan nilai k minimum sebesar 1, yang menunjukkan risiko re-identifikasi tinggi; setelah generalisasi jumlah tersebut menurun menjadi 1.284 dengan k minimum meningkat menjadi 3, dan pada tahap akhir melalui supresi tidak terdapat lagi *equivalence class* dengan k < 5 sehingga k minimum mencapai

5 dengan *suppression rate* hanya 1,15%. Struktur akhir menghasilkan 8.742 *equivalence class* yang menjadi dasar evaluasi lanjutan. Hasil l-*diversity* menunjukkan seluruh *equivalence class* memenuhi $l \geq 2$ (4.216 kelas dengan $l = 2$, 2.891 dengan $l = 3$, dan 1.635 dengan $l \geq 4$), sehingga risiko *attribute disclosure* dapat diminimalkan. Evaluasi t-closeness memperlihatkan mayoritas kelas memiliki EMD ≤ 0,15 dan hanya satu kelas dengan EMD > 0,20 (maksimum 0,20117), yang sangat mendekati ambang batas t = 0,2, sehingga secara umum distribusi atribut sensitif dalam kelompok relatif seragam terhadap distribusi global. Dari sisi data utility, nilai NGIL sebesar 0,079 (7,9%) dan AECS sebesar 6,28 menunjukkan bahwa kehilangan informasi berada pada kategori rendah tanpa terjadi *over-generalization*. Secara keseluruhan, kombinasi k-*anonymity*, l-*diversity*, dan t-*closeness* dalam penelitian ini berhasil mencapai keseimbangan yang baik antara perlindungan privasi dan utilitas data, sehingga *dataset* tetap layak digunakan untuk analisis lanjutan maupun publikasi data sekunder.

4. SIMPULAN

Berdasarkan hasil penelitian, penerapan k-*anonymity* dengan parameter k = 5 berhasil menghilangkan seluruh *equivalence class* dengan ukuran kurang dari lima melalui proses generalisasi dan supresi yang minimal (1,15%). Evaluasi lanjutan menunjukkan bahwa seluruh *equivalence class* memenuhi kriteria *l-diversity* dengan $l \geq 2$, sehingga risiko *attribute disclosure* dapat ditekan. Pengujian *t-closeness* menunjukkan distribusi atribut sensitif pada sebagian besar *equivalence class* mendekati distribusi global, dengan hanya satu kelas yang sedikit melampaui ambang batas t = 0,2. Dari *data utility*, dengan metrik nilai NGIL sebesar 0,079 dan AECS sebesar 6,28 mengindikasikan tingkat kehilangan informasi yang rendah dan tidak terjadi generalisasi berlebihan. Secara keseluruhan, metode anonimisasi yang diterapkan mampu meningkatkan perlindungan privasi secara signifikan dengan tetap mempertahankan kualitas dan kegunaan data, sehingga menghasilkan keseimbangan optimal antara privasi dan utilitas data. Penelitian selanjutnya dapat difokuskan pada optimasi parameter privasi. Salah satu pendekatan yang realistis adalah melakukan pengujian beberapa variasi nilai ambang batas *t* untuk mengidentifikasi titik keseimbangan optimal antara penurunan nilai maksimum EMD dan peningkatan *information loss.* Selain itu, dapat diterapkan strategi generalisasi atau supresi selektif yang hanya ditujukan pada *equivalence class* dengan nilai EMD tertinggi, sehingga perbaikan kedekatan distribusi dapat dicapai tanpa meningkatkan distorsi data secara signifikan.

DAFTAR PUSTAKA

[1] R. Krishna, K. Kelleher, and E. Stahlberg, “Patient Confidentiality in the Research Use of Clinical Medical Databases,” *American Journal of Public Health*, vol. 97, no. 4, pp. 654–658, Apr. 2007, doi: 10.2105/AJPH.2006.093682.

[2] M. Barbaro and T. Zeller Jr., “A face is Exposed for AOL Searcher,” *The New York Times*. [Online]. Aug. 24, 2006. Available: https://www.nytimes.com/2006/08/09/technology/09aol.html. Accessed: Jan. 2026.

[3] P. Samarati and L. Sweeney, “Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement Through Generalization and Suppression.” N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy Beyond K-Anonymity And L-Diversity,” *dalam Proceedings of the 23rd International Conference on Data Engineering (ICDE)*, 2007, pp. 106–115, doi: 10.1109/ICDE.2007.367856.

[4] K. El Emam and F. K. Dankar, “Protecting Privacy Using k-Anonymity,” *Journal of the American Medical Informatics Association*, vol. 15, no. 5, pp. 627–637, Sep. 2008, doi: 10.1197/jamia.M2716. K. El Emam and F. K. Dankar, “Protecting Privacy Using K-Anonymity,” *Journal of the American Medical Informatics Association*, vol. 15, no. 5, pp. 627–637, Sep. 2008, doi: 10.1197/jamia.M2716.

[5] P. Ohm, “Broken Promises of Privacy: Responding to The Surprising Failure of Anonymization.”

[6] A. Majeed and S. Lee, “Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey,” *IEEE Access*, vol. 9, pp. 8512–8545, 2021, doi: 10.1109/ACCESS.2020.3045700

[7] C. Dwork and A. Roth, “The Algorithmic Foundations of Differential Privacy,” *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–487, 2013, doi: 10.1561/0400000042.

[8] C. Dwork, A. Smith, T. Steinke, and J. Ullman, “Exposed! A survey of Attacks on Private Data,” *Annual Review of Statistics and Its Application*, vol. 4, no. 1, pp. 61–84, Mar. 2017, doi: 10.1146/annurev-statistics-060116-054123.

[9] A. Majeed and S. Lee, “Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey,” *IEEE Access*, vol. 9, pp. 8512–8545, 2021, doi: 10.1109/ACCESS.2020.3045700.

[10] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, “A Comprehensive Review on Privacy Preserving Data Mining,” *Springerplus*, vol. 4, no. 1, pp. 1–36, Dec. 2015, doi: 10.1186/s40064-015-1481-x.

[11] H. Lee, S. Kim, J. W. Kim, and Y. D. Chung, “Utility-Preserving Anonymization for Health Data Publishing,” BMC Medical Informatics and Decision Making, vol. 17, no. 1, p. 15, Jul. 2017, doi: 10.1186/s12911-017-0499-0.

[12] V. S. Ivkova and I. R. Opirskyi, “Research of Existing Osint Tools and Approaches in The Context of Personal and State Information Security,” *Computer systems and network*, vol. 7, no. 1, pp. 131–142, Jun. 2025, doi: 10.23939/csn2025.01.131.

[13] K. Oishi, Y. Sei, J. Andrew, Y. Tahara, and A. Ohsuga, "Algorithm to Satisfy L-Diversity by Combining Dummy Records and Grouping," Security and Privacy, vol. 7, no. 3, p. e373, May 2024, doi: 10.1002/spy2.373.
[14] M. Cunha, R. Mendes, and J. P. Vilela, "A Survey of Privacy-Preserving Mechanisms for Heterogeneous Data Types," Computer Science Review, vol. 41, p. 100403, Aug. 2021, doi: 10.1016/j.cosrev.2021.100403.
[15] A. Sepas, A. H. Bangash, O. Alraoui, K. El Emam, and A. El-Hussuna, "Algorithms to Anonymize Structured Medical and Healthcare Data: A Systematic Review," Frontiers in Big Data, vol. 5, p. 984807, Oct. 2022, doi: 10.3389/fbinf.2022.984807.
[16] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-Domain K-Anonymity," in Proc. ACM SIGMOD International Conference on Management of Data, 2005, pp. 49–60.