Penerapan Mikrosoft Presidio untuk Identifikasi dan Perlindungan Informasi Sensitif pada Data Pasien

Grace Sihotang
Fakultas Informatika
Universitas Telkom
Bandung
gracesihotang@students.telkomuniversity.ac.id

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

Ryan Lingga Wicaksono
Fakultas Informatika
Universitas Telkom
Bandung
wicaksonoryan@telkomuniversity.ac.id

Pemanfaatan data pasien dalam bentuk Electronic Medical Records (EMR) memberikan banyak manfaat, terutama untuk integrasi layanan kesehatan antar lembaga. Namun, di sisi lain, keberadaan data sensitif seperti usia dan nomor kontak menimbulkan risiko privasi yang perlu ditangani dengan serius. Penelitian ini bertujuan untuk meningkatkan perlindungan data pasien melalui penerapan Mikrosoft Presidio sebagai alat identifikasi informasi sensitif, serta optimalisasi teknik anonymization berbasis local recoding. Data yang digunakan berasal dari dataset sekunder berisi informasi pasien yang memiliki entitas PII DAN quasi-identifier. Pada tahap awal, Mikrosoft Presidio digunakan untuk mendeteksi dan menyamarkan informasi sensitif. Selanjutnya, local recoding dioptimalkan menggunakan grid search dengan menggunakan nilai ambang usia dan parameter k-anonymity sebagai acuan penilaian. Hasil pengujian menunjukkan bahwa kombinasi parameter terbaik terdapat pada nilai a = 43 dan k = 2, yang memberikan k-anonymity dan I-diversity lebih tinggi dibandingkan metode pembanding, yaitu rounding. Temuan ini menunjukkan bahwa local recording yang dioptimalkan mampu memberikan perlindungan privasi yang lebih baik tanpa menghilangkan utilitas data. Penelitian ini menegaskan bahwa penerapan Mikrosoft Presidio dan optimasi anonymization dapat menjadi pendekatan efektif untuk menjaga kemananan informasi pasien di lingkungan EMR.
Kata kunci: lokal recoding, rekam medis elektronik, pencarian grid, k-anonimity, teknik anonimisasi

Utilizing patient data in the form of Electronic Medical Records (EMR) provides many benefits, especially for the integration of healthcare services between institutions. However, on the other hand, the presence of sensitive data such as age and contact numbers poses privacy risks that need to be addressed seriously. This study aims to improve patient data protection by implementing Microsoft Presidio as a tool for identifying sensitive information, as well as optimizing local recoding-based anonymization techniques. The data used comes from a secondary dataset containing patient information that has PII entities AND quasi-identifiers. In the initial stage, Microsoft Presidio is used to detect and disguise sensitive information. Next, local recoding is optimized using grid search with a threshold value of age and the k-anonymity parameter as a reference for assessment. The test results show that the best parameter combination is at a value of 43 and k = 2, which provides higher k- anonymity and I-diversity compared to the comparison method, namely rounding. These findings indicate that optimized local recording can provide better privacy protection without reducing data utility. This study confirms that the implementation of Microsoft Presidio and anonymization optimization can be an effective approach to maintaining the security of patient information in the EMR environment.
Keywords: local recording, electronic medical records, grid search, k-anonymity, anonymization techniques

1. Pendahuluan

Rekam medis pasien berisi identitas, riwayat kesehatan, dan tindakan medis yang penting untuk diagnosis
serta kesinambungan layanan kesehatan [1]. Peralihan fasilitas kesehatan menuju *Electronic Medical Record* (EMR) dan penggunaan cloud memang meningkatkan efisiensi integritas data, namun juga membuka potensi ancaman privasi ketika data sensitif tidak dilindungi dengan baik [2]. Beberapa atribut seperti usia dan nomor kontak bertindak sebagai *quasi-identifiers* sekaligus PII yang dapat memicu proses re-identifikasi sehingga membutuhkan mekanisme anonimisasi yang memadai [3]. Berbagai pendekatan telah ditawarkan untuk menjaga kerahasiaan data, seperti *interleaving encoder* [4], dan blockchain [5], tetapi metode tersebut belum sepenuhnya mengatasi isu utilitas data yang tetap dibutuhkan untuk analisis atau pembelajaran mesin. Sementara itu, *local recoding* telah diterapkan pada studi genomik [6] dan pengenalan wajah [7], namun penerapannya pada EMR masih memerlukan pengembangan lebih lanjut.

Teknik anonimisasi seperti *grid search* telah digunakan dalam berbagai penelitian, seperti pemodelan *exponential smoothing* dan *spiking neural network* [8], tetapi belum banyak dimanfaatkan untuk meningkatkan performa anonimisasi EMR. Oleh karena itu, penelitian ini berfokus pada optimalisasi *recoding* menggunakan *grid search* serta evaluasi k-anonymity, minimum l-diversity, dan average l-diversity sebagai bagian dari upaya memperkuat privasi data pasien.

Perumusan Masalah

Penelitian ini berangkat dari kebutuhan untuk meningkatkan perlindungan privasi pada data *electronic medical record* (EMR), khususnya terhadap atribut usia dan nomor kontak yang berperan sebagai *quasi- identifiers* serta *personally information* (PII). Meskipun berbagai pendekatan sebelumnya telah dilakukan, teknik yang mampu menjaga privasi sekaligus tetap mempertahankan utilitas data masih perlu dikembangkan. Oleh karena itu, penelitian ini merumuskan beberapa permasalahan utama: bagaimana proses identifikasi dan perlindungan Mikrosoft Presidio, bagaimana mengoptimalkan kinerja teknik anonimisasi local recoding agar memberikan tingkat privasi yang lebih baik, serta bagaimana menentukan kombinasi parameter threshold usia dan nilai k yang paling optimal melalui metode *grid seacrh* berdasarkan evaluasi k-anonymity. Disamping itu, penelitian ini juga

mempertanyakan bagaimana performa *local recoding* yang telah dioptimalkan dibandingkan dengan metode *rounding* melalui pengukuran minimum dan rata-rata l-diversity. Rumusan ini diperlukan untuk memastikan apakah *local recoding* yang telah dioptimalkan mampu memberikan perlindungan yang lebih baik terhadap privasi data pasien tanpa mengorbankan nilai guna data secara signifikan.

Tujuan

Penelitian ini bertujuan menghasilkan mekanisme perlindungan informasi sensitiif pada EMR dengan menilai efektivitas Mikrosoft Presidio dalam mendeteksi dan menyamarkan entitas PII serta mengoptimalkan teknik *local recoding* untuk meningkatkan kualitas anonimisasi. Proses anonimisasi diukur menggunakan k- anonymity, minimum l-diversity, dan average l-diversity, sementara optimasi parameter dilakukan melalui *grid search* untuk menentukan ambang batas usia dan nilai k yang memberikan tingkat anonimisasi optimal. Secara keseluruhan, penelitian ini berfokus pada evaluasi kekuatan anonimisasi dan kemampuan Presidio sehingga dapat menghadirkan model perlindungan privasi pasien yang lebih kuat.

Organisasi Tulisan

Naskah ini terbagi dalam beberapa segmen untuk menyajikan sebuah struktur yang jelas. Setelah Pendahuluan, Bagian II menyampaikan studi terkait yang berfokus pada privasi data kesehatan, teknik anonimisasi, dan penelitian-penelitian relevan yang relevan yang mendukung karya ini. Dalam bagian II, dijelaskan tentang metodologi penelitian yang mencakup pemanfaatan Mikrosoft Presidio untuk mendeteksi PII, penerapan *local recoding* pada data usia, dan upaya optimasi parameter melalui *grid search*. Bagian IV memuat hasil dari eksperimen dan analisis berdasarkan semua skenario pengujian yang telah dilakukan, termasuk perbandingan kinerja antara *local recoding* dan pembulatan untuk kemajuan penelitian di tahap selanjutnya.

2. Studi Terkait

Perkembangan penelitian mengenai perlindungan data medis telah meningkat seiring bertambahnya penggunaan *Electronic Medical Records* (EMR) di berbagai insitusi kesehatan. EMR menyimpan berbagai informasi pribadi dan medis pasien, sehingga memunculkan kebutuhan akan mekanisme perlindungan privasi yang memadai. Sejumlah penelitian mengusulkan pendekatan yang berbeda untuk mengamankan data medis. Pendekatan berbasis enkripsi, seperti yang dikembangkan oleh Fu et al. [4], menitikberatkan pada keamanan akses melalui mekanisme kriptografi. Sementara itu, penelitian ini mengandalkan teknologi *blockchain* untuk memastikan integritas dan keamanan distribusi data antar pihak, sebagaimana dijelaskan oleh Sankar et al. [5].

Meskipun memberikan perlindungan yang kuat, kedua pendekatan tersebut memiliki keterbatassan dalam mempertahankan utilitas data, terutama ketika data perlu dianalisis atau digunakan untuk keperluan penelitian.

Sebagai alternatif, metode anonimisasi menjadi fokus dalam berbagai studi karena mampu menjaga privasi tanpa menghilangkan kegunaan data. Local recoding merupakan salah satu teknik anonimisasi yang telah diterapkan dalam berbagai konteks. Harary et al. [6] memanfaatkan local recoding dalam bidang genomik untuk menjaga kerahasiaan data genetik yang dibagikan lintas institusi. Liu et al. [7] mengaplikasikan pendekatan serupa dalam sistem pengenalan wajah untuk melindungi fitur privat pengguna. Studi-studi tersebut menunjukkan bahwa local recoding menawarkan fleksibilitas dalam melakukan generalisasi secara lokal, namun pemilihan parameter yang tepat menjadi tantangan utama agar privasi dapat tercapai tanpa mengorbankan kualitas data.

Selain pendekatan anonimisasi, peningkatan performa metode juga sering dilakukan melalui optimasi parameter. Grid search merupakan salah satu teknik optimasi sederhana namum efektif yang telah digunakan di berbagai bidang. [8] menerapkan grid search dalam optimasi model peramalan beban listrik, sedangkan [8] memanfaatkan dalam pelatihan *spiking neural network* pada dataset EEG. Penggunaan grid search dalam konteks anonimisasi masih jarang dijelajahi, sehingga membuka peluang penelitian untuk meningkatkan performa local recoding melalui pencarian parameter optimal.

Di sisi lain, identifikasi PII pada data tidak terstruktur juga menjadi bidang yang berkembang pesat. Microsoft Presidio telah digunakan dalam beberapa penelitian sebagai alat untuk mengidentifikasi esntitas sensitif.

Secara keseluruhan, penelitian terdahulu menunjukkan adanya beberapa kesenjangan:

1. Perlunya pendekatan yang dapat menjaga privasi sekaligus mempertahankan utilitas data EMR,
2. Belum adanya optimasi local recoding menggunakan grid search pada konteks data medis, dan
3. Minimnya penelitian yang mengombinasikan pendeteksian otomatis PII melalui alat seperti Mikrosoft Presidio) dengan teknik anonimisasi berbasis metrik privasi seperti *k-anonymity* dan *l-diversity*.

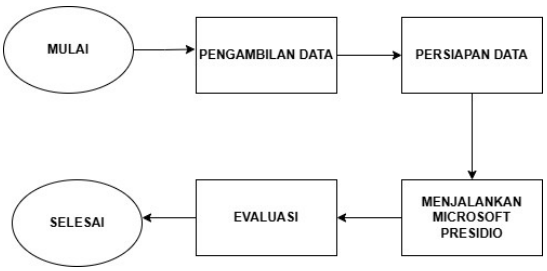
Oleh karena itu, penelitian ini menempatkan dirinya pada irisan dari ketiga arah penelitian tersebut, dengan mengintegrasikan deteksi PII, anonimisasi local recoding, dan optimasi grid search untuk memperkuat perlindungan privasi pada data pasien.

3. Sistem yang Dibangun

Penelitian ini menerapkan pendekatan terstruktur untuk mencapai sasaran yang telah ditentukan. Atribut seperti usia dan nomor telepon dipilih sebagai inti perhatian karena keduanya berfungsi sebagai sebagai quasi- identifiers dan PII dalam kumpulan data sekunder yang diambil dari Kaggle. Dalam penelitian ini, kemampuan pengkodean lokal ditingkatkan melalui metode pencarian grid, dengan k-anonymity berfungsi sebagai ukuran evaluasi. Semua kombinasi nilai yang diselidiki ditampilkan dalam bentuk peta tiga dimensi untuk menggambarkan lanskap kecocokan. Selanjutnya, hasil dair metode pengkodean lokal dibandingkan dengan pembulatan menggunakan metrik tambahan, yaitu minimum l-diversity dan rata-rata l-diversity. Proses penelitian secara keseluruhan

`

digambarkan melalui diagram proses yang terdapat di gambar 1.



GAMBAR 1
Diagram penelitian pelestarian privasi rekam medis rumah sakit.

3.1 Perlindungan Privasi pada EMR

Perlindungan privasi secara umum bertujuan untuk menjaga kerahasiaan data pribadi yang dilakukan publikasi, tanpa mengurangi manfaat data tersebut untuk analisis lebih mendalam [9]. EMR itu sendiri adalah kumpulan data kesehatan pasien yang bisa diakses secara umum [10]. Dalam hal ini, *quasi-identifiers* adalah informasi yang tampak tidak berbahaya, tetapi bisa digunakan untuk melacak identitas individu jika dipadukan dengan informasi lainnya [11]. Di sisi lain, PII adalah informasi yang secara langsung menandakan idenntitas sesorang sehingga memerlukan perlindungan yang lebih ketat [2]. Serangan privasi pada EMR bisa berupa serangan inferensi, contohnya adalah usaha untuk mendapatkan atribut sensitif dari rekaman suara pasien [12].

Dataset yang dipakai dalam penelitian ini diperoleh dari platform Kagglle, yaitu file *patients.csv*. dataset ini terdiri dari 50 baris dan 14 kolom yang mencakup berbagai informasi identitas seperti nama, alamat, tanggal lahir, nomor identitas, dan atribut lainnya [11]. Dari daftar kolom tersebut, atribut usia dianggap sebagai quasi-identifier, sedangkan nomor kontak dikategorikan sebagai PII. Mengingat keduanya rentan terhadap pelanggaran privasi, penerapan alat perlindungan seperti Mikrosoft Presidio menjadi sangat penting [13].

3.2 Anonimisasi dengan Mekanisme Local Recoding

Anonimisasi merupakan metode untuk meningkatkan privasi (PET) yang menyembunyikan informasi pribadi dengan cara menggeneralisasikan data. Sebagai contoh,

TABEL 1
Studi terkait mengenai Perlindungan Privasi pada EMR.

Referensi	**EMR Privacy**	**Local Recoding**	**Grid Search**
Saxena et al. [8]	✓	✗	✗
Ganadily et al. [10]	✓	✗	✗
Akram et al. [10]	✗	✓	✗
Yang et al. [11]	✗	✓	✗
Metode Usulan	✓	✓	✓

Anonimisasi adalah salah satu metode dalam teknologi yang meningkatkan privasi *privacy-enhancing technology* (PET) yang bertujuan untuk menyembunyikan data pribadi melalui proses generalisasi nilai-nilai data [8]. Contohnya, apabila sebuah database menyimpan alamat lengkap seseorang seperti nomor rumah dan nama jalan, anonimisasi dilakukan denga mengganti informasi tersebut menjadi tingkat lokasi yang lebih umum seperti desa atau kecamatan. Pendekatan ini memberikan perlindungan terhadap data sensitif sekaligus tetap menjaga kemanfaatan data untuk analisis yang diperlukan [3].

Dalam penelitian ini, *local recoding* diterapkan sebagai mekanisme utama untuk anonimisasi. Tidak seperti generalisasi global yang menerapkan aturan yang sama pada seluruh dataset, *local recoding* memungkinkan setiap bagian data untuk digeneralisasi dengan aturan yang berbeda berdasarkan konteksnya [10]. Setiap record t ditempatkan dalam sebuah kelas ekuivalensi C(t). Apabila jumlah anggota dlam kelas tersebut kurang dari nilai ambang k, maka record tersebut akan digeneralisasi dengan menggunakan fungsi G(t). Formula mekanismenya disajikan pada Persamaan (1):

Pendekatan ini menawarkan kebebasan dalam melindungi privasi, karena hanya elemen data yang tidak memenuhi k-anonimity yang akan mengalami generalisasi, sehingga meminimalkan kehilangan kegunaan data.

Sebagai metode pembanding, penelitian ini juga menganalisis teknik generalisasi numerik berupa *rounding*. Metode ini menurunkan tingkat keakuratan data angka untuk menyamarkan nilai-nilai asli tanpa menghilangkan informasi sepenuhnya [3]. Misalnya, usia 35 dan 37 dapat dibulatkan menjadi 30. Meskipun tampak sederhana, pembulatan memberikan referensi penting dalam mengevaluasi konsistensi dan efektivitas *local recoding.*

Selain generalisasi, penelitian ini menerapkan teknik masking pada atribut nomor telepon, yang merupakan bagian dari PII. Masking bekerja dengan menyembunyikan sebagian karakter dalam sebuah string sehingga nilai aslinya tidak dapat dikenali, tetap menjaga beberapa karakter awal untuk memastikan utilitas data [8]. Penelitian ini menggunakan *partial redaction,* yaitu teknik masking yang mempertahankan n karakter dan mengganti karakter lainnya dengan simbol penanda. Rumus masking dijelaskan pada Persamaan (2):

Dimana nilai p ditentukan melalui optimasi [11] Penerapan *local recoding, rouding,* dan masking secara terintegrasi memberikan pendekatan anonimisasi yang lebih adaptif, dengan tujuan mencapai keseimbangan optimal antara privasi dan utilitas data EMR.

3.3 Optimasi Grid Search dengan Evaluasi K-Anonymity

Grid search adalah teknik optimasi mendasar berfungsi dengan menciptakan semua kombinasi hyperparameter yang telah ditetapkan sebelumnya. Pendekatan ini bersifat menyeluruh, karena tidak melibatkan pemilihan nilai secara acak atau melewati kombinasi tertentu. Kombinasi hyperparameter yang dianggap paling sesuai adalah yang menghasilkan nilai fungsi *fitness* terbaik yang berhasil dicapai [13]. Dalam studi ini, dua hyperparameter utama yang diuji adalah batas usia dan nilai k sebagai ambang untuk pembentukan kelas ekuivalensi C(t) dalam mekanisme *local recoding.* Kedua parameter ini menentukan apakah suatu record harus digeneralisasi dengan menggunakan fungsi G(t).

Untuk mengevaluasi hasil dari proses anonimisasi, studi ini

`

memanfaatkan metrik *k-anonimity* sebagai ukuran fitness. K-anonymity menjamin bahwa setiap record dalam dataset disertai minimal k record lain yang memiliki pola atribut serupa. Misalnya, sebuah dataset dianggap memiliki k-anonymity sebesar 3 jika setiap entri data disertai oleh dua entri lain yang tidak dapat dibedakan. Nilai k-anonymity yang lebih tinggi berarti perlindungan privasi terhadap data tersebut menjadi kuat.

GAMBAR 2
Metrik Privasi untuk Evaluasi Anonimisasi

Algorithm 1 Grid Search for Local Recoding Optimization

Require: D, A, K, d, $k \in K$, $a \in A$
Ensure: $a*$, $k*$
Initialize results list $R \leftarrow \emptyset$
for each a in A **do**
  for each k in K **do**
    Compute $C(t)$ in D using quasi-identifiers (age, contact number) with Equation 1
    $m \leftarrow \min_{c \in C} |c|$ ▷ Smallest group size
    if $m < k$ **then**
      $t' \leftarrow G(t) \cup M(s)$ ▷ from Equation 1 and Equation 2
    else ▷ No recoding needed
      $t' \leftarrow t$
    end if
    $k_{\text{local}} \leftarrow$ Compute k-anonymity of D
    Append (a, k) to P
    Append k_{local} to K_{local}
  end for
end for
The following equation obtains the optimum result:

$$(a*, k*) = \arg\max_{(a,k) \in P} K_{\text{local}} \quad (3)$$

Proses optimasi grid search yang digunakan diilustrasikan melalui Algoritma 1. Algoritma ini memanfaatkan beberapa hyperparameter penting. Dataset asli direpresentasikan sebagai D, sedangkan d menunjukkan jumlah digit nomor telepon yang masih dipertahankan setelah proses masking. Himpunan K berisi daftar nilai k yang akan diuji sebagai ambang pembentukan kelas ekuivalensi, sementara A merupakan himpunan nilai batas usia yang menjadi kandidat generalisasi. Hasil keluaran dari proses ini adalah nilai dari ambang usia optimal *a** dan nilai *k** yang digunakan dalam fungsi C(t).

Setiap perpaduan hyperparameter (a, k) dianalisis dengan teliti melalui perhitungan nilai *k-anonymity* untuk dataset yang telah melalui tahap generalisasi sesuaikan dengan ketentuan *local recoding*. Nilai-nilai yang diperoleh selanjutnya disimpan dan dibandingkan untuk mencari kombinasi paling efisien. Kombinasi parameter yang paling unggul ditentukan berdasarkan nilai k-anonymity tertinggi yang didapat dari seluruh iterasi grid search.

Selain evaluasi yang didasarkan pada k-anonymity, studi ini juga melakukan perbandingan kinerja *local recoding* dengan teknik *rouding* dengan menggunakan metrik *l-diversity*. Berbeda dengan k-anonymity yang hanya fokus pada keseragaman atribut quasi-identifier, l-diversity mengevaluasi variasi nilai atribut sensitif pada setiap kelompok ekuivalensi [8]. Penelitian ini menerapkan *minimum l-diversity* untuk mendeteksi kelas dengan tingkat privasi paling rendah serta *average l-diversity* untuk menilai tingkat privasi secara keseluruhan [9]. Ringkasan tentang metrik yang diterapkan dalam analisis ini dapat dilihat pada tabel II.

TABEL 2
Metrik Privasi untuk Evaluasi Anonimisasi

Nama	Definisi	Makna
k-anonymity	Ukuran minimum dari sebuah kelompok (kelas ekuivalensi) yang memiliki nilai quasi-identifier yang sama .	**Tinggi** : Perlindungan privasi lebih kuat karena setiap individu tersembunyi diantara lebih banyak orang. **Rendah** : Kelompok kecil meningkatkan risiko re-identifikasi.
l-diversity	Jumlah minimum nilai sensitif yang berbeda di dalam setiap kelas ekuivalensi.	**Tinggi** : Perlindungan lebih baik terhadap pengungkapan atribut; informasi pribadi lebih sulit ditebak. **Rendah** : Nilai sensitif dapat ditebak meskipun k-anonymity terpenuhi.

4. Evaluasi

4.1 Hasil Pengujian

Pengujian pertama dilakukan untuk mengevaluasi kemampuan Mikrosoft Presidio dalam menyembunyikan PII pada data EMR. Hasil pada tabel III menunjukkan bahwa identitas sensitif berhasil disamarkan, meskipun pola penyamaran yang homogen dapat mengurangi variasi data untuk kebutuhan analisis lanjutan

Pengujian berikutnya berfokus pada optimasi *local recoding* menggunakan *grid search* terhadap dataset EMR dari Kaggle. Dua parameter diuji, yaitu batas usia (40-44) dan nilai k (0-10), dengan k-anonymity sebagai fungsi evaluasi. Hasil visualisasi pada Gambar 3 menunjukkan kombinasi paling optimal pada batas usia 43 dan k=2.

Perbandingan dengan metode *rounding* menunjukkan bahwa *local recoding* yang dioptimasi mampu menghasilkan k-anonymity lebih tinggi dibandingkan *rounding,* yang hanya mencapai nilai 2. Nilai minimum l-diversity pada *local recoding* juga lebih baik (2 dibandingkan 1), menunjukkan risiko re-identifikasi yang lebih rendah.

Selain itu, average l-diversity pada *local recoding* (3.12) lebih tinggi dibanding *rounding* (1.95). perbedaan nilai rata-rata dan nilai minimum yang cukup jauh mengindikasikan bahwa sebagian besar kelas ekuivalensi pada *local recoding* pada *local recoding* memberikan perlindungan yang lebih kuat.

TABEL 3
Contoh Anonimisasi oleh Mikrosoft Presidio pada EMR.

Teks asli	Setelah Disamarkan
Nama pasien adalah John Smith, lahir pada Januari 1990	Nama pasien adalah [PERSON], lahir pada [DATE TIME]
Nomor SSN pasien adalah 123-45-6789, dan pasien tinggal di Boston	Nomor pasien adalah [US_SSN] dan pasien tinggal di [LOCATION]

4.2 Analisis Hasil Pengujian

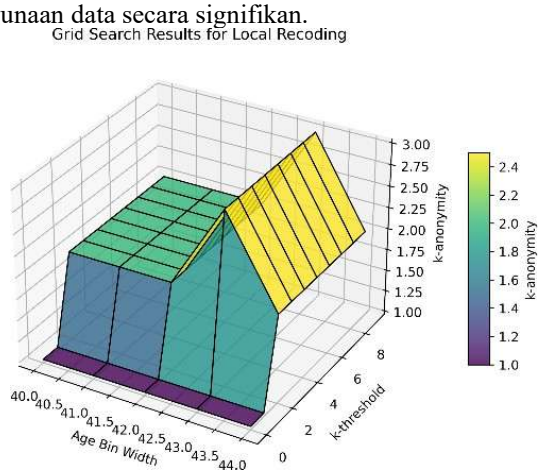
Analisis ini menjelaskan temuan dari keseluruhan rangkaian uji coba yang telah dilakukan, yang meliputi proses awal anonimisasi dengan menggunakan Mikrosoft Presidio, optimasi local recoding dengan grid search, serta perbandingan kinerja local recoding terhadap metode pembandingnya, yaitu rounding. Setiap analisis ditujukan untuk menilai tingkat privasi yang dihasilkan, terutama melalui metrik k-anonymity dan l-diversity. Uji coba awal dengan Mikrosoft Presidio menunjukkan bahwa alat ini mampu menyamarkan OII secara efektif menggunakan label-label tertentu, sehingga informasi sensitif dapat terlindungi sepenuhnya. Namun, penyamaran yang seragam mengakibatkan berkurangnya keragaman data. Situasi ini menimbulkan konsekuensi bahwa meskipun privasi terjaga, kegunaan data untuk analisis lebih lanjut atau pelatihan model bisa berkurang karena setiap nilai sensitif digeneralisasi menjadi token identik. Fenomena ini memperkuat pentingnya pendekatan anonimisasi yang dapat menjaga utilitas data sekaligus tidak mengorbankan tingkat
perlindungan privasi.

Hasil dari pengujian optimasi local recoding melalui grid search menunjukkan cara hubungan antara parameter ambang usia dan nilai k sebagai indikator untuk menentukan generalisasi pada data. Visualisasi dari lanskap fitness memperlihatkan bahwa nilai k=1, maka k-anonymity yang dihasilkan adalah terendah, sedangkan kombinasi tertentu dari parameter a dan k dapat secara signifikan mendongkrak nilai k-anonymity. Dari hasil pencarian, didapatkan bahwa konfigurasi yang paling optimal terletak pada a=43 dan k=2, yang memberikan keseimbangan terbaik antara kegunaan dan privasi. Ini menunjukkan bahwa pemilihan ambang generalisasi yang sesuai sangat memengaruhi ukuran kelompok ekuivalensi yang dihasilkan.

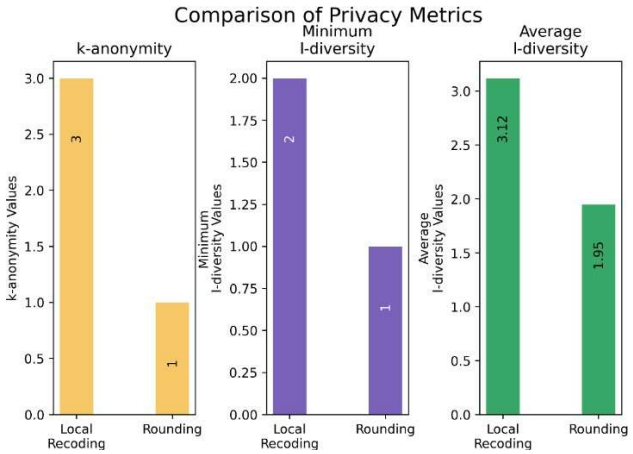
Selanjutnya, perbandingan antara local recoding dan rounding memperlihatkan bahwa local recoding memberikan performa privasi yang lebih baik, baik dari aspek k-anonymity maupun l-diversity. Rounding menghasilkan nilai k-anonymity sebesar 2, yang berarti generalisasi angka melalui pembulatan memiliki keterbatasan karena tidak membentuk kelompok ekuivalensi yang cukup besar. Sebaliknya, local recoding yang telah dioptimalkan mampu meningkatkan k-anonymity, sehingga risiko identifikasi ulang menjadi lebih rendah.

Dari sisi l-diversity, local recoding juga menunjukkan keunggulan. Nilai minimum l-diversity untuk local recoding adalah 2, sedangkan rounding hanya menghasilkan nilai 1, yang menunjukkan bahwa ada kelompok ekuivalensi yang kurang memiliki variasi dalam data sensitif. Keadaan ini berisiko karena individu dalam kelompok tersebut masih mungkin diidentifikasi informasinya meskipun secara teknis sudah memenuhi syarat k—anonymity. Di sisi lain, nilai rata-rata l-diversity yang diperoleh melalui local recoding lebih tinggi dibandingkan dengan pembulatan, sekali lagi membuktikan bahwa pendekatan ini lebih efektif dalam menjaga keragaman informasi sensitif di setiap kategori ekuivalensi.

Secara keseluruhan, pembahasan ini menunjukkan bahwa local recoding, terutama setelah jalur optimasi menggunakan grid search, lebih baik dalam melindungi privasi data medis jika dibandingkan dengan metode penyamaran awal atau teknik pembulatan. Strategi ini tidak hanya memperbaiki nilai k-anonymity, tetapi juga meningkatkan struktur l-diversity, yang sangat penting untuk memastikan keamanan atribut sensitif. Temuan ini menunjukkan bahwa local recoding adalah metode yang lebih efisien untuk situasi anonimisasi EMR yang memerlukan perlindungan yang kuat tanpa mengorbankan kegunaan data secara signifikan.



GAMBAR 3
Lanskap fitness hasil optimasi grid search



GAMBAR 4
Perbandingan kinerja antara local recoding dan rounding

5. Kesimpulan

Penelitian ini menunjukkan bahwa penerapan anonimisasi dengan menggunakan Mikrosoft Presidio dan mekanisme local recoding yang dioptimalkan menggunakan grid search mampu memperkuat perlindungan privasi pada data EMR. Presidio terbukti dapat secara efisien menyamarkan informasi peribadi yang dapat diidentifikasi, meskipun keluaran yang terlalu konsisten dapat mengurangi variasi data. Optimasi local recoding menghasilkan kombinasi parameter yang terbaik, age threshold 43 dan nilai k=2 yang memberikan nilai k-anonymity lebih tinggi dibandingkan parameter lainnya. Dibandingkan metode pembulatan (rounding), local recoding menunjukkan performa privasi yang lebih baik, ditunjukkan oleh nilai k-anonymity, minimum l-diversity, dan average l-diversity yang lebih tinggi. Secara keseluruhan, metode yang diterapkan dalam penelitian ini efektif untuk meningkatkan

privsai data medis tanpa menghilangkan utilitas data secara signifikan.

REFERENSI

[1] M. T. Neame *et al.*, "Successful integration of an automated patient-reported outcome measure within a hospital electronic patient record," *Rheumatol. Adv. Pract.*, vol. 6, no. 3, p. rkac065, Sept. 2022, doi: 10.1093/rap/rkac065.

[2] K.-H. Nguyen *et al.*, "Cashing in: cost-benefit analysis framework for digital hospitals," *BMC Health Serv. Res.*, vol. 24, no. 1, p. 694, May 2024, doi: 10.1186/s12913-024-11132-7.

[3] J. Andrew, R. J. Eunice, and J. Karthikeyan, "An anonymization-based privacy-preserving data collection protocol for digital health data," *Front. Public Health*, vol. 11, p. 1125011, Mar. 2023, doi: 10.3389/fpubh.2023.1125011.

[4] J. Fu, N. Wang, and Y. Cai, "Privacy-Preserving in Healthcare Blockchain Systems Based on Lightweight Message Sharing," *Sensors*, vol. 20, no. 7, p. 1898, Mar. 2020, doi: 10.3390/s20071898.

[5] SunithaBJ, K. Sankar, A. Ayesha, and M. Islabudeen, "Different Approaches on Security, Privacy and Efficient Sharing of Electronic Health Records Using Blockchain Technology," in *Advances in Parallel Computing*, D. J. Hemanth, T. N. Nguyen, J. Indumathi, and S. Lakshmanan, Eds., IOS Press, 2022. doi: 10.3233/APC220007.

[6] M. Harary, "SHIELD: Secure Haplotype Imputation Employing Local Differential Privacy," 2023, *arXiv*. doi: 10.48550/ARXIV.2309.07305.

[7] Y. Liu *et al.*, "Local Features Meet Stochastic Anonymization: Revolutionizing Privacy-Preserving Face Recognition for Black-Box Models," 2024, *arXiv*. doi: 10.48550/ARXIV.2412.08276.

[8] D. Penkov, P. Koprinkova-Hristova, N. Kasabov, S. Nedelcheva, S. Ivanovska, and S. Yordanov, "Grid Search Optimization of Novel SNN-ESN Classifier on a Supercomputer Platform," in *Large-Scale Scientific Computations*, vol. 13952, I. Lirkov and S. Margenov, Eds., in Lecture Notes in Computer Science, vol. 13952. , Cham: Springer Nature Switzerland, 2024, pp. 435–443. doi: 10.1007/978-3-031-56208-2_45.

[9] M. A. Rafif Tarigan, A. G. Putrada, and R. L. Wicaksono, "ElGamal Homomorphic Encryption with SMOTE for PET in Occupancy Monitoring by XGBoost," in *2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)*, Bandung, Indonesia: IEEE, Feb. 2025, pp. 1–6. doi: 10.1109/ICADEIS65852.2025.10933291.

[10] N. Alrebdi, A. Alabdulatif, C. Iwendi, and Z. Lian, "SVBE: searchable and verifiable blockchain-based electronic medical records system," *Sci. Rep.*, vol. 12, no. 1, p. 266, Jan. 2022, doi: 10.1038/s41598- 021-04124-8.

[11] S. Prabowo, A. G. Putrada, I. D. Oktaviani, and M. Abdurohman, "Camera-Based Smart Lighting

System that complies with Indonesia's Personal Data Protection Act," in *2023 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)*, Bali, Indonesia: IEEE, Aug. 2023, pp. 1–6. doi: 10.1109/ICADEIS58666.2023.10271086.

[12] T. Feng, R. Peri, and S. Narayanan, "User-Level Differential Privacy against Attribute Inference

Attack of Speech Emotion Recognition in Federated Learning," 2022, doi: 10.48550/ARXIV.2204.02500.

[13] S. Prabowo *et al.*, "Privacy-Preserving Tools and Technologies: Government Adoption and Challenges," *IEEE Access*, vol. 13, pp. 33904–33934, 2025, doi: 10.1109/ACCESS.2025.3540878.

`