ANALISIS PENGARUH KEBIASAAN MEROKOK DAN AKTIVITAS BEGADANG TERHADAP RISIKO PENYAKIT PARU-PARU MENGGUNAKAN SUPPORT VECTOR MACHINE

Athallah Yasyfi Imran, Muhammad Rafli Maulana, Gabriel Lifiano Jamot Munthe, Deni Athallah Ubaid, Muhammad Yasir Alghifari, Rizki Adriansyah,Ken Ditha Tania, Allsela Meiriza
Sistem Informasi, Universitas Sriwijaya
Jalan Palembang-Prabumulih km 32 Palembang, Indonesia
09031282227106@student.unsri.ac.id

ABSTRAK

Penelitian ini bertujuan menganalisis pengaruh kebiasaan merokok dan aktivitas begadang terhadap risiko penyakit paru-paru menggunakan Support Vector Machine (SVM). Kebiasaan merokok dan kurang tidur telah diidentifikasi sebagai faktor risiko gangguan paru-paru, namun interaksi kompleks antara keduanya sering diabaikan dalam penelitian konvensional. Metode statistik tradisional cenderung tidak mampu menangani hubungan nonlinier atau efek pengacau tersembunyi. Dataset terdiri dari 30.001 sampel dengan 10 atribut, termasuk usia, jenis kelamin, kebiasaan merokok, dan aktivitas begadang. Data diproses melalui tahapan Knowledge Discovery in Database meliputi seleksi data, praproses dengan label encoding dan scaling, transformasi data, pemodelan SVM, dan evaluasi. Hasil penelitian menunjukkan model SVM mencapai akurasi 89,28%, dengan presisi 0,83 dan recall 1,00 untuk kelas tidak berisiko, serta presisi 1,00 dan recall 0,78 untuk kelas berisiko. Temuan ini menegaskan kombinasi kebiasaan merokok dan pola tidur irregular signifikan meningkatkan risiko penyakit paru-paru, serta memberikan rekomendasi berbasis bukti bagi praktisi kesehatan dalam merancang intervensi preventif.

Kata kunci : *support vector machine, kebiasaan merokok, aktivitas begadang, penambangan data, penemuan pengetahuan dalam basis data, pembelajaran mesin.*

1. PENDAHULUAN

Di era digital yang ditandai dengan integrasi masif *big data* dan kecerdasan buatan (AI) dalam bidang kesehatan, pendekatan prediktif berbasis *machine learning* semakin mendapat perhatian luas. Tren ini tidak hanya mencerminkan kebutuhan akan solusi diagnosis yang cepat dan akurat, tetapi juga menggarisbawahi pentingnya pemodelan ulang terhadap faktor-faktor risiko klasik dengan pendekatan komputasional yang lebih adaptif. Salah satu isu yang menonjol adalah risiko penyakit paru-paru yang berkaitan dengan kebiasaan merokok dan aktivitas begadang. Kedua kebiasaan tersebut telah lama diketahui sebagai faktor yang berdampak negatif terhadap fungsi paru, namun sebagian besar penelitian terdahulu cenderung menganalisisnya secara terpisah, tanpa mempertimbangkan kompleksitas interaksinya dalam kehidupan nyata.

Permasalahan utama yang muncul adalah belum optimalnya metode prediktif dalam mendeteksi interaksi nonlinier antara faktor gaya hidup yang saling berkaitan. Metode statistik tradisional, seperti regresi linier, sering kali tidak mampu mengakomodasi pola hubungan multivariabel yang kompleks, termasuk efek pengacau tersembunyi yang memengaruhi akurasi interpretasi data. Padahal, beberapa temuan epidemiologis terbaru menunjukkan bahwa kombinasi antara merokok dan kurang tidur dapat mempercepat penurunan fungsi paru hingga 40% dibandingkan jika hanya satu faktor risiko yang dimiliki. Situasi ini menandakan bahwa pendekatan yang lebih fleksibel dan cerdas sangat dibutuhkan dalam analisis risiko penyakit paru.

Salah satu metode yang memiliki potensi besar dalam menangani permasalahan ini adalah algoritma *Support Vector Machine* (SVM). SVM dikenal mampu mengklasifikasikan data berdimensi tinggi dengan membentuk batas pemisah optimal, serta mengatasi ketidakteraturan pola data melalui teknik *kernel* yang adaptif. Metode ini juga dinilai lebih robust dalam mengurangi bias prediksi yang sering muncul akibat tumpang tindih data. Di sisi lain, berbagai penelitian sebelumnya menunjukkan bahwa performa algoritma klasifikasi dapat ditingkatkan secara signifikan melalui optimasi fitur. Misalnya, Isnaini [1] membandingkan tiga algoritma dan menemukan bahwa XGBoost sedikit lebih unggul dibanding Random Forest, sedangkan Sumanto [2] mengembangkan konfigurasi spesifik pada Random Forest yang menghasilkan nilai AUC hingga 0,993. Selanjutnya, Dahlan [3] mengimplementasikan teknik *forward selection* untuk optimalisasi fitur pada algoritma Random Forest, yang terbukti mampu meningkatkan akurasi model dari 89,45% menjadi 92,46%, dengan peningkatan presisi dan sedikit penurunan pada recall.

Berdasarkan latar belakang dan permasalahan yang telah diuraikan, penelitian ini bertujuan untuk menganalisis pengaruh kombinasi kebiasaan merokok dan aktivitas begadang terhadap risiko penyakit paru-paru dengan pendekatan *Knowledge Discovery in Database* (KDD). Untuk mencapai tujuan tersebut, penelitian ini akan melalui beberapa tahapan:

eksplorasi dan preprocessing data, seleksi fitur menggunakan metode statistik dan visualisasi, serta penerapan algoritma *Support Vector Machine* (SVM) untuk klasifikasi. Evaluasi performa model dilakukan melalui metrik akurasi, presisi, dan recall, guna memastikan efektivitas pendekatan yang digunakan dalam mendeteksi pola risiko yang tersembunyi dan kompleks.

2. TINJAUAN PUSTAKA

2.1. Penilitian Terdahulu

Penelitian oleh Lindberg dalam [4] menunjukkan bahwa individu dengan berbagai gejala insomnia (kesulitan memulai tidur, kesulitan mempertahankan tidur, terbangun dini hari) atau kantuk berlebihan di siang hari memiliki kemungkinan lebih rendah untuk mencapai keberhasilan berhenti merokok jangka panjang, bahkan setelah memperhitungkan berbagai faktor lain seperti usia, BMI, dan kondisi kesehatan; sementara itu, pada subjek yang awalnya tidak memiliki gangguan tidur, merokok terus-menerus meningkatkan risiko kesulitan memulai tidur selama masa tindak lanjut dibandingkan dengan mereka yang telah berhenti merokok, menunjukkan hubungan dua arah di mana insomnia mempersulit upaya berhenti merokok dan merokok itu sendiri dapat memicu masalah tidur, sehingga penanganan gangguan tidur sebaiknya dimasukkan dalam program berhenti merokok untuk memutus siklus negatif ini. dari metode penelitian adalah memformulasikan permasalahan yang diteliti dengan lebih rinci (sedapat mungkin ditulis secara matematis) dan menjelaskan metode yang diusulkan. Apabila menggunakan sebuah algoritma, dapat dijelaskan di bagian ini, beserta dengan *state of the art.*

Penelitian yang dilakukan oleh Isnaini dalam [1]menggunakan algoritma Random Forest, XGBoost, dan Logistic Regression untuk menentukan algoritma yang paling efektif dalam melakukan prediksi penyakit paru-paru. Hasil dari tiap algoritma menghasilkan tingkat akurasi yang berbeda-beda dengan pengukuran F-score 93,70% untuk Random Forest; 93,83% untuk XGBoost; 86,88% untuk Logistic Regression. Prediksi ini didasarkan pada dataset yang mencakup faktor-faktor utama, seperti usia, jenis kelamin, bekerja, kebiasaan merokok, rumah tangga, aktivitas begadang, aktivitas olahraga, asuransi, penyakit bawaan.

Penelitian [2]oleh Sumanto merancang menggunakan algoritma Random Forest untuk deteksi dan prediksi cerdas penyakit paru-paru. Algoritma Random Forest diimplementasikan dengan 10 pohon keputusan dan enam atribut yang dipertimbangkan pada setiap pembagian data.Model ini diuji menggunakan validasi silang dengan 10 lipatan, dan hasil pengujian menunjukkan nilai AUC sebesar 0,993, yang mengindikasikan tingkat akurasi yang sangat tinggi. Matriks kebingungan digunakan untuk mengevaluasi kinerja model, dengan mengukur akurasi, presisi, recall, F1-Score, dan AUC. Model ini menunjukkan akurasi yang tinggi, dengan nilai ROC AUC 0,453 untuk prediksi adanya penyakit paru-paru dan 0,547 untuk prediksi ketiadaan penyakit paru-paru. Hasil ini menunjukkan bahwa algoritma Random Forest dapat menjadi alat yang efektif dalam mengidentifikasi penyakit paru-paru.

2.2. Machine Learning

Machine Learning (ML) telah menjadi bidang yang berkembang pesat dalam beberapa tahun terakhir, didorong oleh kemajuan dalam komputasi paralel dan distribusi, serta ketersediaan data dalam jumlah besar. Subasi et al.[5] memberikan tinjauan komprehensif tentang lanskap modern ML, mencakup pembelajaran mesin tradisional, pembelajaran terdistribusi, dan pembelajaran federasi. Mereka menyoroti bagaimana algoritma ML diterapkan dalam berbagai domain, termasuk penelitian ilmiah dan produk konsumen, serta tantangan yang dihadapi dalam implementasi model-model tersebut.

2.3. Support Vector Machine (SVM)

Salah satu algoritma ML yang banyak digunakan adalah Support Vector Machine (SVM). SVM dikenal karena kemampuannya dalam klasifikasi dan regresi dengan mencari hyperplane optimal yang memisahkan data ke dalam kelas-kelas yang berbeda. Dalam konteks medis, SVM telah diterapkan untuk mendiagnosis berbagai penyakit. Sebagai contoh, penelitian oleh Yuliadi et al. [6] menunjukkan penerapan algoritma SVM dalam diagnosis retinopati diabetik, yang menghasilkan akurasi tinggi dalam mendeteksi penyakit tersebut.

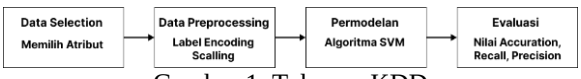
2.4. Prediksi Penyakit

Penerapan ML dalam prediksi penyakit telah menjadi fokus utama dalam penelitian kesehatan. Liu et al. [7] menyajikan survei tentang penggunaan ML untuk prediksi risiko penyakit menular, membahas berbagai model yang digunakan, termasuk pendekatan statistik, data-driven, dan epidemiologi-inspirasi. Selain itu, studi oleh Nurrokhman [8] membandingkan algoritma SVM dan Neural Network dalam klasifikasi penyakit hati, menemukan bahwa SVM memiliki kinerja lebih baik dengan akurasi mencapai 87,65%.

3. METODE PENELITIAN

Penelitian ini melalui tahapan dari metode Knowledge Discovery in Database yang didasarkan pada lima tahapan, Data Selection, Data Preprocessing, Data Transformation, Data Modelling, dan Evaluation.

3.1. Knowledge Discovery in Database



Gambar 1. Tahapan KDD

Metodologi penelitian yang digambarkan pada Gambar 1 ini dirancang melalui serangkaian tahapan

sistematis yang dimulai dengan seleksi data melalui pemilihan atribut yang relevan. Proses praproses data selanjutnya mencakup dua prosedur kritis, yaitu label encoding untuk mentransformasi variabel kategorik dan scaling untuk mengstandarisasi rentang nilai numerik, yang keduanya bertujuan mempersiapkan dataset agar kompatibel dengan analisis mendalam. Pemodelan selanjutnya mengimplementasikan algoritma Support Vector Machine (SVM), sebuah metode klasifikasi canggih yang mampu mengeksplorasi kompleksitas hubungan antarvariabel dengan mengoptimalkan hyperplane pemisah. Evaluasi akhir penelitian difokuskan pada tiga metrik fundamental: akurasi, recall, dan presisi, yang secara komprehensif akan mengukur kapabilitas model dalam memprediksi risiko penyakit paru-paru berdasarkan kebiasaan merokok dan pola aktivitas begadang.

3.2. Pengambilan Dataset

Penelitian ini menggunakan dataset mengenai prediksi atau deteksi penyakit paru paru berdasarkan kriteria yang terdapat di dalam dataset. Dataset ini diperoleh dari website Kaggle dengan dataset bernama Dataset Predict Terkena Penyakit Paru-paru.

3.3. Seleksi Fitur

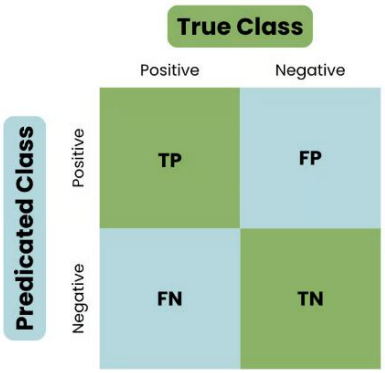
Feature selection merupakan teknik pemilihan fitur, *feature selection* dapat digunakan untuk menyeleksi banyaknya fitur yang tidak relevan pada dataset sebagaimana yang dinyatakan oleh Parenden dalam [9]. Teknik ini diterapkan pada penelitian sebagai teknik memilih fitur yang relevan dalam prediksi penyakit paru-paru. Proses dimulai dengan fase data selection yang melibatkan pengambilan dan pemilahan data secara selektif serta penerapan pelabelan yang tepat, sehingga memastikan hanya data yang relevan yang diproses untuk analisis selanjutnya, seperti halnya yang dinyatakan oleh Voutama dalam [10].

Menurut Budidarma dalam [11] label Encoding merupakan metode mengubah data bertipe kategorik menjadi format numerik. Dalam data yang terdapat pada fitur Merokok di dalam dataset, nilai "Ya" dikodekan menjadi 1 dan "Tidak" menjadi 0, sementara dalam fitur Bekerja, nilai "Bekerja" dapat dikodekan sebagai 1 dan "Tidak Bekerja" sebagai 0.

3.4. Support Vector Machine

Yahya dalam [12] mengungkap bahwa *Support Vector Machine (SVM)* merupakan metode machine learning yang digunakan untuk klasifikasi dan regresi dengan prinsip utama mencari bidang pemisah terbaik dalam ruang berdimensi N atau jumlah data yang digunakan. Sebelum membangun melatih model menggunakan SVM, peneliti memilih model yang diambil dari hasil hyperparameter tuning dan dilatih menggunakan data training yang telah disiapkan.

3.5. Pengukuran Model



Gambar 2. Confusion Matrix

Gambar 2 menampilkan bentuk matriks dari Confusion Matrix. Suryati dalam [13] menyatakan bahwa tahap evaluasi model merupakan pengukuran tingkat akurasi model yang telah dibangun dengan data latih. Lalu melakukan analisis performa model dalam mengklasifikasikan individu berisiko dan tidak berisiko menggunakan confusion matrix dan matrix accuracy, precisison, recall, dan f1-score. Dalam Confusion matrix memiliki empat istilah untuk mengevaluasi akurasi dan kesalahan prediksi dalam model klasifikasi yaitu True Positive (TP) terjadi ketika model berhasil mengklasifikasikan kelas positif dengan benar, sedangkan True Negative (TN) menunjukkan bahwa model dapat memprediksi kelas negatif secara akurat. False Positive (FP) terjadi apabila kelas negatif diklasifikasikan sebagai positif, sedangkan False Negative (FN) terjadi ketika model salah mengklasifikasikan kelas positif sebagai negatif, sebagaimana yang diungkap oleh Wafa dalam [14].

4. HASIL DAN PEMBAHASAN

Dalam penelitian ini, penulis menyajikan hasil dan pembahasan yang diperoleh menggunakan algoritma *Support Vector Machine (SVM)*. Analisis data dilakukan menggunakan metode pembelajaran mesin di Google Colab dengan bahasa pemrograman Python. Model dikembangkan dengan menggunakan beberapa library utama, seperti pandas untuk pengolahan data, numpy untuk perhitungan numerik, serta matplotlib dan seaborn untuk visualisasi data.

4.1. Pengambilan Data

Pengumpulan Data Tahap pertama penelitian ini mengumpulkan data, peneliti memanfaatkan dataset public yang tersedia di kaggle.com dataset tersebut berisi data prediksi terkena penyakit paru-paru, Dataset ini terdiri dari 10 Atribut dan salah satunya adalah atribut class, data terdiri dari 30.001 baris.

4.2. Prapemrosesan Data

Tabel 1. Feature Selection

Sebelum	Sesudah
(['No', 'Usia', 'Jenis_Kelamin', 'Merokok', 'Bekerja', 'Rumah_Tangga', 'Aktivitas_Begadang', 'Aktivitas_Olahraga', 'Asuransi', 'Penyakit_Bawaan', 'Hasil'], dtype='object')	(['Usia', 'Jenis_Kelamin', 'Merokok', 'Bekerja', 'Rumah_Tangga', 'Aktivitas_Begadang', 'Aktivitas_Olahraga', 'Asuransi', 'Penyakit_Bawaan', 'Hasil'], dtype='object')

Proses preprocessing yang diperlihatkan di Tabel 1 dilakukan untuk memastikan data yang digunakan sebelum pelatihan model memiliki kualitas yang baik untuk dianalisis. Tahap ini mencakup, Feature Selection yang berperan penting dalam memilih fitur-fitur yang relevan dan mengeliminasi fitur yang redundan sehingga meningkatkan efisiensi komputasi serta mengurangi dimensionalitas data.

Tabel 2. Label Encoding

Sebelum	Sesudah
Aktivitas_Olahraga	Aktivitas_Olahraga
Pasif	1
Aktif	0
Aktif	0
Aktif	0
Pasif	1

Label Encoding yang diperlihatkan pada Tabel 2 merupakan langkah krusial untuk mengubah data kategorikal menjadi bentuk numerik yang dapat diproses oleh algoritma pembelajaran mesin, memungkinkan model untuk memahami dan mempelajari pola dari data kategorikal tersebut. Data Scaling berfungsi untuk menstandarisasi rentang nilai fitur-fitur numerik yang berbeda menjadi skala yang seragam, mencegah fitur dengan nilai besar mendominasi perhitungan dan memastikan kontribusi setiap fitur setara dalam proses pembelajaran model, sehingga meningkatkan stabilitas dan konvergensi algoritma yang digunakan.

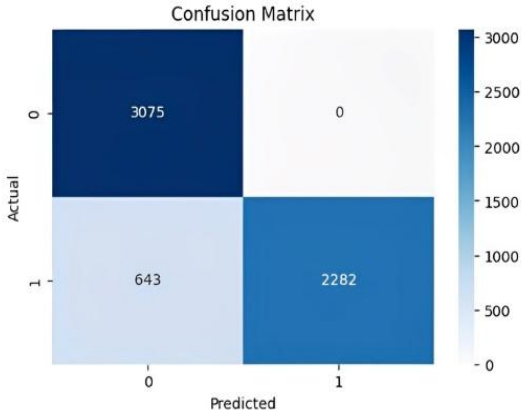
4.3. Pemodelan Data

Model SVM melakukan klasifikasi dengan dua langkah utama. Pertama, model terbaik diambil dari hasil GridSearchCV sebelumnya (best_estimator_) dan dilatih menggunakan data training yang telah disiapkan. Setelah pelatihan selesai, program memberikan konfirmasi melalui pesan "Optimized Model training completed". Kedua, model yang telah dioptimalkan langsung digunakan untuk membuat prediksi terhadap data pengujian (X_test), menghasilkan label kelas yang diprediksi dan menampilkan 30 prediksi pertama sebagai sampel. Melalui pendekatan ini, program secara efisien mengintegrasikan proses training dan prediksi dalam alur kerja yang kohesif untuk menghasilkan klasifikasi yang lebih akurat berdasarkan parameter optimal yang telah ditemukan sebelumnya.

Tabel 3. Prediksi

Predictions: [0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0]

Representasi dari Tabel 3 merupakan hasil prediksi dari model SVM teroptimalisasi yang menunjukkan 30 label kelas pertama dari data pengujian. Angka 0 dan 1 yang muncul pada larik mengindikasikan klasifikasi biner, di mana setiap sampel data diklasifikasikan ke dalam satu dari dua kategori.



Gambar 3 Confusion Matrix Model

Gambar 3 menampilkan matriks konfusi yang memvisualisasikan prediksi model. Dari 3075 sampel kelas 0, model memprediksi semuanya dengan benar (true positives), tetapi dari 2925 sampel kelas 1, model salah mengklasifikasikan 643 sebagai kelas 0 (false negatives) dan hanya memprediksi 2282 dengan benar (true positives). Hal ini menjelaskan mengapa recall untuk kelas 1 lebih rendah (0,78) meskipun memiliki presisi sempurna.

4.4. Evaluasi Model

Performa model SVM teroptimalisasi diukur melalui perhitungan metrik-metrik evaluasi yang komprehensif. Program mengkalkulasi nilai akurasi dengan membandingkan prediksi model (y_pred) terhadap nilai sebenarnya (y_test), memberikan persentase klasifikasi yang benar dari keseluruhan data pengujian. Selanjutnya, program menghasilkan classification report yang menampilkan metrik presisi, recall, dan f1-score untuk setiap kelas, memberikan gambaran lebih mendalam tentang performa model. Metrik-metrik evaluasi ini memudahkan pengembang untuk mengidentifikasi kekuatan dan kelemahan model dalam mengklasifikasikan data dari kelas-kelas berbeda. Keluaran dari bagian ini berupa nilai akurasi numerik dan tabel classification report yang memfasilitasi analisis komprehensif terhadap kinerja model. Informasi evaluasi ini sangat berharga untuk pengambilan keputusan terkait penerimaan model atau kebutuhan penyempurnaan lebih lanjut.

Tabel 4. Evaluasi Model

Kelas	Precision	Recall	F1-Score	Support
0	0.83	1.00	0.91	3.075
1	1.00	0.78	0.88	2.925
Accuracy	-	-	0.89	6.000
Macro Avg.	0.91	0.89	0.89	6.000

Evaluasi model SVM yang diperoleh di Tabel 4 menunjukkan akurasi sebesar 89,28%. Model menunjukkan kinerja yang seimbang, di mana kelas 0 memiliki presisi 0,83 dan recall sempurna 1,00, sedangkan kelas 1 memiliki presisi sempurna 1,00 tetapi recall lebih rendah 0,78. F1-score untuk kedua kelas cukup tinggi (0,91 dan 0,88), menunjukkan keseimbangan yang baik antara presisi dan recall.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengembangkan model prediktif berbasis Support Vector Machine (SVM) yang mencapai akurasi 89,28% dalam menganalisis pengaruh simultan kebiasaan merokok dan aktivitas begadang terhadap risiko penyakit paru-paru, dengan kinerja seimbang yang ditunjukkan oleh F1-score 0,91 dan 0,88 untuk kedua kelas klasifikasi, mengkonfirmasi bahwa kombinasi kedua faktor gaya hidup tersebut signifikan meningkatkan risiko gangguan paru-paru. Untuk pengembangan lebih lanjut, direkomendasikan mengeksplorasi teknik optimalisasi fitur seperti forward selection yang telah terbukti meningkatkan akurasi dalam penelitian sejenis, menggabungkan SVM dengan algoritma lain dalam kerangka ensemble learning untuk memperbaiki nilai recall pada kelas berisiko, serta mengintegrasikan data tambahan seperti riwayat penyakit keluarga dan faktor lingkungan, dengan hasil penelitian ini berpotensi diimplementasikan dalam sistem pendukung keputusan klinis dan kampanye kesehatan publik setelah divalidasi pada dataset yang lebih representatif, mendorong pengembangan pendekatan interdisipliner yang menggabungkan aspek biomedis, perilaku, dan komputasional untuk pencegahan dan penanganan penyakit paru-paru secara holistik.

DAFTAR PUSTAKA

[1] B. S. C. Putra, I. Tahyudin, B. A. Kusuma, and K. N. Isnaini, "Efektivitas Algoritma Random Forest, XGBoost, dan Logistic Regression dalam Prediksi Penyakit Paru-paru," *Techno.Com*, vol. 23, no. 4, pp. 909–922, Nov. 2024, doi: 10.62411/tc.v23i4.11705.

[2] A. Christian, Hariyanto, A. Yani, and Sumanto, "ANALISIS MACHINE LEARNING UNTUK PREDIKSI PENYAKIT PARU-PARU MENGGUNAKAN RANDOM FOREST," *Journal of Innovation And Future Technology (IFTECH)*, vol. 7, no. 1, pp. 122–133, Feb. 2025.

[3] F. Hamami and I. A. Dahlan, "KLASIFIKASI CUACA PROVINSI DKI JAKARTA MENGGUNAKAN ALGORITMA RANDOM FOREST DENGAN TEKNIK OVERSAMPLING," *Jurnal Teknoinfo*, vol. 16, no. 1, pp. 87–92, Jan. 2022, doi: 10.33365/jti.v16i1.1533.

[4] S. A. Hägg *et al.*, "Smokers with insomnia symptoms are less likely to stop smoking," *Respir Med*, vol. 170, no. 106069, Jun. 2020, doi: 10.1016/j.rmed.2020.106069.

[5] O. Subasi, O. Bel, J. Manzano, and K. Barker, "The Landscape of Modern Machine Learning: A Review of Machine, Distributed and Federated Learning," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.03120

[6] Y. Yuliadi, F. Dzil Ikram, M. Julkarnain, F. Hamdan, and H. Nuryadi, "Application of the Support Vector Machine (SVM) Algorithm for the Diagnosis of Diabetic Retinopathy," *Brilliance: Research of Artificial Intelligence*, vol. 3, no. 2, pp. 416–422, Jan. 2024, doi: 10.47709/brilliance.v3i2.3436.

[7] M. Liu, Y. Liu, and J. Liu, "Machine Learning for Infectious Disease Risk Prediction: A Survey," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.03037

[8] J. Khatib Sulaiman and U. Amikom Yogyakarta, "Perbandingan Algoritma Support Vector Machine dan Neural Network untuk Klasifikasi Penyakit Hati Ma'mur Zaky Nurrokhman," *Indonesian Journal of Computer Science Attribution*, vol. 12, no. 4, p. 2096.

[9] N. Parenden, Nurfadillah, and M. F. B. Bunga, "ANALISIS PERBANDINGAN METODE FEATURE SELECTION BACKWARD METHOD DAN STEPWISE METHOD," *CENDERAWASIH Journal of Statistics and Data Science*, vol. 2, no. 2, pp. 80–82, Jun. 2024.

[10] D. L. Rianti, Y. Umaidah, and A. Voutama, "Tren Marketplace Berdasarkan Klasifikasi Ulasan Pelanggan Menggunakan Perbandingan Kernel Support Vector Machine," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 6, no. 1, pp. 98–105, Aug. 2021, doi: 10.30998/string.v6i1.9993.

[11] C. Herdian, A. Kamila, and I. G. Agung Musa Budidarma, "Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi," *Technologia: Jurnal Ilmiah*, vol. 15, no. 1, pp. 93–108, Jan. 2024, doi: 10.31602/tji.v15i1.13457.

[12] R. Deswandi Yahya, S. Adi Wibowo, and N. Vendyansyah, "ANALISIS SENTIMEN UNTUK DETEKSI UJARAN KEBENCIAN PADA MEDIA SOSIAL TERKAIT PEMILU 2024 MENGGUNAKAN METODE SUPPORT VECTOR MACHINE," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 2, pp. 1182–1189, Apr. 2024, doi: 10.36040/jati.v8i2.9076.

[13] S. Styawati, A. Nurkholis, A. A. Aldino, S. Samsugi, E. Suryati, and R. P. Cahyono,

"Sentiment Analysis on Online Transportation Reviews Using Word2Vec Text Embedding Model Feature Extraction and Support Vector Machine (SVM) Algorithm," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, IEEE, Jan. 2022, pp. 163–167. doi: 10.1109/ISMODE53584.2022.9742906.

[14] H. S. W. Hovi, A. Id Hadiana, and F. Rakhmat Umbara, "Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM)," *Informatics and Digital Expert (INDEX)*, vol. 4, no. 1, pp. 40–45, Jan. 2023, doi: 10.36423/index.v4i1.895