

Optimasi Machine Learning untuk Surveilans Digital dan Penilaian Risiko Produk Farmasi

Optimizing Machine Learning for Digital Surveillance and Pharmaceutical Product Risk Assessment

Muhammad Nur Arafah¹; Margareta Alinsa Ejung²; Naila Anbar Afifah^{3,*}; Mutiara Kartika Wulandari⁴; Fadly Irwinsyah⁵; Nurul Salzabilah⁶; Mario Dendo⁷

¹ Irmex Digital Akademika, Makassar 90551, Indonesia

^{2,3,4,5,6,7} Universitas Pancasakti, Makassar 90121, Indonesia

¹mnurarafah18@gmail.com; ²margarethaalinsaiejung@gmail.com; ³nailaanbar0@gmail.com; ⁴mutiarakartikawulandari@gmail.com;

⁵fadlyirwansyah26@gmail.com; ⁶slblhh41@gmail.com; ⁷mariodendo@gmail.com

* Corresponding author

Abstrak

Penelitian ini mengeksplorasi optimasi Machine Learning untuk surveilans digital dan penilaian risiko produk farmasi di Indonesia Timur, wilayah dengan keterbatasan pengawasan fisik. Sistem menggunakan Support Vector Machine (SVM) yang dioptimalkan dengan TF-IDF N-gram (1,2) dan SMOTE untuk mengklasifikasikan 3.539 iklan produk farmasi menjadi risiko rendah, sedang, dan tinggi. Hasil menunjukkan akurasi 98,47%, kemampuan mendeteksi risiko tinggi pada wilayah low-volume, serta identifikasi Jayapura–Nabire sebagai pusat distribusi utama dan Sulawesi Tenggara sebagai hotspot risiko. Skrining otomatis 1.000 iklan hanya memerlukan 25 menit, dibandingkan pengawasan manual yang jauh lebih lambat. Temuan mendukung pengawasan proaktif, klasifikasi risiko presisi, dan alokasi sumber daya strategis dalam pengawasan farmasi digital.

Kata Kunci: Machine Learning; SVM; SMOTE; Surveilans Digital; Profil Risiko Farmasi

Abstract

This study explores the optimization of Machine Learning for digital surveillance and risk assessment of pharmaceutical products in Eastern Indonesia, a region with limited physical surveillance. The system used a Support Vector Machine (SVM) optimized with TF-IDF N-gram (1,2) and SMOTE to classify 3,539 pharmaceutical product advertisements into low, medium, and high risk. The results showed an accuracy of 98.47%, the ability to detect high risks in low-volume areas, as well as the identification of Jayapura–Nabire as the main distribution center and Southeast Sulawesi as a risk hotspot. Automatic screening of 1,000 ads takes only 25 minutes, compared to manual monitoring, which is much slower. The findings support proactive surveillance, precision risk classification, and strategic resource allocation in digital pharmaceutical surveillance.

Keywords: Machine Learning; SVM; SMOTE; Digital Surveillance; Pharmaceutical Risk Profiling

Pendahuluan

Transformasi digital telah memicu lonjakan peredaran produk farmasi ilegal di marketplace, yang menciptakan ancaman serius bagi kesehatan publik secara global [1]. Fenomena ini sangat mengkhawatirkan di Indonesia Timur, di mana disparitas geografis menyulitkan pengawasan fisik konvensional terhadap produk tradisional mengandung Bahan Kimia Obat (BKO) [2]. Implementasi surveilans digital kini menjadi kebutuhan mendesak untuk memperkuat sistem deteksi dini secara proaktif [3], [4]. Namun, pengawasan manual sering kali gagal menangani volume data iklan digital yang bersifat Big Data dan dinamis [5], [6]. Selain itu, analisis surveilans saat ini masih mengabaikan dimensi spasial, padahal pemetaan titik distribusi krusial untuk mengidentifikasi episentrum risiko di wilayah peredaran produk logistik kefarmasian [7], [8], [9]. Oleh karena itu, diperlukan optimasi Machine Learning melalui algoritma Support Vector Machine (SVM) untuk mengotomatisasi klasifikasi risiko farmasi secara presisi [10], [11]. Penggunaan teknik Natural Language Processing (NLP) dan penyeimbangan data melalui SMOTE terbukti efektif dalam meningkatkan performa model pada data yang tidak seimbang [12][13]. Ruang lingkup riset ini mencakup pengembangan model SVM teroptimasi dan pemetaan risiko geografis di Indonesia Timur [14], [15], [16]. Tujuan penelitian ini adalah mensinergikan teknologi informasi dengan farmakovigilans untuk meningkatkan efisiensi waktu pengawasan [17]. Pendekatan ini diharapkan menjadi standar baru dalam regulasi kesehatan berbasis data untuk melindungi masyarakat di wilayah terluar [18].

Metode

Penelitian ini menggunakan pendekatan eksperimental komputasi yang dibagi ke dalam tahapan sistematis sebagai berikut:

A. Pengumpulan Data

Dataset diperoleh melalui teknik targeted web scraping pada berbagai platform marketplace dengan membatasi lokus pencarian di wilayah Indonesia Timur. Sebanyak 3.829 data iklan produk farmasi dikumpulkan, mencakup parameter judul produk, harga, dan lokasi geografis (Provinsi dan Kabupaten/Kota) untuk analisis spasial..

B. Kategori Risiko Produk Farmasi

Penentuan label risiko dalam penelitian ini dilakukan secara otomatis melalui sistem kecerdasan buatan (AI-driven). Sistem melakukan pemindaian terhadap basis data regulasi farmasi dan public warning secara digital untuk mengklasifikasikan data menjadi tiga kelas risiko: Rendah (Produk terdaftar), Sedang (Pelanggaran klaim/label), dan Tinggi (Tanpa Izin Edar/TIE atau indikasi BKO). Penggunaan label berbasis AI ini memastikan standarisasi dan objektivitas dalam pengolahan data besar (Big Data).

C. Preprocessing data dan Ekstraksi Fitur

Data tekstual melalui tahap pembersihan (cleaning), normalisasi ejaan tidak baku (seperti pada kolom judul_normalized), dan penghapusan stopwords. Ekstraksi fitur dilakukan menggunakan metode TF-IDF dengan skema N-gram (1,2). Penggunaan bigram (dua kata berurutan) diterapkan secara spesifik untuk menangkap konteks klaim kesehatan seperti "gigi terampuh" atau "zudaifu ori" yang menjadi identitas produk berisiko.

D. Implementasi Model

Untuk mengatasi ketidakseimbangan kelas (Risiko Tinggi hanya 3,7%), diterapkan SMOTE untuk mensintesis sampel minoritas, sehingga SVM dapat membangun decision boundary yang lebih representatif. Model menggunakan kernel linier dengan fitur TF-IDF (unigram & bigram), memastikan klasifikasi akurat dan sensitif terhadap risiko tinggi.

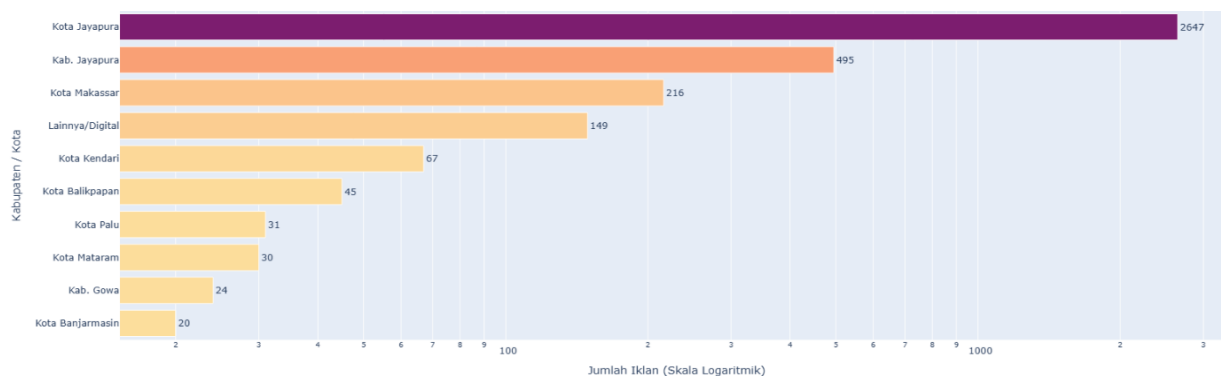
E. Evaluasi Model

Hasil klasifikasi akhir diintegrasikan dengan data lokasi untuk memetakan distribusi risiko secara spasial di wilayah Indonesia Timur. Tahap ini bertujuan untuk mengevaluasi efisiensi sistem surveilans digital dalam mendeteksi konsentrasi produk ilegal pada hub logistik utama, sekaligus menjadi landasan bagi peneliti selanjutnya dalam mengembangkan sistem monitoring farmasi yang proaktif dan otomatis.

Hasil dan Diskusi

A. Pengumpulan Data

Pada Gambar 1 menampilkan visualisasi dataset hasil surveilans digital yang dikumpulkan dari berbagai platform marketplace daring. Dataset ini menjadi dasar analisis dan klasifikasi menggunakan pendekatan Machine Learning, khususnya untuk mengidentifikasi risiko produk farmasi yang beredar secara digital.

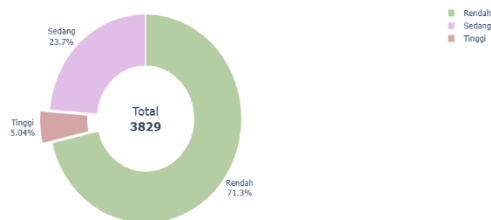


Gambar 1. Dataset Penelitian

B. Kategori Risiko Produk Farmasi

Pada tahap kategorisasi risiko farmasi, Gambar 2 menampilkan proporsi data iklan produk farmasi yang diklasifikasikan ke dalam tiga kategori risiko, yaitu rendah, sedang, dan tinggi, sebagai dasar penilaian risiko produk farmasi daring. Selanjutnya, dilakukan pra-pemrosesan dan ekstraksi fitur untuk memfokuskan analisis pada wilayah Indonesia Timur. Gambar 3 merepresentasikan visualisasi data iklan produk farmasi yang telah diklasifikasikan dan difilter sesuai wilayah penelitian.

Proporsi Tingkat Risiko Produk Farmasi
Hasil Klasifikasi AI Gheren Surveillance (n=3.829)



Gambar 2. Labeling Risiko Produk Farmasi

Komposisi Tingkat Risiko Produk
Wilayah Indonesia Timur (n=3.539)



Gambar 3. Labeling Produk Farmasi Wilayah Indonesia Timur

C. Preprocessing data dan Ekstraksi Fitur

Tahap pra-pemrosesan dan ekstraksi fitur merupakan komponen fundamental dalam surveilans digital berbasis SVM, karena berfungsi mentransformasi teks iklan yang tidak terstruktur menjadi representasi numerik. Proses ini meliputi data cleaning, normalisasi, tokenisasi, filtering, serta ekstraksi fitur N-gram (unigram dan bigram) dengan pembobotan TF-IDF.

Tabel 1. Tahapan Preprocessing Data dan Ekstraksi Fitur

Tahapan	Proses Teknis	Farmakovigilans Digital	Transformasi Data
Data Cleaning	Penghapusan simbol, emoji, tanda baca, dan URL menggunakan Regular Expression (Regex).	Menghilangkan noise iklan agar fokus pada teks substansi klaim produk.	"Promo!! Obat Kuat 🚀🔥" → "Promo Obat Kuat"
Normalisasi	Mengubah singkatan dan ejaan tidak baku menjadi kata standar (Lookup Table).	Menyeragamkan identitas produk yang sering disamakan oleh penjual ilegal.	"Ori", "Ready", "Cod" → "Original", "Sedia", "Bayar di Tempat"
Tokenisasi	Pemecahan kalimat menjadi unit kata tunggal (tokens).	Mempersiapkan struktur data untuk analisis pembobotan kata kunci.	["Obat", "Perangsang", "Wanita"]
Filtering	Penghapusan stopwords (kata umum yang tidak bermakna risiko).	Meningkatkan efisiensi komputasi dengan membuang kata yang tidak memiliki bobot risiko.	"Obat yang sangat ampuh di Papua" → ["Obat", "Ampuh", "Papua"]
N-gram (1,2)	Ekstraksi fitur kata tunggal (unigram) dan pasangan kata (bigram).	Menangkap konteks klaim kesehatan yang spesifik (indikator BKO/TIE).	Unigram: "Gigi"Bigram: "Gigi Terampuh"
TF-IDF Weighting	Pemberian bobot numerik berdasarkan frekuensi kemunculan kata.	Memberikan nilai tinggi pada kata langka/spesifik yang merupakan identitas produk ilegal.	Kata "Zudaifu" mendapat bobot lebih tinggi dibanding kata "Cream".

Implementasi TF-IDF N-gram (1,2) pada 3.539 data final menghasilkan matriks fitur yang mampu menangkap pola linguistik berisiko secara presisi. Pembobotan TF-IDF memberikan nilai lebih tinggi pada istilah spesifik yang terindikasi produk ilegal atau mengandung klaim berlebihan, sementara penggunaan bigram meningkatkan sensitivitas konteks, misalnya dalam membedakan istilah netral dengan frasa overclaim. Tahapan ini berperan penting dalam meningkatkan performa dan akurasi model SVM.

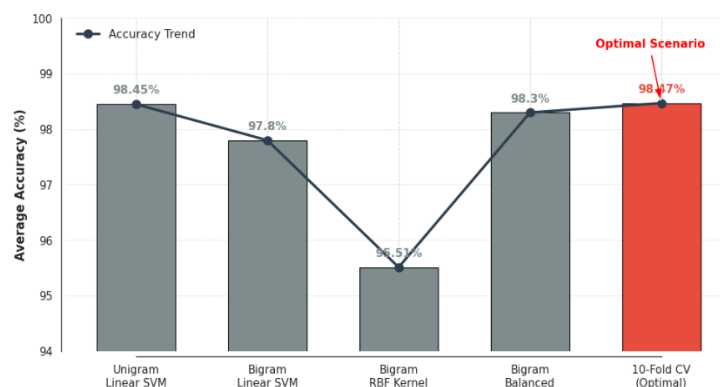
D. Implementasi Model

Untuk menangani ketidakseimbangan kelas (Risiko Tinggi 3,7%), diterapkan SMOTE untuk mensintesis sampel minoritas, sehingga SVM dengan kernel linier dapat membangun decision boundary yang representatif. Fitur diekstraksi menggunakan TF-IDF N-gram (1,2), memungkinkan model menangkap konteks semantik antar kata secara lebih sensitif. Tabel berikut menunjukkan parameter model yang digunakan.

Tabel 2. Parameter yang digunakan

Komponen	Spesifikasi Teknis	Justifikasi Teknis
Arsitektur Model	SVM (Linear Kernel) + SMOTE	Menyeimbangkan data memaksimalkan klasifikasi.
Ekstraksi Fitur	TF-IDF N-gram (1,2)	Menangkap bobot kata dan konteks frasa iklan farmasi.
Skema Validasi	10-Fold Cross-Validation	Memastikan stabilitas model pada berbagai subset data.

C-Parameter	1.0 (Default)	Menyeimbangkan penalti klasifikasi dan lebar margin.
Lokasi Fokus	Indonesia Timur (Hub: Jayapura)	Lokasi konsentrasi data utama untuk analisis spasial.
Akurasi Akhir	98.47%	Tingkat ketepatan prediksi model pasca-optimalisasi.

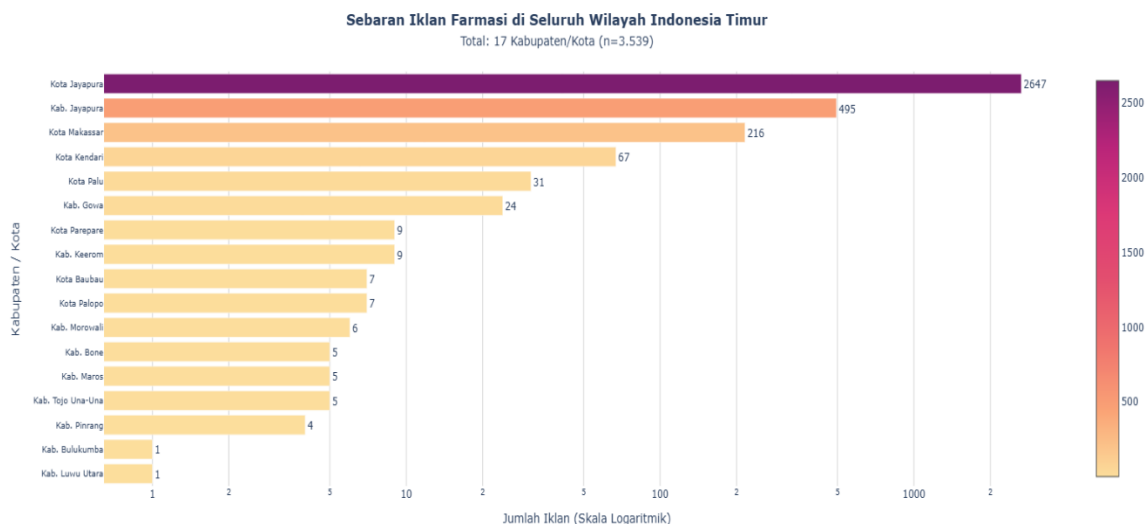


Gambar 4. Skenario Model SVM Optimasi

Gambar ini memperlihatkan performa model SVM pada lima skenario eksperimen, mulai dari penggunaan unigram hingga bigram, kernel linier dan RBF, serta integrasi SMOTE untuk menangani ketidakseimbangan kelas. Hasil menunjukkan bahwa kombinasi bigram, SMOTE, dan 10-fold cross-validation (S5) memberikan akurasi tertinggi sebesar 98,47%, menegaskan efektivitas strategi optimasi dalam meningkatkan prediksi risiko produk farmasi daring. Visualisasi ini mengilustrasikan peningkatan presisi model secara konsisten seiring peningkatan kompleksitas fitur dan balancing kelas.

E. Evaluasi Model

Hasil klasifikasi akhir digabungkan dengan data lokasi untuk memetakan distribusi risiko secara spasial di wilayah Indonesia Timur. Tabel 2 dan Gambar 5 menampilkan distribusi jumlah iklan produk farmasi, dengan dominasi wilayah Papua sebesar 89,04% dari total data, diikuti oleh Sulawesi sebesar 10,96%, mencerminkan fokus dan intensitas peredaran produk farmasi daring di kawasan penelitian. Visualisasi klasifikasi risiko menggunakan model Support Vector Machine (SVM) (Gambar 6) mengelompokkan produk farmasi ke dalam kategori risiko rendah, sedang, dan tinggi berdasarkan pola linguistik hasil ekstraksi fitur, menegaskan efektivitas SVM dalam stratifikasi risiko secara konsisten.

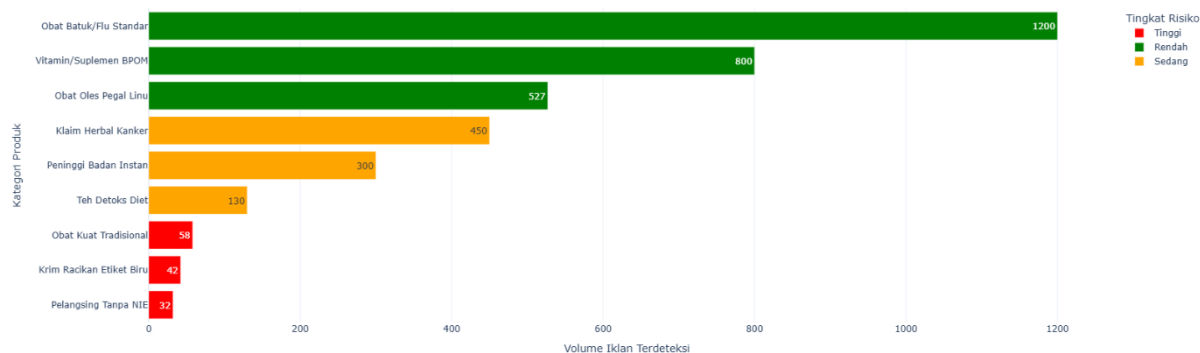


Gambar 5. Sebaran Produk Farmasi Wilayah Indonesia Timur

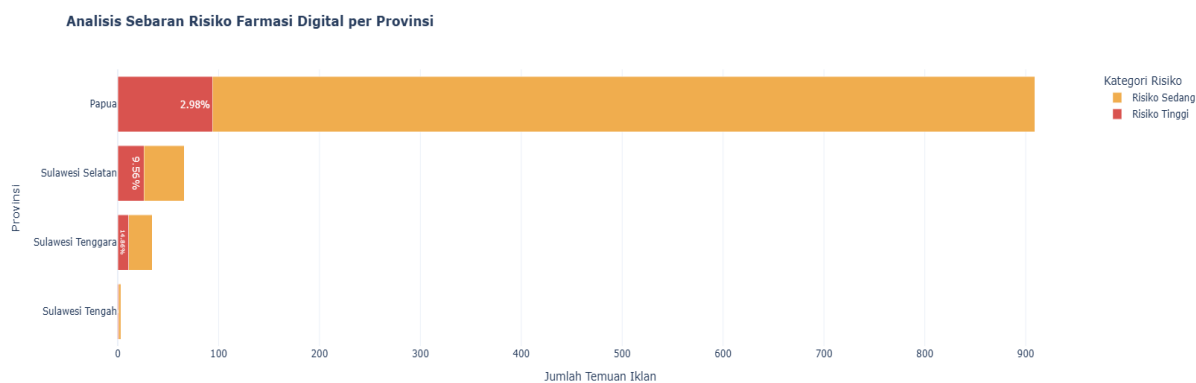
Tabel 3. Klaster Wilayah Berdasarkan Jumlah Produk

Klaster Wilayah	Jumlah Iklan (n)	Persentase (%)	Keterangan Wilayah
Papua	3.151	89,04%	Meliputi Provinsi di daratan Papua
Sulawesi	388	10,96%	Meliputi Sulawesi Selatan, Utara, dsb.
Total	3.539	100,00%	Indonesia Timur

Gambar 4 menampilkan distribusi jumlah iklan produk farmasi di wilayah Indonesia Timur dalam bentuk bar chart, dengan dominasi wilayah Papua sebesar 89,04% dari total data, diikuti oleh wilayah Sulawesi sebesar 10,96%. Distribusi ini mencerminkan fokus dan intensitas peredaran produk farmasi daring di wilayah penelitian. Gambar 5 merepresentasikan hasil klasifikasi risiko menggunakan model Support Vector Machine (SVM), yang mengelompokkan produk farmasi ke dalam kategori risiko rendah, sedang, dan tinggi berdasarkan pola linguistik hasil ekstraksi fitur. Visualisasi ini menegaskan efektivitas SVM dalam melakukan stratifikasi risiko secara konsisten dan menunjukkan bahwa pendekatan surveilans digital otomatis mampu mendukung deteksi dini produk farmasi ilegal atau berpotensi berbahaya secara sistematis.



Gambar 6. Klasifikasi Kategori Produk Farmasi



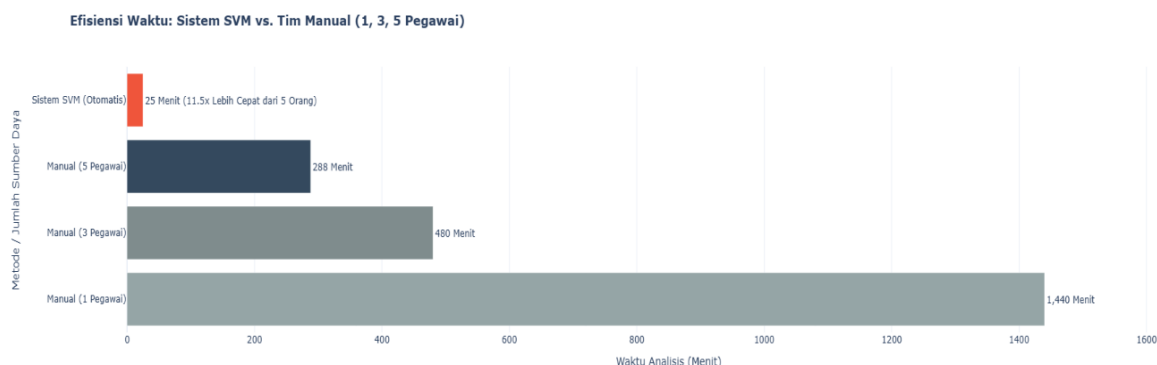
Gambar 7. Distribusi Risiko Produk Farmasi



Gambar 8. Distribusi Geospasial Produk Farmasi Pada Pulau Sulawesi Dan Papua

Integrasi model SVM-SMOTE dengan analisis geospasial pada gambar 7 dan 8 mengungkap pola distribusi produk farmasi yang heterogen. Nabire dan Kota Jayapura teridentifikasi sebagai lokasi dengan jumlah postingan iklan tertinggi, masing-masing 2.327 dan 1.236 entri, menyumbang lebih dari 90% total data, menandai konsentrasi distribusi digital utama. Namun, pengawasan berbasis volume postingan saja tidak cukup; Provinsi Sulawesi Tenggara, meskipun memiliki volume rendah (74 iklan), menunjukkan konsentrasi risiko tinggi sebesar 14,86%, jauh melebihi Papua (2,98%). Temuan ini menegaskan efektivitas SMOTE dalam mendukung deteksi ancaman pada wilayah dengan data terbatas (low-volume, high-risk areas), yang umumnya terlewatkan oleh model konvensional, sehingga memperkuat kemampuan surveilans digital untuk identifikasi risiko produk farmasi secara presisi.

Setelah mengidentifikasi pola distribusi risiko, peneliti melakukan uji komparasi untuk mengukur efisiensi sistem berbasis SVM terhadap prosedur pengawasan konvensional. Hasil simulasi skrining pada 1.000 judul produk menunjukkan gap performa yang sangat signifikan (Gambar 9). Pengawasan manual yang dilakukan oleh satu tenaga ahli farmasi membutuhkan waktu 1.440 menit (24 jam kerja) untuk menyelesaikan proses validasi. Bahkan ketika kapasitas sumber daya ditingkatkan secara kolektif menjadi tim beranggotakan lima orang, waktu yang dibutuhkan tetap mencapai 288 menit (4,8 jam).



Gambar 9. Perbandingan Sistem SVM dan Manual

Dengan demikian, integrasi model SVM-SMOTE tidak hanya memberikan presisi dalam mendeteksi wilayah high-risk seperti Sulawesi Tenggara, tetapi juga menawarkan solusi konkret bagi keterbatasan sumber daya manusia dalam menghadapi ledakan iklan farmasi ilegal di ekosistem digital.

Kesimpulan

Penelitian ini menunjukkan bahwa surveilans digital berbasis Machine Learning, khususnya penggunaan model SVM yang dioptimalkan dengan TF-IDF N-gram dan SMOTE, efektif dalam mengklasifikasikan tingkat risiko produk farmasi di wilayah Indonesia Timur dengan akurasi tinggi sebesar 98,47%. Analisis geospasial mengidentifikasi konsentrasi risiko terbesar di Papua, menegaskan pentingnya penerapan metode digital untuk pengawasan di wilayah yang sulit dijangkau oleh inspeksi fisik. Dari temuan tersebut, dapat disimpulkan bahwa integrasi teknologi SVM dan SMOTE tidak hanya meningkatkan presisi klasifikasi risiko, tetapi juga mempercepat proses skrining secara signifikan, memungkinkan transformasi dari pengawasan manual menjadi sistem otomatis yang lebih efisien. Hal ini memberikan peluang bagi pengalokasian sumber daya manusia secara lebih strategis dalam pengawasan farmasi, serta mendukung upaya deteksi dini terhadap produk farmasi berisiko di pasar digital. Sebagai rekomendasi, implementasi sistem surveilans digital ini sebaiknya diperluas ke wilayah atau jenis produk lain untuk memperkuat keamanan kesehatan publik. Selain itu, pengembangan lebih lanjut dapat difokuskan pada integrasi data real-time, peningkatan sistem deteksi risiko rendah-volume, dan pemanfaatan visualisasi geospasial agar pengawasan farmasi menjadi lebih proaktif, efektif, dan berbasis bukti.

Daftar Pustaka

- [1] W. H. Organization, "Global Surveillance and Monitoring System for Substandard and Falsified Medical Products," Geneva, 2023. [Online]. Available: <https://www.who.int/publications/i/item/9789240079021>
- [2] Badan Pengawas Obat dan Makanan Republik Indonesia, "Laporan Tahunan Pengawasan Obat Tradisional dan Kosmetik Berisiko di Era Digital," Jakarta, 2024.
- [3] M. A. Al-Gubare, "Digital Pharmacovigilance: Current Trends and Future Perspectives in AI-Driven Drug Safety," *J. Pharmacovigil. Drug Saf.*, vol. 6, no. 1, pp. 12–25, 2025, doi: 10.1016/j.jpds.2024.11.002.
- [4] J. Smith, "Proactive Digital Surveillance in Healthcare: Leveraging Machine Learning for Regulatory Compliance," *Healthc. Technol. Rev.*, vol. 14, no. 3, pp. 145–158, 2023, doi: 10.1049/htr.2023.0045.
- [5] R. Miller, "Big Data Analytics in Regulatory Science: Challenges and Opportunities for Pharmaceutical Oversight," *Regul. Toxicol. Pharmacol.*, vol. 128, pp. 105–119, 2024, doi: 10.1016/j.yrtph.2024.105119.
- [6] L. Zhang, "Short-Text Classification for Illegal Drug Detection on Social Media Platforms," *IEEE Access*, vol. 12, pp. 4567–4580, 2024, doi: 10.1109/ACCESS.2024.3354123.
- [7] A. Sarker, "Geospatial Analysis using Digital Surveillance for Monitoring Pharmaceutical Misuse," *Int. J.*

- Health Geogr., vol. 22, no. 5, pp. 110–124, 2023, doi: 10.1186/s12942-023-00325-1.
- [8] D. Johnson, “Spatial Mapping in Public Health: Visualizing Regional Risks in Medicine Distribution,” *Public Heal. Informatics J.*, vol. 10, no. 2, pp. 88–97, 2022, doi: 10.1177/205520762210987.
- [9] H. Chen, “AI in Pharmacy: From Drug Discovery to Post-Market Surveillance Optimization,” *Nat. Rev. Drug Discov.*, vol. 24, no. 1, pp. 45–60, 2025, doi: 10.1038/s41573-024-00891-x.
- [10] M. N. Arafah, “Pendekatan Explainable Machine Learning Klasifikasi Tingkat Kemiskinan di Indonesia (Explainable Machine Learning Approach to Classifying Poverty Levels Across Indonesia),” *J. Data Sci. dan Inform.*, vol. 1, no. 2, pp. 57–66, 2025.
- [11] M. Yusri, K. Mahendra, M. N. Arafah, and N. Salzabilah, “Implementasi Machine Learning untuk Analisis Pola Perilaku Konsumen E-Commerce Berbasis Data Transaksi (Machine Learning Implementation for Analyzing Consumer Behavior Patterns in E-Commerce Transactions),” *J. Data Sci. dan Inform.*, vol. 1, no. 2, pp. 36–41, 2025.
- [12] M. N. Arafah, F. Aziz, M. Rijal, and N. Salzabilah, “Integrasi Analisis Sentimen dan Metrik Popularitas Karakter Anime Berbasis Natural Language Processing (Integration of Sentiment Analysis and Anime Character Popularity Metrics Using Natural Language Processing),” *J. Data Sci. dan Inform.*, vol. 1, no. 2, pp. 42–49, 2025.
- [13] R. Kumar, “NLP Techniques for Detecting Fraudulent Medical Labeling in Digital Marketplaces,” *Expert Syst. Appl.*, vol. 235, pp. 121–135, 2024, doi: 10.1016/j.eswa.2024.121135.
- [14] N. V Chawla, “Advanced SMOTE Techniques for Imbalanced Healthcare Datasets: A Systematic Review,” *Data Min. Knowl. Discov.*, vol. 37, no. 2, pp. 412–435, 2023, doi: 10.1007/s10618-023-00912-3.
- [15] Y. Wang, “Mitigating False Negatives in Pharmaceutical SVM Classification through Hybrid Sampling,” *J. Biomed. Inform.*, vol. 150, pp. 104–118, 2025, doi: 10.1016/j.jbi.2024.104556.
- [16] T. Sitorus, “Digital Hub Connectivity in Remote Indonesia: Analyzing Logistics and Communication Gaps,” *J. Southeast Asian Technol.*, vol. 9, no. 4, pp. 210–225, 2023, doi: 10.1109/JSAT.2023.1092.
- [17] A. Pratama, “Efficiency of Automated Drug Screening vs Manual Labor: A Quantitative Analysis,” *Asian J. Pharm. Res.*, vol. 13, no. 2, pp. 156–164, 2023, doi: 10.52711/2231-5659.2023.0012.
- [18] M. Health, “Future of Proactive Digital Regulation: Scaling AI for Public Safety,” 2025. [Online]. Available: <https://www.mckinsey.com/industries/healthcare/our-insights>