**INDEXIA: Informatic and Computational Intelligent Journal**
Vol.07, No.02, November 2025, hlm. 110-119
P-ISSN: 2657-0424
E-ISSN: 2657-0432

# Prediksi Jenis Kebocoran Data Kesehatan di AS Berdasarkan Laporan HIPAA Menggunakan LightGBM dan Kerangka OSEMN

**Noeni Indah Sulistiyani[1,*], Raihan Ade Purnomo[2)], Nasya Rohmatunisa[3)], Renjiro Maheswara Pujo[4)], Rizal Broer Bahaweres[5)], Nashrul Hakiem[6)]**

[1, 2, 3, 4, 5,6)] Prodi Teknik Informatika, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta
Jl. Ir H. Juanda No.95, Ciputat, Kec. Ciputat Tim., Kota Tangerang Selatan, Banten 15412
e-mail: noeni.indahs22@mhs.uinjkt.ac.id[1)], raihan.adepurnomo22@mhs.uinjkt.ac.id[2)], nasya.rohmatunisa22@mhs.uinjkt.ac.id[3)], renjiro.pujo22@mhs.uinjkt.ac.id[4)], rizalbroer@uinjkt.ac.id[5)], hakiem@uinjkt.ac.id[6)]
* corresponding author

(Naskah masuk: 09 Juli 2025 Diterima untuk diterbitkan: 28 Juli 2025)

**ABSTRAK**

*Meningkatnya insiden kebocoran data pada sektor kesehatan di Amerika Serikat mendorong perlunya analisis komprehensif terhadap pola, tren, dan prediksi jenis pelanggaran yang terjadi. Penelitian ini menganalisis 1654 laporan pelanggaran data berdasarkan publikasi resmi HIPAA dari tahun 2009 hingga 2016. Dengan pendekatan kerangka kerja OSEMN (Obtain, Scrub, Explore, Model, and Interpret), dilakukan eksplorasi data deskriptif dan text mining untuk mengidentifikasi distribusi insiden berdasarkan waktu, lokasi geografis, jenis entitas, serta lokasi penyimpanan data. Selanjutnya, model prediksi jenis pelanggaran data (Type of Breach) dibangun menggunakan algoritma Light Gradient Boosting Machine (LightGBM) yang dikombinasikan dengan teknik preprocessing, one-hot encoding, SMOTE untuk penyeimbangan kelas, dan tuning hyperparameter melalui GridSearchCV. Evaluasi menggunakan F1-score macro menunjukkan bahwa model mampu melakukan klasifikasi multi-kelas dengan performa baik, khususnya pada kelas mayoritas. Temuan ini memberikan kontribusi penting dalam pemahaman risiko keamanan informasi kesehatan dan menjadi dasar pengembangan sistem deteksi dini berbasis data historis.*

***Kata Kunci:*** *HIPAA, LightGBM, OSEMN, Prediksi Pelanggaran, Text Mining.*

***ABSTRACT***

*The increasing incidence of data breaches in the United States healthcare sector necessitates a comprehensive analysis of the patterns, trends, and predictions of breach types. This study analyzes 1654 data breach reports based on official HIPAA publications from 2009 to 2016. Utilizing the OSEMN (Obtain, Scrub, Explore, Model, and Interpret) framework, descriptive data exploration and text mining were conducted to identify the distribution of incidents by time, geographical location, entity type, and data storage location. Subsequently, a predictive model for the type of data breach was built using the Light Gradient Boosting Machine (LightGBM) algorithm combined with preprocessing techniques, one-hot encoding, SMOTE for class balancing, and hyperparameter tuning via GridSearchCV. Evaluation using the macro F1-score demonstrated that the model is capable of performing multi-class classification with good performance, particularly for majority classes. These findings provide a significant contribution to understanding healthcare information security risks and serve as a basis for developing early detection systems based on historical data.*

***Keywords****: HIPAA, LightGBM, OSEMN, Breach Prediction, Text Mining.*

## I. PENDAHULUAN

Transformasi digital yang pesat di sektor kesehatan telah membawa efisiensi dan inovasi dalam penyediaan layanan medis, namun di sisi lain juga menghadirkan tantangan signifikan terkait keamanan informasi kesehatan yang dilindungi (*Protected Health Information*/PHI). PHI, yang mencakup data demografi, nomor Jaminan Sosial, dan informasi klinis seperti diagnosis medis, tersimpan baik dalam rekam medis kertas maupun elektronik atau disebut sebagai *Electronic Health Record* (EHR) [1]. Sifat sensitif dan nilai tinggi dari data ini menjadikannya target utama bagi para pelaku kejahatan siber. Di Amerika Serikat, keamanan PHI diatur secara ketat oleh *Health Insurance Portability and Accountability A*ct (HIPAA) tahun 1996, yang kemudian diperkuat oleh *Health Information Technology for Economic and Clinical Health* (HITECH) tahun 2009 dan *Omnibus Final Rule* tahun 2013 [3], [4]. Regulasi ini mewajibkan entitas yang tercakup untuk melaporkan setiap insiden kebocoran yang

melibatkan penggunaan atau pengungkapan informasi kesehatan yang tidak sah, terutama yang memengaruhi 500 individu atau lebih [4].

Laporan kebocoran data yang diamanatkan oleh HIPAA dan dipublikasikan secara daring oleh *Office for Civil Rights* (OCR) Departemen Kesehatan dan Layanan Kemanusiaan AS (HHS), menyediakan sumber daya yang sangat berharga untuk memahami lanskap ancaman yang terus berkembang [4]. Dataset publik ini mencatat insiden-insiden yang memengaruhi setidaknya 500 individu sejak tahun 2009 hingga Desember 2016, dengan lebih dari 860 entri yang merinci nama entitas yang terdampak, negara bagian, jenis entitas yang tercakup, tanggal pengajuan pelanggaran, jenis pelanggaran, lokasi informasi yang diretas, keberadaan mitra bisnis, dan deskripsi insiden. Meskipun demikian, penelitian komprehensif yang secara sistematis menganalisis pola-pola kebocoran data dalam konteks laporan HIPAA masih terbatas, terutama yang mencakup periode data yang ekstensif dan memanfaatkan teknik analisis data yang beragam [1], [2].

Studi-studi terbaru memperlihatkan bahwa masifnya adopsi EHR secara nasional juga berdampak pada tren peningkatan insiden kebocoran data kesehatan. Hossain dan Hong [13] mencatat bahwa dari tahun 2010 hingga 2018 tercatat 2.529 insiden kebocoran data kesehatan di Amerika Serikat yang memengaruhi hampir 195 juta individu. Sebagian besar insiden melibatkan penyedia layanan kesehatan, dengan tipe pelanggaran terbesar adalah pencurian dan serangan siber (hacking/IT incident). Bahkan, kasus berskala besar (>1 juta data) mayoritas disebabkan kompromi pada sistem internal atau jaringan server [13]. Hal ini menunjukkan pentingnya penguatan mekanisme perlindungan data serta evaluasi berkelanjutan terhadap kebijakan keamanan di era digital.

Selain itu, integrasi sistem kesehatan dengan teknologi web juga membuka celah baru kebocoran PHI melalui web tracking. Huo et al. [14] menemukan bahwa sekitar 14% portal pasien daring di Amerika Serikat secara tidak sengaja membocorkan PHI melalui penggunaan web tracker pihak ketiga seperti Google Analytics dan Facebook Pixel, yang tidak hanya mencatat aktivitas pengunjung namun juga berpotensi membocorkan identitas, hasil laboratorium, hingga jadwal kunjungan pasien. Studi ini menyoroti bahwa sebagian besar operator kesehatan tidak menyadari besarnya risiko dari integrasi skrip pihak ketiga tersebut, serta menunjukkan intervensi vendor terbukti lebih efektif daripada sekadar pemberitahuan kepada institusi [14].

Berbagai pendekatan telah dikembangkan untuk menganalisis dan mengklasifikasikan insiden pelanggaran data dalam sistem layanan kesehatan. Nazurah et al. [7] membangun sistem prediktif kesehatan menggunakan *pipeline* OSEMN (*Obtain*, *Scrub*, *Explore*, *Model, and Interpret*) dan machine learning. Koczkodaj et al. [8] menerapkan *text mining* berbasis LDA pada catatan pelanggaran data kesehatan, dengan temuan utama bahwa serangan siber menjadi penyebab dominan dalam beberapa tahun terakhir. Raghupathi et al. [9] menggunakan visualisasi data untuk menganalisis distribusi pelanggaran, sedangkan Reddy et al. [10] menyoroti peningkatan insiden ransomware dan perlunya incident response plan. Aljawarneh et al. [11] membandingkan berbagai metode klasifikasi dalam deteksi serangan siber dan menekankan pentingnya *ensemble model* pada distribusi kelas tidak merata.

Penggunaan LightGBM sebagai algoritma utama dalam penelitian ini berakar dari pengembangan oleh Ke et al. [12], yang memperkenalkan teknik *Gradient-based One-Side Sampling* (GOSS) dan *Exclusive Feature Bundling* (EFB) untuk meningkatkan efisiensi dan akurasi. Studi lain membuktikan LightGBM efektif pada klasifikasi berskala besar dengan fitur kategorikal yang kompleks [15]. Namun, aplikasi spesifiknya pada prediksi pelanggaran data kesehatan berbasis laporan HIPAA masih jarang ditemukan di literatur. Dengan demikian, penelitian ini mengisi celah tersebut dengan membangun *pipeline* supervised learning berbasis LightGBM yang mencakup seluruh tahap preprocessing, penyeimbangan data, pelatihan model, evaluasi, serta interpretasi hasil untuk memprediksi jenis pelanggaran data secara akurat dan efisien.

## II. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dalam analisis prediksi jenis pelanggaran data (*data breach*) yang terjadi pada sektor kesehatan di Amerika Serikat. Pendekatan utama yang digunakan adalah kerangka kerja OSEMN yang diperkenalkan oleh Hilary Mason dan Chris Wiggins sebagai salah satu taksonomi paling praktis dalam ilmu data modern [5]. OSEMN dipilih karena memiliki struktur tahapan kerja yang jelas, mencakup seluruh proses dari pengumpulan hingga interpretasi data.

### A. Bahan Penelitian

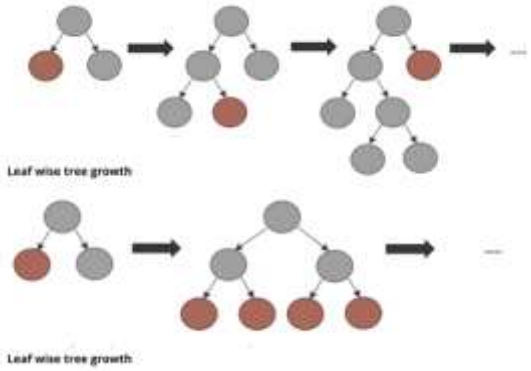
Data yang digunakan dalam penelitian ini diperoleh dari situs *Kaggle*, dengan judul dataset "Major US Health Data Breaches". Dataset ini berisi laporan insiden pelanggaran data pada sektor kesehatan di Amerika Serikat yang

dilaporkan kepada Department of Health and Human Services (HHS) melalui sistem HIPAA selama periode 2009 hingga 2016. Data dikumpulkan dalam format *CSV* dan terdiri atas 10 kolom atribut penting seperti *Name of Covered Entity, State, Covered Entity Type, Breach Submission Date, Type of Breach, Location of Breached Information, Individuals Affected, Web Description, Business Associate Present, dan Webpost Date.* Dataset ini dipilih karena representatif terhadap karakteristik pelanggaran data sektor kesehatan di AS, serta bersifat terbuka dan terpercaya untuk dianalisis secara ilmiah.

Selanjutnya dilakukan pembersihan dan filltering data untuk menghilangkan *missing values* yang ada. Jumlah dataset setelah melalui proses pembersihan dan filltering terdiri atas 1.654 baris data dengan tujuh kategori kelas target. Untuk mengatasi ketidakseimbangan kelas, digunakan teknik *Synthetic Minority Oversampling Technique* (SMOTE), sehingga setelah proses penyeimbangan, masing-masing kelas memiliki 912 sampel dan total data menjadi 6.384 baris (7 kelas × 912 sampel per kelas).

B. Tahapan Penelitian Menggunakan OSEMN

1. *Obtain*
   Tahap pertama adalah pengambilan data (*obtain*), yaitu proses mengunduh dan mengekstrak dataset dari platform *Kaggle* menggunakan API berbasis Python. File *CSV* dimuat ke dalam lingkungan Google Colaboratory menggunakan library pandas, kemudian dikonversi ke format *DataFrame* untuk memudahkan pemrosesan selanjutnya. Proses pengambilan data dilakukan dengan mempertimbangkan validitas sumber dan keberlanjutan format data mentah yang dibutuhkan [5].
2. *Scrub*
   Setelah data diperoleh, tahap selanjutnya adalah pembersihan (*scrub*). Pada tahap ini, dilakukan berbagai proses praproses data, termasuk menghapus nilai kosong (missing values) dan baris duplikat, menyamakan format teks menggunakan string normalization, serta mengonversi format tanggal ke format datetime. *Outlier* pada kolom numerik seperti *Individuals Affected* dianalisis dengan metode IQR (Interquartile Range), dan apabila perlu, dilakukan transformasi untuk mengurangi bias. Proses ini penting agar model statistik yang dibangun di tahap selanjutnya dapat berjalan dengan data yang bersih dan konsisten [5].
3. *Explore*
   Tahap eksplorasi data (*explore*) dilakukan untuk memahami distribusi dan pola dalam dataset. Visualisasi data dilakukan terhadap variabel seperti *Type of Breach, Covered Entity Type, dan Location of Breached Information*. Analisis tren dilakukan terhadap tahun pelanggaran untuk melihat dinamika kejadian dari waktu ke waktu. Selain itu, dilakukan pula eksplorasi berbasis teks (*text mining*) terhadap atribut *Web Description* untuk mengidentifikasi kata kunci dan pola topik menggunakan teknik seperti *WordCloud* dan *Latent Dirichlet Allocation* (LDA). Hasil eksplorasi ini menjadi landasan awal dalam memahami kemungkinan hubungan antar fitur yang akan dimodelkan.
4. Model
   Tahap pemodelan bertujuan untuk membangun sistem prediksi terhadap jenis pelanggaran data (*Type of Breach*) berdasarkan fitur-fitur penting dalam laporan insiden. Fitur yang digunakan mencakup lokasi negara bagian (*State*), jenis entitas yang terlibat (*Covered Entity Type*), keterlibatan pihak ketiga (*Business Associate Present*), serta jumlah individu yang terdampak (*Individuals Affected*). Metode yang digunakan adalah *Light Gradient Boosting Machine* (LightGBM), yaitu algoritma gradient boosting decision tree yang dirancang untuk efisiensi dan kecepatan dalam klasifikasi multi-kelas berskala besar [6].
   Pada dasarnya, LightGBM membangun model dengan menggabungkan banyak pohon keputusan secara bertahap (iteratif), di mana setiap pohon baru yang dibuat bertujuan untuk memperbaiki kesalahan prediksi pohon-pohon sebelumnya.
   Salah satu keunggulan utama LightGBM dibandingkan metode boosting tradisional adalah pendekatan leaf-wise tree growth, yaitu pertumbuhan pohon difokuskan pada daun (leaf) dengan penurunan loss terbesar. Dengan strategi ini, pohon dapat tumbuh lebih dalam di cabang-cabang yang paling kompleks, sehingga mampu menangkap pola data yang sulit. Namun, untuk menghindari *overfitting*, parameter seperti max_depth tetap diatur sesuai kebutuhan.
   Secara arsitektur, setiap iterasi pada LightGBM menghasilkan pohon keputusan baru yang dioptimalkan berdasarkan gradien residual dari prediksi sebelumnya. Model akhir merupakan kombinasi seluruh pohon yang dibangun selama proses pelatihan, sehingga prediksi yang dihasilkan lebih akurat dan robust terhadap data berskala besar serta fitur kategorikal.




Gambar 1. Model LightGBM

Sebelum pelatihan model, data terlebih dahulu diproses dalam beberapa tahapan. Pertama, dilakukan pembersihan data dengan menghapus baris yang memiliki nilai kosong pada kolom fitur maupun target. Variabel kategorikal kemudian diubah menjadi representasi numerik menggunakan teknik *one-hot encoding* pada fitur, dan *label encoding* pada target. Untuk memastikan performa model yang adil terhadap semua kelas, dilakukan penyaringan terhadap label yang sangat jarang muncul (≤1 sampel), sehingga hanya label yang representatif yang dilibatkan dalam pelatihan.

Untuk membangun model prediksi tipe pelanggaran data, penelitian ini menerapkan algoritma LightGBM yang dikenal efisien dalam mengelola data skala besar dengan fitur kategorikal yang kompleks, serta didukung teknik GOSS dan EFB [15].

Karena distribusi kelas yang tidak seimbang, teknik SMOTE digunakan untuk meningkatkan representasi kelas minoritas. SMOTE secara sintetis menciptakan data baru dari sampel kelas minor sehingga distribusi label menjadi lebih seimbang. Selanjutnya, fitur numerik distandarisasi menggunakan *StandardScaler* untuk meningkatkan stabilitas pembelajaran model.

Pembagian dataset untuk pelatihan dan pengujian dilakukan menggunakan *stratified split* dengan rasio 80:20. Rasio pembagian 80% data untuk pelatihan dan 20% data untuk pengujian merupakan praktik yang umum digunakan dalam riset machine learning untuk mencapai keseimbangan antara kecukupan data pelatihan dan kemampuan generalisasi model [16].

Model LightGBM dilatih menggunakan *GridSearchCV* untuk mencari kombinasi terbaik dari beberapa hyperparameter seperti n_estimators, learning_rate, dan max_depth. Proses tuning dilakukan dengan *cross-validation* sebanyak tiga kali dan menggunakan metrik evaluasi *F1-score macro*, yang sesuai untuk klasifikasi multi-kelas dengan distribusi label yang tidak merata.

Pemilihan hyperparameter optimal dilakukan secara sistematis melalui GridSearchCV dengan skema 3-fold *cross-validation* dan evaluasi menggunakan *F1-score macro*. Tabel berikut merangkum hyperparameter yang dicoba serta nilai terbaik hasil tuning:

**Tabel 1.** Hyperparameter LightGBM

| Hyperparameter | Nilai yang Dicoba | Nilai Terbaik |
|---|---|---|
| n_estimators | 100, 200 | 200 |
| learning_rate | 0.05, 0.1 | 0.1 |
| max_depth | 10, 20 | 20 |
| class_weight | 'balanced' | 'balanced' |
| random_state | 42 | 42 |

Kombinasi parameter terbaik diperoleh pada n_estimators=200, learning_rate=0.1, dan max_depth=20. Dengan konfigurasi ini, model mencapai performa optimal dalam klasifikasi multi-kelas pada dataset yang telah seimbang.

5. Interpret

Tahap *interpret*asi dilakukan untuk menafsirkan hasil klasifikasi secara kuantitatif dan memberikan pemahaman terhadap model yang dibangun. Setelah model optimal diperoleh dari proses Grid Search,

dilakukan evaluasi terhadap kinerja model menggunakan metrik akurasi, F1-score macro, dan *classification report* yang mencakup precision dan recall pada setiap kelas. Selain itu, digunakan *confusion matrix* untuk mengidentifikasi distribusi kesalahan prediksi antar kategori.

Visualisasi *confusion matrix* dihasilkan dalam bentuk *heatmap* untuk memudahkan identifikasi pola prediksi yang keliru, khususnya pada kelas yang sering tertukar. Interpretasi model juga diperkuat dengan menampilkan *feature importance* dari model LightGBM, untuk mengetahui kontribusi masing-masing fitur dalam keputusan klasifikasi.

Selain evaluasi kuantitatif, dilakukan juga simulasi prediksi menggunakan data baru untuk menunjukkan kemampuan model dalam mengklasifikasikan jenis pelanggaran pada kasus aktual. Proses ini melibatkan input data manual dari pengguna (seperti *State*, *Covered Entity Type*, dan *Individuals Affected*), yang kemudian diproses melalui *pipeline* preprocessing dan diprediksi oleh model terlatih. Hasil prediksi ditampilkan dalam bentuk label pelanggaran data yang paling mungkin terjadi.

Dengan pendekatan ini, interpretasi tidak hanya dilakukan untuk mengevaluasi performa teknikal, tetapi juga untuk memberikan *insight* praktis yang dapat digunakan oleh pengambil keputusan, seperti lembaga pengawas atau pengelola sistem informasi kesehatan. Model yang dibangun memungkinkan pengguna memprediksi risiko jenis pelanggaran data secara cepat, akurat, dan berbasis fitur historis yang terdokumentasi.

## III. HASIL DAN PEMBAHASAN

Penelitian ini menganalisis insiden kebocoran data pada sektor kesehatan di Amerika Serikat dari tahun 2009 hingga 2016 menggunakan laporan HIPAA, dengan pendekatan deskriptif, *text mining*, dan pemodelan prediktif. Proses analisis mengikuti kerangka kerja OSEMN.

### A. Hasil Tahap Eksplorasi Data (*Explore*)

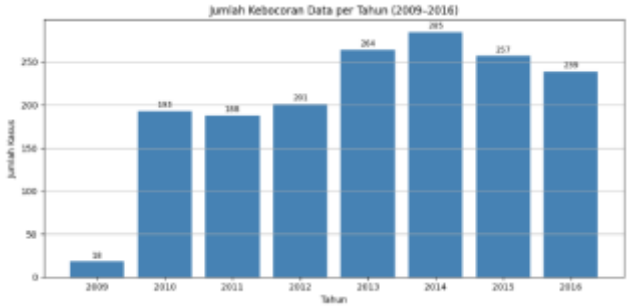
Tahap eksplorasi data memberikan pemahaman awal mengenai distribusi dan pola dalam dataset "Major US Health Data Breaches" yang diperoleh dari *Kaggle*. Dataset ini terdiri dari 1654 entri dengan 10 kolom atribut penting seperti seperti *Name of Covered Entity, State, Covered Entity Type, Breach Submission Date, Type of Breach, Location of Breached Information, Individuals Affected, Web Description, Business Associate Present,* dan *Webpost Date.*

1. Tren Temporal Insiden Kebocoran Data

   Grafik pada Gambar 2 menunjukkan dinamika jumlah insiden kebocoran data kesehatan dari waktu ke waktu. Terlihat adanya peningkatan yang cukup signifikan terutama pada tahun-tahun akhir pengamatan, dengan puncak insiden terjadi sekitar tahun 2014. Peningkatan ini sejalan dengan masifnya adopsi teknologi digital di sektor kesehatan. Meski digitalisasi membawa kemudahan dan efisiensi, hal ini juga menimbulkan tantangan baru, khususnya terkait perlindungan informasi kesehatan pasien yang bersifat sensitif.

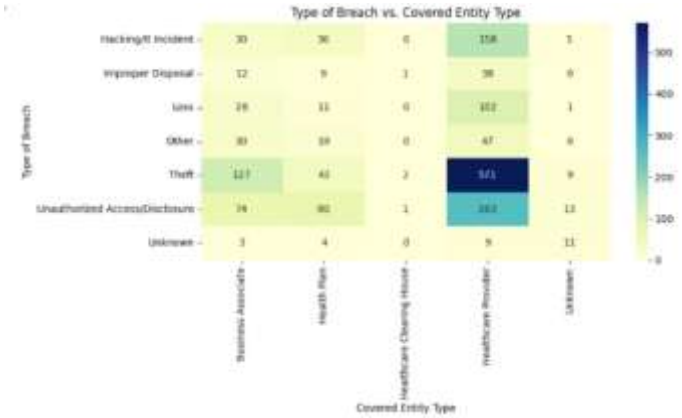
2. Distribusi Geografis Insiden Kebocoran Data

   Distribusi insiden secara geografis menunjukkan bahwa beberapa negara bagian mencatat jumlah kebocoran data yang jauh lebih tinggi dibandingkan lainnya. Negara bagian seperti California, Texas, dan New York merupakan wilayah dengan frekuensi insiden paling tinggi. Hal ini kemungkinan besar berkaitan dengan tingginya jumlah fasilitas layanan kesehatan dan populasi di wilayah-wilayah tersebut, sehingga risiko paparan data juga menjadi lebih besar. Distribusi insiden ini dapat dilihat pada Gambar 3.

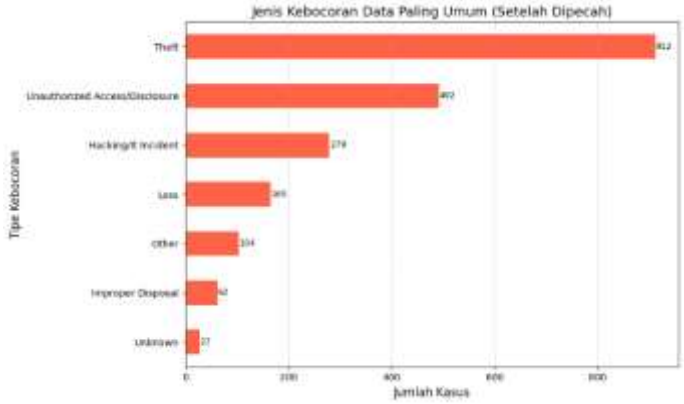



Gambar 2. Grafik jumlah insiden kebocoran data setiap tahunnya dari 2009 hingga 2016.


Gambar 3. Sebaran Geografis Insiden Kebocoran Data Kesehatan di AS (2009–2016)




Gambar 4. Distribusi Insiden Kebocoran Data Berdasarkan Tipe Entitas Tercakup




Gambar 5. Diagram jenis kebocoran paling umum

3. Frekuensi Berdasarkan Tipe Entitas Tercakup (*Covered Entity Type*)

   Berdasarkan jenis entitas yang terlibat, penyedia layanan kesehatan merupakan pihak yang paling banyak mengalami kebocoran data. Hal ini menandakan bahwa institusi pelayanan langsung kepada pasien memiliki risiko yang lebih tinggi, mungkin karena besarnya volume data yang mereka kelola atau lemahnya sistem keamanan yang diterapkan. Visualisasi distribusi insiden kebocoran data berdasarkan tipe entitas tercakup tersaji pada Gambar 4.

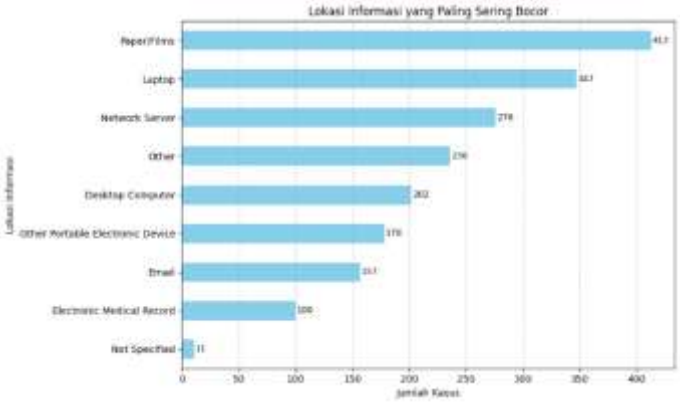
4. Frekuensi Berdasarkan Tipe Pelanggaran (*Type of Breach*)

   Jenis pelanggaran yang paling sering ditemukan adalah Theft dan Unauthorized Access/Disclosure. Dalam dua tahun terakhir, serangan jenis ini mengalami lonjakan yang cukup tajam, sejalan dengan meningkatnya kompleksitas infrastruktur digital dan keterbukaan sistem informasi yang digunakan oleh

institusi kesehatan. Jenis kebocoran paling umum disajikan dalam diagram bar seperti ditunjukkan pada Gamibar 5.

5. Frekuensi Berdasarkan Lokasi Informasi yang Diretas (Location of Breached Information)

Data yang paling sering diretas tersimpan yaitu paper/films dan Laptop. Temuan ini menunjukkan bahwa sistem digital masih memiliki celah keamanan yang rentan terhadap akses tidak sah, meskipun insiden pada media fisik seperti dokumen cetak atau perangkat penyimpanan lokal juga masih terjadi.




Gambar 6. Diagram lokasi penyimpanan data saat terjadi kebocoran.


Gambar 7. *WordCloud* kata kunci serta topik utama hasil LDA.

6. Analisis Teks (*Text Mining*) dari *Web Description*

Analisis teks dari kolom deskripsi insiden mengungkap kata-kata yang sering muncul seperti **"CE," "PHI," "covered," "media," "employee," "notification," "provided," "procedure," "protected,"** dan **"description."** Kata-kata ini mengindikasikan aspek-aspek umum yang terkait dengan insiden, seperti entitas yang terlibat (CE, PHI, employee), tindakan (provided, protected, covered), serta proses atau informasi (media, notification, procedure, description). Sementara itu, analisis topik menggunakan LDA berhasil mengelompokkan deskripsi ke dalam lima topik utama, seperti serangan phishing, insiden dari pihak internal, perangkat yang hilang/dicuri, serangan ransomware, dan pengungkapan data yang tidak disengaja. Temuan ini semakin menguatkan bahwa risiko terbesar datang dari serangan siber yang terus berkembang.

B. Hasil Tahap Pemodelan (*Model*)

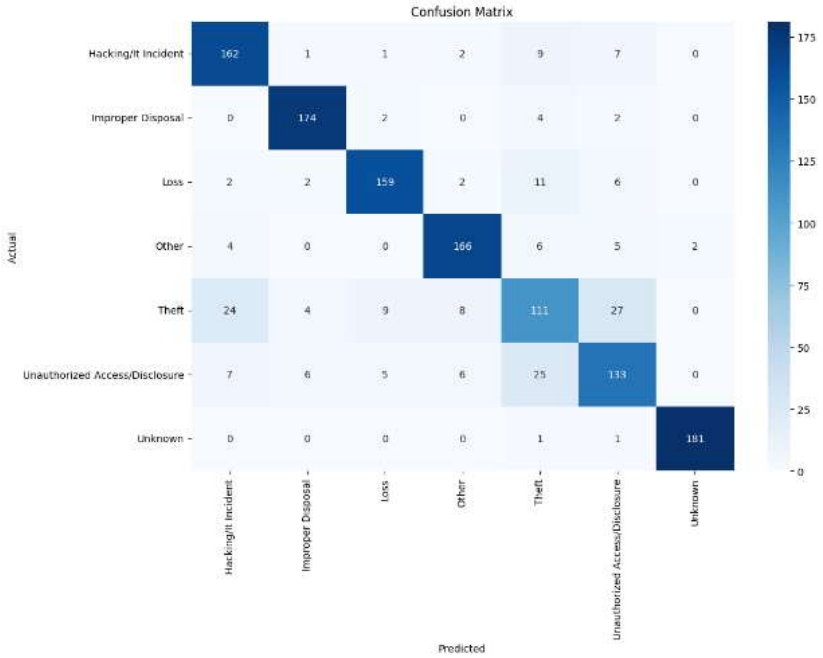
1. Evaluasi Visual Model melalui *Confusion Matrix*

Evaluasi performa model dilakukan melalui yang memperlihatkan hubungan antara label aktual dan prediksi model untuk tujuh kelas pelanggaran, yaitu: *Hacking/IT Incident*, *Improper Disposal*, *Loss*, *Other*, *Theft*, *Unauthorized Access/Disclosure*, dan *Unknown*. Dari matriks tersebut terlihat bahwa model memberikan hasil prediksi yang cukup akurat untuk sebagian besar kelas. Misalnya, terdapat 162 prediksi benar dari 217 kasus aktual untuk kelas *Hacking/IT Incident*, dan 174 dari 178 kasus untuk *Improper Disposal*. Sementara itu, kesalahan klasifikasi lebih tinggi terjadi pada kelas *Theft*, di mana 111 kasus berhasil diprediksi dengan benar, tetapi cukup banyak juga yang salah diklasifikasikan ke kelas lain seperti *Unauthorized Access/Disclosure* dan *Hacking/IT Incident*. Matriks ini juga memperlihatkan bahwa prediksi pada kelas *Unknown* cukup baik (181 benar dari 183), namun nilai aktualnya sangat sedikit. Matriks ini

memberikan gambaran visual mengenai kekuatan dan kelemahan model dalam membedakan antar kelas pelanggaran.

2. Evaluasi Kuantitatif melalui *Classification Report*

Evaluasi lanjutan dilakukan menggunakan metrik kuantitatif berupa *classification report* yang mencakup nilai *precision*, *recall*, dan *f1-score* untuk masing-masing kelas. Berdasarkan hasil yang ditampilkan, akurasi keseluruhan model mencapai 83,55%. Kelas *Theft* memiliki performa paling tinggi dengan nilai f1-score sebesar 0,87, disusul oleh *Unauthorized Access/Disclosure* dengan f1-score sebesar 0,83, serta *Improper Disposal* dan *Hacking/IT Incident* masing-masing dengan f1-score sebesar 0,82 dan 0,81. Nilai rata-rata tertimbang (*weighted average*) untuk f1-score adalah 0,83, menunjukkan bahwa model bekerja cukup baik secara keseluruhan. Namun, terdapat kelemahan pada kelas *Unknown* dengan nilai f1-score hanya sebesar 0,23, mengindikasikan bahwa model belum mampu mengidentifikasi kategori ini secara akurat. Secara umum, *classification report* ini menegaskan bahwa model cukup andal dalam memprediksi mayoritas jenis pelanggaran data, meskipun masih ada ruang perbaikan, terutama untuk kelas minoritas dan kelas yang memiliki karakteristik ambigu.




Gambar 9. *Confusion Matrix*

✅ Akurasi Prediksi (label tunggal saja): 83.55%

📋 Classification Report (tanpa label gabungan):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Hacking/IT Incident | 0.78 | 0.85 | 0.81 | 217 |
| Improper Disposal | 0.82 | 0.82 | 0.82 | 50 |
| Loss | 0.86 | 0.78 | 0.82 | 115 |
| Other | 0.84 | 0.82 | 0.83 | 76 |
| Theft | 0.85 | 0.88 | 0.86 | 707 |
| Unauthorized Access/Disclosure | 0.83 | 0.80 | 0.81 | 386 |
| Unknown | 0.90 | 0.39 | 0.55 | 23 |
| accuracy | | | 0.84 | 1574 |
| macro avg | 0.84 | 0.76 | 0.79 | 1574 |
| weighted avg | 0.84 | 0.84 | 0.83 | 1574 |

Gambar 10. Akurasi Prediksi

C. Hasil Tahap Interpretasi (*Interpret*)

Model klasifikasi yang dibangun untuk memprediksi jenis kebocoran data (*Type of Breach*) menunjukkan performa yang cukup baik. Berdasarkan hasil evaluasi pada data uji internal (hasil pembagian dari train_test_split), model berhasil mencapai akurasi sebesar 85,04% dan nilai macro F1-score sebesar 84,89%. Ini mengindikasikan bahwa model tidak hanya akurat secara keseluruhan, tetapi juga mampu mempertahankan keseimbangan kinerja antar berbagai kelas target, termasuk kelas-kelas dengan distribusi data yang lebih kecil. Hasil *classification report* memperlihatkan bahwa model sangat baik dalam mengklasifikasikan beberapa kategori seperti Improper Disposal dan Unknown yang masing-masing memiliki nilai precision dan recall di atas 90%, bahkan mendekati sempurna. Selain itu, kategori *Loss* dan *Other* juga menunjukkan performa yang stabil, dengan nilai F1 yang mendekati 90%.

Namun, untuk kelas Theft dan Unauthorized Access/Disclosure, performa model masih relatif lebih rendah, dengan F1-score berkisar antara 63%–73%. Ini menunjukkan adanya potensi kebingungan model dalam membedakan pola atau fitur dari kedua tipe kebocoran tersebut. Hal ini mungkin disebabkan oleh adanya kemiripan karakteristik antar fitur seperti jumlah individu yang terdampak, tipe entitas yang terlibat, atau status *Business Associate*. *Confusion matrix* pada evaluasi ini juga memperkuat analisis tersebut, di mana terlihat sebagian besar kesalahan prediksi berasal dari kelas-kelas tersebut.

Lebih lanjut, ketika model diuji menggunakan data eksternal (file prediksi terpisah yang tidak ikut dilatih sebelumnya), hasilnya masih konsisten dengan akurasi prediksi sebesar 83,55%. Ini menunjukkan bahwa model memiliki kemampuan generalisasi yang baik terhadap data baru yang belum pernah dilihat sebelumnya. Nilai precision dan recall pada sebagian besar kelas masih tinggi, kecuali pada label Unknown yang memiliki recall rendah (39%) meskipun precision-nya cukup tinggi (90%). Hal ini mengindikasikan bahwa model sangat selektif dalam memprediksi Unknown dan cenderung menghindari prediksi tersebut jika tidak yakin, yang berdampak pada menurunnya recall. Secara keseluruhan, model terbukti efektif dalam menangani klasifikasi kebocoran data, dan mampu mempertahankan kinerja baik tidak hanya pada data pelatihan, namun juga pada data eksternal, yang menjadikannya layak untuk diintegrasikan dalam sistem pendukung keputusan mitigasi insiden kebocoran informasi.

## IV. KESIMPULAN

Penelitian ini berhasil mengungkap pola dan tren insiden kebocoran data pada sektor kesehatan di Amerika Serikat melalui analisis dataset laporan HIPAA tahun 2009–2016 dengan pendekatan kerangka OSEMN. Hasil eksplorasi menunjukkan bahwa jenis pelanggaran paling umum adalah serangan siber, dengan peningkatan signifikan pada tahun-tahun akhir observasi, serta dominasi insiden pada negara bagian dengan populasi tinggi seperti California dan Texas. Melalui penerapan algoritma LightGBM dalam pemodelan prediktif, model mampu mengklasifikasikan jenis pelanggaran data dengan akurasi dan stabilitas yang baik, terutama pada kelas mayoritas, dengan bantuan teknik SMOTE untuk menyeimbangkan distribusi kelas. Fitur "*Individuals Affected*" terbukti sebagai indikator paling signifikan dalam membedakan jenis pelanggaran. Temuan ini tidak hanya memperkuat pemahaman terhadap lanskap ancaman keamanan informasi kesehatan, tetapi juga menawarkan pendekatan prediktif yang potensial untuk mendukung pengambilan keputusan dan mitigasi risiko di sektor kesehatan. Ke depan, penelitian ini dapat dikembangkan lebih lanjut dengan menambahkan data terbaru, fitur kontekstual yang lebih kaya, serta eksplorasi metode klasifikasi lain untuk meningkatkan akurasi pada kelas minoritas.

## DAFTAR PUSTAKA

[1] D. Molitor, A. Saharia, V. Raghupathi, dan W. Raghupathi, "Exploring the characteristics of data breaches: A descriptive analytic study," J. Inf. Secur. , vol. 15, hal. 168–195, 2024.

[2] R. F. Parks dan L. Adams, "Analyzing security breaches in the U.S.: A business analytics case-study," Inf. Syst. Educ. J., vol. 14, no. 2, hal. 43–48, 2016.

[3] M. H. Gabriel, A. Noblin, A. Rutherford, A. Walden, dan K. Cortelyou-Ward, "Data breach locations, types, and associated characteristics among US hospitals," Am. J. Manag. Care , vol. 24, no. 2, hal. 78–84, Feb. 2018.

[4] V. Liu, M. A. Musen, dan T. Chou, "Data breaches of protected health information in the United States," JAMA , vol. 313, no. 14, hal. 1471–1473, Apr. 2015.

[5] H. Mason dan C. Wiggins, A Taxonomy of Data Science: OSEMN Framework , 2010. [Online]. Tersedia: https://www.dataists.com/2010/09/a-taxonomy-of-data-science/

[6] DQLab, "Step by step tugas data scientist dengan framework OSEMN," DQLab, 2024. [Online]. Tersedia: https://dqlab.id/step-by-step-tugas-data-scientist-dengan-framework-osemn

[7] N. Nazurah et al ., "HealthyHeart: Visualisasi Prediktif Kondisi Jantung Menggunakan Machine Learning dan OSEMN," J. Informatika , vol. 7, no. 2, 2023.

[8] W. W. Koczkodaj, M. Nowacki, W. Pedrycz, dan D. Strzalka, “Text mining analysis of over 392 million compromised healthcare records,” Healthc. Anal. , 2025.

[9] W. Raghupathi, V. Raghupathi, dan A. Saharia, “Analyzing health data breaches: A visual analytics approach,” Health Inf. Manag. J. , vol. 52, no. 1, 2023.

[10] J. Reddy, N. Elsayed, Z. ElSayed, dan M. Ozer, “A review on data breaches in healthcare security systems,” Int. J. Cyber Health Secur. , vol. 4, no. 2, 2023.

[11] S. Aljawarneh, M. Aldwairi, dan M. B. Yassein, “Anomaly-based intrusion detection system through feature selection and hybrid model,” J. Comput. Sci. , vol. 25, hal. 63–75, 2018.

[12] G. Ke et al ., “LightGBM: A highly efficient gradient boosting decision tree,” Adv. Neural Inf. Process. Syst. , vol. 30, 2017.

[13] M. M. Hossain and Y. A. Hong, "Trends and characteristics of protected health information breaches in the United States," 2022.

[14] M. Huo, M. Bland, and K. Levchenko, "All Eyes On Me: Inside Third Party Trackers’ Exfiltration of PHI from Healthcare Providers’ Online Systems," Proc. 21st Workshop on Privacy in the Electronic Society (WPES ’22), ACM, Los Angeles, CA, USA, 2022.

[15] M. Zhu, Y. Zhang, Y. Gong, K. Xing, X. Yan, and J. Song, "Ensemble Methodology: Innovations in Credit Default Prediction Using LightGBM, XGBoost, and LocalEnsemble," 2024.

[16] A. Shepard and N. Naheed, “Application of Data Transformation Techniques and Train-Test Split,” 2021 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 58-63, Dec. 2021, doi: 10.1109/CSCI54926.2021.00079.