Vol. 7 No.02 2025
E-ISSN: 2685-6921
Universitas Ma'arif Nahdlatul Ulama Kebumen

Evaluasi Model *Machine learning* untuk Prediksi Keparahan Kanker Berdasarkan Data *Real-world* Global

Sudriyanto Sudriyanto[1*], Abdul Fatah[1], Moh Dafa Wahna Putra[1]

[1]*Universitas Nurul Jadid, Probolinggo, Indonesia*

sudriyanto@unuja.ac.id[*]

Received: 26/11/2025 | *Revised: 04/12/2025* | *Accepted: 16/12/2025*

Abstrak

Kanker merupakan salah satu penyebab utama kematian di dunia dan menimbulkan beban yang besar bagi sistem kesehatan. Informasi mengenai tingkat keparahan kanker sangat penting untuk mendukung penentuan prioritas penanganan dan perencanaan sumber daya. Penelitian ini bertujuan membangun dan membandingkan model klasifikasi tingkat keparahan kanker berbasis pembelajaran mesin menggunakan data pasien kanker global periode 2015–2024. Dataset terdiri dari 50.000 pasien dengan berbagai atribut demografis, gaya hidup, lingkungan, klinis, serta skor keparahan (Target *Severity Score*). Dataset yang digunakan dalam penelitian ini berasal dari platform data terbuka Kaggle (www.kaggle.com) yang memuat data pasien kanker global periode 2015–2024. Skor keparahan dikonversi menjadi variabel biner dengan dua kelas, yaitu keparahan rendah dan tinggi. Tahapan penelitian meliputi pra-proses data (pembersihan, transformasi variabel kategorikal dengan *one-hot encoding*, standardisasi), pembagian data menjadi data latih dan data uji dengan proporsi 80:20 secara stratified, serta pembangunan tiga model klasifikasi, yaitu Logistic Regression, K-Nearest Neighbors (K-NN), dan Support Vector Machine (SVM) dengan kernel RBF. Kinerja model dievaluasi menggunakan akurasi, presisi, recall, F1-score, dan confusion matrix, serta divalidasi dengan 5-fold cross validation. Hasil percobaan menunjukkan bahwa *Logistic regression* menghasilkan akurasi 99,82%, presisi 99,86%, recall 99,78%, dan F1-score 99,82%, dengan kesalahan klasifikasi yang sangat kecil. SVM memperoleh akurasi 98,22% dengan kinerja yang juga tinggi, sedangkan K-NN hanya mencapai akurasi sekitar 79,42%. Hasil validasi silang mengonfirmasi bahwa *Logistic regression* memiliki rata-rata akurasi tertinggi dan paling stabil. Dengan demikian, *Logistic regression* direkomendasikan sebagai model utama untuk prediksi tingkat keparahan kanker pada dataset ini dan berpotensi dikembangkan lebih lanjut sebagai komponen sistem pendukung keputusan klinis.

Kata kunci: kanker, pembelajaran mesin, regresi logistik, k-nearest neighbors (K-NN), support vector machine (SVM), klasifikasi keparahan

Abstract

Cancer is one of the leading causes of death worldwide and places a significant burden on healthcare systems. Information on cancer severity is crucial for prioritizing treatment and resource planning. This study aims to develop and compare machine learning-based cancer severity classification models using global cancer patient data from 2015–2024. The dataset comprises 50,000 patients with various demographic, lifestyle, environmental, and clinical attributes, as well as severity scores (Target Severity Score). The dataset used in this study was obtained from the open data platform Kaggle (www.kaggle.com), which contains global cancer patient data from 2015 to 2024. The severity score is converted into a binary variable with two classes: low and high severity. The research steps include data preprocessing (cleaning, categorical transformation of variables with one-hot encoding, standardization), data division into training and testing data with a stratified 80:20 ratio, and the development of three classification models: Logistic Regression, K-Nearest Neighbors (K-NN), and Support Vector Machine (SVM) with RBF kernel. Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix, and validated with 5-fold cross-validation. Experimental results showed that Logistic regression achieved 99.82% accuracy, 99.86% precision, 99.78% recall, and 99.82% F1-score, with very small classification errors. SVM achieved 98.22% accuracy with also high performance, while K-NN only achieved an accuracy of around 79.42%. Cross-validation results confirmed that Logistic regression had the highest average accuracy and the most stability. Thus, Logistic regression is recommended as the primary model for predicting cancer severity in this dataset and has the potential for further development as a component of a clinical decision support system.

Keywords: cancer, machine learning, Logistic Regression, K-Nearest Neighbors, Support Vector Machine

Pendahuluan

Kanker masih menjadi salah satu penyebab utama kematian di dunia dan menjadi beban kesehatan global yang terus meningkat. Laporan epidemiologi menunjukkan tren peningkatan insidensi dan mortalitas hampir di seluruh kawasan, terutama di negara berpendapatan menengah yang mengalami transisi demografi dan epidemiologi (Cruz & Wishart, 2006; Kourou et al., 2015). Selain membahayakan nyawa, kanker menimbulkan konsekuensi sosial-ekonomi yang serius bagi pasien, keluarga, dan sistem kesehatan di berbagai negara (Hero, 2021).

Beban kanker yang tinggi tersebut tidak terlepas dari kompleksitas faktor risiko yang terlibat, mulai dari faktor genetik, gaya hidup, hingga faktor lingkungan dan sosial-ekonomi. Variasi kombinasi faktor-faktor tersebut menyebabkan perbedaan yang cukup besar dalam angka kesakitan dan angka kesintasan (*survival*) antar individu maupun antar negara (Cruz & Wishart, 2006; Kourou et al., 2015). Selain tingginya angka kesakitan dan kematian, kanker juga menimbulkan tantangan tersendiri dalam penatalaksanaan klinis. Dokter dituntut tidak hanya menetapkan diagnosis, tetapi juga menilai tingkat keparahan penyakit untuk menentukan prioritas tindakan, estimasi kebutuhan sumber daya, dan komunikasi risiko kepada pasien. Penilaian

tingkat keparahan yang akurat sangat penting dalam penentuan pilihan terapi, penjadwalan tindak lanjut, dan perencanaan pembiayaan layanan kesehatan (Cruz & Wishart, 2006; Delen et al., 2005).

Perkembangan teknologi informasi dan ketersediaan data kesehatan berskala besar membuka peluang pemanfaatan pendekatan *machine learning* untuk mendukung penilaian tingkat keparahan kanker. Berbagai studi menunjukkan bahwa model *machine learning* mampu memanfaatkan pola non-linier dan interaksi antar variabel yang sulit ditangkap oleh pendekatan statistik tradisional (Cruz & Wishart, 2006; Esteva et al., 2017; Kourou et al., 2015). Di Indonesia, penerapan *machine learning* pada diagnosis dan klasifikasi penyakit juga mulai berkembang, termasuk untuk penyakit kanker (Cahyana & Nurlayli, 2023; Desiani et al., 2025; Juliani, 2024; Mubarog et al., 2019; Septiany et al., 2024; Warnilah et al., 2024).

Dalam konteks tersebut, ketersediaan data pasien kanker global periode satu dekade terakhir yang memuat variabel demografis, klinis, dan komorbiditas memberikan kesempatan untuk membangun model prediksi tingkat keparahan berbasis data *real-world*. Pada penelitian ini digunakan dataset sekunder pasien kanker global yang diperoleh dari platform data terbuka Kaggle (www.kaggle.com) (Kaggle, 2024), yang memuat informasi demografis, klinis, komorbiditas, dan faktor risiko lainnya. Penggunaan data *real-world* seperti ini penting untuk menghasilkan model yang lebih mencerminkan variasi klinis sehari-hari dan berpotensi tetap andal ketika diterapkan pada populasi yang lebih luas.

Pengelompokan keparahan kanker ke dalam kategori "tinggi" dan "rendah" pada dasarnya merupakan masalah klasifikasi biner. Berbagai algoritma telah digunakan untuk permasalahan serupa, seperti *logistic regression*, *k-nearest neighbors* (K-NN), *decision tree*, *random forest*, dan *support vector machine* (SVM) (Cahyana & Nurlayli, 2023; Desiani et al., 2025; Delen et al., 2005; Kourou et al., 2015; Panda et al., 2022). *Logistic regression* sering dipilih karena kemudahan interpretasi koefisien dan kemampuan menghasilkan probabilitas kejadian, sehingga banyak digunakan dalam penelitian klinis (Panda et al., 2022). Di sisi lain, K-NN dan SVM menawarkan kemampuan menangkap hubungan non-linier yang lebih kompleks dan berpotensi memberikan kinerja klasifikasi yang lebih baik pada pola data tertentu (Chazar & Widhiaputra, 2020; Kourou et al., 2015).

Meskipun demikian, sebagian besar penelitian sebelumnya baik di tingkat nasional maupun internasional lebih banyak berfokus pada klasifikasi jenis kanker atau status jinak-ganas, serta seringkali menggunakan ukuran dataset yang relatif terbatas (Adiningrum et al., 2023; Chazar & Widhiaputra, 2020; Maulani & Fatah, 2025; Mubarog et al., 2019; Nurnawati, 2022; Oktafiani et al., 2023; Wardhana et al., 2023). Kajian yang secara sistematis membandingkan kinerja beberapa algoritma *machine learning* pada data pasien kanker *real-world* berskala besar dengan fokus khusus pada klasifikasi tingkat keparahan masih relatif sedikit, padahal informasi tingkat keparahan sangat penting untuk pengambilan keputusan klinis dan manajerial (Cruz & Wishart, 2006; Delen et al., 2005; Kourou et al., 2015).

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja tiga algoritma machine learning, yaitu *logistic regression*, K-NN, dan SVM kernel RBF, dalam memprediksi tingkat keparahan kanker (tinggi dan rendah) menggunakan data pasien kanker global berskala besar. *Logistic regression* dipilih karena banyak

digunakan dalam penelitian klinis, mudah diinterpretasikan, dan mampu menghasilkan estimasi risiko individual yang jelas bagi klinisi. K-NN dipilih untuk merepresentasikan pendekatan berbasis jarak yang sensitif terhadap pola lokal pada data, sedangkan SVM dengan kernel RBF mewakili algoritma non-linier berkapasitas tinggi yang mampu menangkap batas pemisah kelas yang kompleks. Berbeda dengan penelitian yang berfokus pada algoritma pohon keputusan seperti decision tree atau random forest, studi ini secara sengaja memusatkan perhatian pada tiga algoritma dasar yang secara luas direkomendasikan sebagai baseline kuat pada data klinis, sehingga hasilnya lebih mudah dibandingkan dengan literatur yang ada dan dapat menjadi pijakan sebelum mengevaluasi model yang lebih kompleks. Pemilihan ketiga algoritma ini juga mempertimbangkan ketersediaan dataset *real-world* berskala besar dari repositori terbuka Kaggle (Kaggle, 2024), yang lazim digunakan sebagai benchmark untuk studi komparatif antar model dasar.

Metodologi Penelitian

Metode penelitian yang digunakan dalam studi ini adalah penelitian kuantitatif dengan pendekatan supervised *machine learning* untuk masalah klasifikasi biner. Variabel target adalah tingkat keparahan kanker yang dibagi menjadi dua kelas, yaitu keparahan rendah dan keparahan tinggi, sedangkan variabel prediktor terdiri atas karakteristik demografis, faktor klinis, komorbiditas, dan faktor terkait lainnya. Data yang digunakan merupakan data *real-world* pasien kanker global dengan total 50.000 entri yang telah melalui proses pembersihan awal (cleaning) dan penyesuaian format, dan diperoleh dari platform data terbuka Kaggle (www.kaggle.com) (Kaggle, 2024).

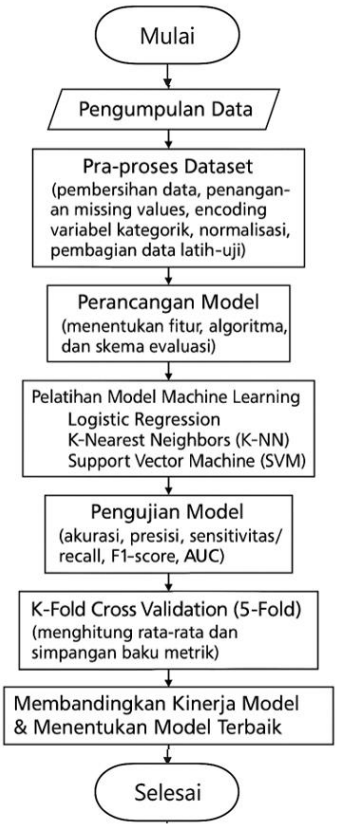
Algoritma pertama yang dievaluasi adalah *logistic regression*. Secara konseptual, *Logistic regression* memodelkan hubungan antara variabel prediktor dan probabilitas kejadian kelas tertentu melalui fungsi logit, sehingga menghasilkan parameter yang dapat ditafsirkan sebagai log odds dari kejadian tersebut (Panda et al., 2022). Pendekatan ini banyak digunakan dalam penelitian klinis karena mampu memberikan estimasi risiko individual dan interval kepercayaan yang jelas, serta relatif stabil pada ukuran sampel besar (Cruz & Wishart, 2006; Delen et al., 2005).

Algoritma kedua adalah k-*nearest neighbors* (K-NN). K-NN melakukan klasifikasi suatu observasi baru berdasarkan mayoritas kelas dari k tetangga terdekatnya di ruang fitur. Nilai k dipilih melalui proses hyperparameter tuning menggunakan validasi silang, dengan mempertimbangkan kompromi antara varians dan bias model. K-NN sering digunakan sebagai pembanding pada studi klasifikasi medis karena implementasinya sederhana dan sensitif terhadap struktur lokal data, sehingga dapat memberikan gambaran awal mengenai separabilitas kelas pada ruang fitur (Kourou et al., 2015).

Algoritma ketiga adalah *support vector machine* (SVM) dengan kernel *radial basis function* (RBF). SVM bekerja dengan mencari hyperplane pemisah yang memaksimalkan margin antar kelas pada ruang berdimensi tinggi, sedangkan penggunaan kernel RBF memungkinkan model untuk menangkap hubungan non-linier antara variabel prediktor dan kelas *output* (Chazar & Widhiaputra, 2020; Kourou et al., 2015). SVM banyak dilaporkan memiliki kinerja yang baik pada berbagai permasalahan klasifikasi kanker, terutama ketika jumlah fitur cukup besar dan

terdapat kemungkinan batas pemisah kelas yang tidak linier (Cruz & Wishart, 2006; Kourou et al., 2015).

Tahapan penelitian dimulai dari pengumpulan dan integrasi data, dilanjutkan dengan pra-proses meliputi penanganan *missing values*, pengkodean variabel kategorik menggunakan *one-hot encoding*, serta normalisasi fitur numerik agar berada pada rentang skala yang sebanding. Selanjutnya, data dibagi menjadi data latih dan data uji dengan perbandingan 80:20 menggunakan teknik stratified sampling untuk menjaga proporsi kelas keparahan tinggi dan rendah tetap seimbang pada kedua subset. Ketiga algoritma kemudian dilatih pada data latih dan dievaluasi pada data uji menggunakan metrik akurasi, presisi, sensitivitas (*recall*), F1-score, dan *Area Under the ROC Curve* (AUC). Untuk menilai kestabilan kinerja model, dilakukan pula validasi silang (*5-fold cross validation*) sehingga setiap observasi memperoleh kesempatan yang seimbang untuk menjadi bagian dari data latih maupun data uji (Cruz & Wishart, 2006; Kourou et al., 2015). Adapun alur penelitian ditampilkan pada Gambar 1.



Gambar 1. Metode Research and Development (R&D)

Pengumplan Data

Penelitian diawali dengan tahap pengumpulan data, yaitu mengambil dataset sekunder pasien kanker global periode 2015–2024 yang tersimpan dalam berkas *global_cancer_patients_2015_2024.csv*. Dataset ini diunduh dari repositori data terbuka Kaggle (www.kaggle.com) (Kaggle, 2024) dan berisi 50.000 baris data pasien dari berbagai negara dengan sejumlah atribut yang merepresentasikan karakteristik demografis (usia, jenis kelamin, negara/region, tahun), faktor risiko genetik dan lingkungan (skor risiko genetik dan tingkat polusi

udara), faktor gaya hidup (tingkat konsumsi alkohol, kebiasaan merokok, dan tingkat obesitas), faktor klinis dan ekonomi (jenis kanker, stadium kanker, biaya pengobatan dalam dolar Amerika, serta lama kesintasan dalam tahun), serta satu variabel skor keparahan kanker (*Target Severity Score*). Seluruh data tersebut menjadi dasar bagi pembangunan model klasifikasi tingkat keparahan kanker pada tahap-tahap berikutnya.

Pra-Proses Dataset

Setelah data terkumpul, dilakukan pra-proses dataset untuk memastikan data siap diolah oleh algoritma klasifikasi. Langkah pertama adalah pembersihan data, yaitu menghapus atribut yang tidak memiliki nilai prediktif, dalam hal ini kolom identitas pasien (Patient ID) yang hanya berfungsi sebagai nomor unik. Pada tahap ini juga dilakukan pemeriksaan terhadap adanya nilai yang hilang atau tidak wajar pada setiap atribut; bila ditemukan, data tersebut ditangani sesuai kaidah pra-proses (misalnya melalui imputasi atau pengabaian), sehingga dataset yang digunakan pada tahap pemodelan berada dalam kondisi konsisten dan bersih.

Langkah berikutnya adalah pembentukan variabel target. Awalnya, tingkat keparahan kanker direpresentasikan dalam bentuk skor kontinu (*Target Severity Score*). Untuk keperluan klasifikasi biner, skor ini kemudian dikonversi menjadi dua kategori: keparahan rendah dan keparahan tinggi. Batas pemisah ditetapkan menggunakan nilai median skor keparahan pada seluruh data. Pasien dengan skor sama atau di atas median diklasifikasikan ke dalam kategori keparahan tinggi (label 1), sedangkan pasien dengan skor di bawah median dimasukkan ke dalam kategori keparahan rendah (label 0). Pendekatan ini menghasilkan distribusi kelas yang relatif seimbang antara kategori rendah dan tinggi, sehingga mengurangi potensi bias model terhadap salah satu kelas. Setelah variabel target biner terbentuk, skor kontinu semula tidak lagi digunakan sebagai fitur dan dikeluarkan dari himpunan prediktor.

Tahap pra-proses kemudian dilanjutkan dengan transformasi variabel kategorikal. Atribut seperti jenis kelamin, negara/*region*, jenis kanker, dan stadium kanker pada mulanya berupa data kategori (teks) yang tidak dapat langsung diolah oleh algoritma numerik. Oleh karena itu, setiap kategori dikonversi menjadi representasi numerik melalui teknik *one-hot encoding* sehingga tiap kategori diwakili oleh variabel biner tersendiri. Untuk mencegah terjadinya multikolinearitas, satu kategori referensi pada tiap atribut dihilangkan (pendekatan *drop first*), sementara kategori lainnya tetap dimasukkan. Hasil transformasi ini membuat seluruh fitur berada dalam bentuk numerik dengan jumlah atribut meningkat menjadi puluhan variabel prediktor.

Selanjutnya dilakukan pemisahan data menjadi data latih dan data uji. Sebanyak 80% dari seluruh data digunakan sebagai data latih untuk proses pembelajaran model, sedangkan 20% sisanya digunakan sebagai data uji untuk mengukur kemampuan generalisasi model. Pembagian dilakukan dengan pendekatan *stratified sampling*, yaitu mempertahankan proporsi kelas keparahan rendah dan tinggi yang sama pada data latih dan data uji, sehingga distribusi kelas tetap seimbang di kedua subset. Sebelum masuk ke tahap pemodelan, seluruh fitur numerik distandarisasi menggunakan teknik standardisasi (mengubah skala fitur sehingga memiliki rataan nol dan simpangan baku satu). Standardisasi ini sangat penting, terutama untuk algoritma yang sensitif terhadap skala data seperti *K-Nearest Neighbors* dan *Support Vector Machine*, agar tidak ada satu fitur yang mendominasi perhitungan jarak maupun pembentukan hiperbidang pemisah.

Perancangan dan Implementasi Model

Tahap berikutnya adalah perancangan dan implementasi model klasifikasi. Penelitian ini membandingkan tiga algoritma *supervised learning*, yaitu *Logistic Regression*, *K-Nearest Neighbors* (K-NN), dan *Support Vector Machine* (SVM) dengan kernel *Radial Basis Function* (RBF). Ketiga algoritma ini dipilih karena mewakili metode yang umum digunakan pada penelitian medis dan telah terbukti efektif untuk masalah klasifikasi biner.

Pada *Logistic Regression*, hubungan antara variabel prediktor dan probabilitas pasien memiliki keparahan tinggi dimodelkan melalui fungsi logit. Model ini menggunakan regularisasi L2 untuk mencegah *overfitting*, dengan parameter regulasi disetel pada nilai moderat sehingga menyeimbangkan kompleksitas model dan kemampuan generalisasi. *Logistic regression* diharapkan memberikan interpretasi yang jelas mengenai pengaruh relatif setiap fitur terhadap peluang keparahan tinggi.

Pada algoritma K-NN, kelas suatu pasien ditentukan berdasarkan mayoritas kelas dari sejumlah tetangga terdekat dalam ruang fitur. Nilai jumlah tetangga (misalnya k = 5) ditetapkan sebagai konfigurasi awal berdasarkan praktik umum pada literatur. Semua fitur yang telah distandarisasi digunakan untuk menghitung jarak antar pasien menggunakan metrik jarak Euclidean. Metode ini diharapkan mampu menangkap pola lokal dalam data, yaitu kedekatan karakteristik antar pasien yang memiliki tingkat keparahan serupa.

Sedangkan pada SVM, penelitian ini menggunakan kernel RBF yang mampu memetakan data ke ruang berdimensi lebih tinggi sehingga pemisahan kelas yang tidak linear dapat dilakukan menggunakan hiperbidang yang optimal. Parameter regulasi dan parameter kernel diatur pada nilai yang lazim digunakan sebagai titik awal (misalnya nilai C yang menyeimbangkan margin dan kesalahan klasifikasi, serta gamma yang mengatur jangkauan pengaruh masing-masing titik data). SVM diharapkan mampu menghasilkan batas pemisah yang kuat antara kelas keparahan rendah dan tinggi pada ruang fitur yang kompleks.

Ketiga algoritma tersebut diimplementasikan dengan skema yang seragam: *input* data terlebih dahulu distandarisasi, kemudian dimasukkan ke algoritma klasifikasi sesuai konfigurasi masing-masing. Hal ini memastikan bahwa perbandingan performa antar model dilakukan secara adil dengan pra-proses yang sama.

Pengujian Model

Setelah model selesai dilatih menggunakan data latih, tahap berikutnya adalah pengujian model pada data uji. Pada tahap ini, setiap observasi pada data uji dimasukkan ke dalam model untuk memperoleh prediksi kelas keparahan, kemudian hasil prediksi tersebut dibandingkan dengan label sebenarnya untuk menghitung berbagai metrik kinerja. Penggunaan data uji yang tidak pernah dilihat model selama pelatihan bertujuan untuk memperoleh estimasi kinerja yang lebih objektif dan bebas dari *overfitting*.

Metrik utama yang digunakan meliputi akurasi, presisi, sensitivitas (*recall*), F1-score, dan *Area Under the ROC Curve* (AUC) untuk kelas keparahan tinggi. Selain itu, confusion matrix juga ditampilkan untuk masing-masing model guna memberikan gambaran yang lebih rinci mengenai pola kesalahan, khususnya terkait kesalahan klasifikasi pada pasien dengan keparahan tinggi. Dalam konteks klinis, kesalahan mengklasifikasikan pasien yang sebenarnya berat sebagai

ringan (*false negative*) dianggap lebih serius dibandingkan kesalahan sebaliknya, sehingga sensitivitas dan F1-score kelas keparahan tinggi memperoleh perhatian khusus (Cruz & Wishart, 2006; Kourou et al., 2015).

Validasi Silang (K-Fold Cross Validation)

Untuk memastikan bahwa kinerja model tidak hanya bergantung pada satu kali pembagian data latih dan data uji, dilakukan validasi silang *(k-fold cross validation)*. Pada penelitian ini digunakan skema 5-*fold*, di mana keseluruhan data dibagi secara acak menjadi lima lipatan (*fold*) dengan proporsi kelas yang relatif seimbang. Pada setiap iterasi, empat *fold* digunakan sebagai data latih dan satu *fold* sisanya sebagai data uji, sehingga setiap data mendapat giliran berada di set uji tepat satu kali.

Proses pelatihan dan pengujian diulang sebanyak lima kali untuk masing-masing model dan pada setiap pengulangan dihitung nilai akurasi, presisi, recall, F1-score, dan AUC. Dari kelima nilai tersebut kemudian dihitung rata-rata dan simpangan baku untuk menilai kestabilan kinerja model. Model yang baik tidak hanya memiliki rata-rata kinerja tinggi, tetapi juga variabilitas yang rendah antar lipatan, yang menunjukkan bahwa model tidak terlalu sensitif terhadap variasi dalam pembagian data (Cruz & Wishart, 2006; Kourou et al., 2015).

Pembandingan Hasil Model dan Penarikan Kesimpulan

Tahap terakhir penelitian adalah pembandingan hasil model dan penarikan kesimpulan. Pada tahap ini, seluruh metrik yang diperoleh dari pengujian pada data uji dan validasi silang dibandingkan untuk menentukan model yang paling unggul secara keseluruhan. Selain mempertimbangkan kinerja kuantitatif, pembahasan juga memperhatikan aspek interpretabilitas model, kemudahan implementasi di lingkungan klinis, serta implikasi praktisnya bagi pengambilan keputusan. Hasil perbandingan tersebut kemudian dirangkum menjadi rekomendasi model yang paling tepat digunakan sebagai dasar pengembangan sistem pendukung keputusan untuk prediksi keparahan kanker.

Hasil dan Pembahasan

Hasil Evaluasi Model pada Data Uji

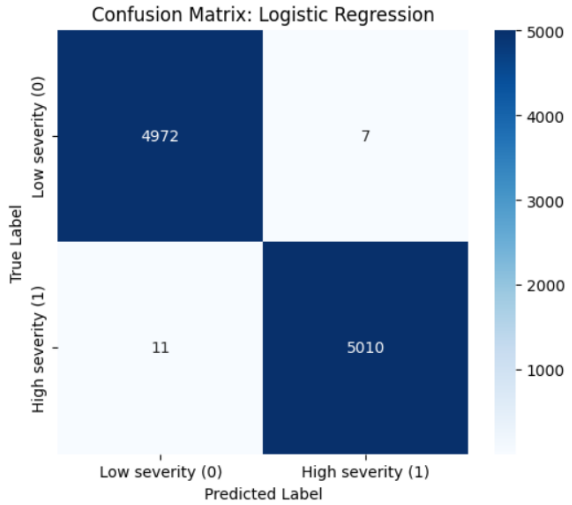
Data sebanyak 50.000 record dibagi menjadi 40.000 data latih dan 10.000 data uji dengan komposisi kelas yang seimbang antara kategori keparahan rendah (*severity_high* = 0) dan keparahan tinggi (*severity_high* = 1). Tiga model klasifikasi yang diuji adalah *Logistic Regression*, *K-Nearest Neighbors* (K-NN), dan *Support Vector Machine* (SVM) dengan kernel RBF.

Ringkasan kinerja ketiga model pada data uji ditunjukkan pada Tabel 1. Metrik yang digunakan meliputi akurasi, presisi, recall, dan F1-score untuk kelas keparahan tinggi sebagai kelas positif.

Tabel 1. Hasil evaluasi model pada data uji

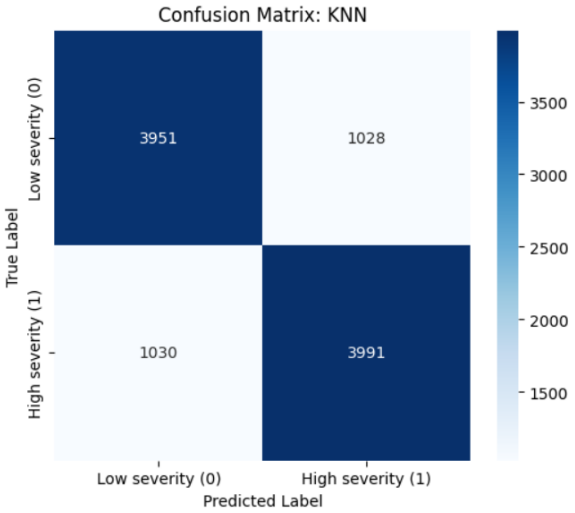
Model	Akurasi	Presisi	Recall	F1-score
Logistic Regression	0,9982	0,9986	0,9978	0,9982
K-NN	0,7942	0,7952	0,7949	0,795
SVM (RBF)	0,9822	0,984	0,9805	0,9822

Berdasarkan Tabel 1 terlihat bahwa *Logistic regression* memberikan kinerja paling tinggi dibandingkan dua model lainnya, dengan akurasi hampir 100% dan nilai presisi, *recall*, serta F1-score yang sangat seimbang. Model SVM (RBF) juga menunjukkan performa yang sangat baik dengan akurasi di atas 98%, sedangkan K-NN memiliki akurasi dan F1-score sekitar 79% sehingga jauh tertinggal dari dua model lainnya. Visualisasi confusion matrix untuk masing-masing model ditunjukkan pada Gambar 2, Gambar 3, dan Gambar 4.



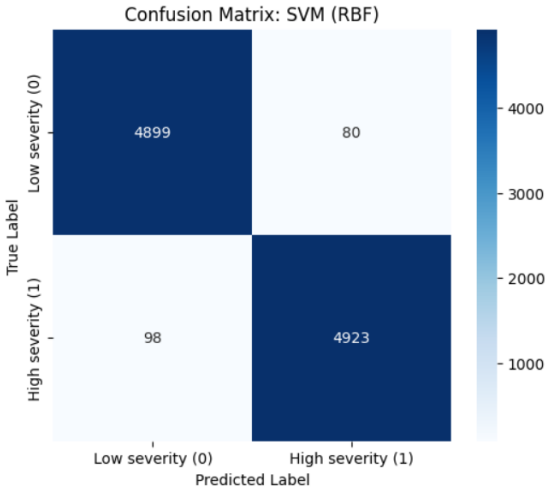
Gambar 2. Confusion matrix model *Logistic regression* pada data uji

Logistic regression menghasilkan 4.972 prediksi benar untuk kelas keparahan rendah dan 5.010 prediksi benar untuk kelas keparahan tinggi. Jumlah kesalahan klasifikasi sangat kecil, yaitu 7 *false positive* (keparahan rendah diprediksi tinggi) dan 11 *false negative* (keparahan tinggi diprediksi rendah). Dominasi nilai pada diagonal utama menunjukkan bahwa model mampu memisahkan kedua kelas secara sangat akurat.



Gambar 3. *Confusion matrix model K-Nearest Neighbors* (K-NN) pada data uji

K-NN memperlihatkan pola yang berbeda. Meskipun jumlah prediksi benar masih cukup besar (3.951 untuk keparahan rendah dan 3.991 untuk keparahan tinggi), terdapat lebih dari 1.000 *false positive* dan lebih dari 1.000 *false negative*. Besarnya nilai di luar diagonal utama menjelaskan mengapa metrik kinerja K-NN jauh lebih rendah dibandingkan dua model lainnya.



Gambar 4. *Confusion matrix model Support Vector Machine* (SVM) kernel RBF pada data uji

SVM (RBF) (Gambar 4) menunjukkan hasil yang berada di antara *Logistic regression* dan K-NN. Model ini memberikan 4.899 prediksi benar untuk kelas keparahan rendah dan 4.923 prediksi benar untuk kelas keparahan tinggi, dengan 80 *false positive* dan 98 *false negative*. Jumlah kesalahan ini masih sedikit dan sejalan dengan nilai akurasi serta F1-score yang tinggi. Dari hasil tersebut dapat disimpulkan bahwa pada dataset ini *Logistic regression* merupakan model yang paling efektif dalam mengklasifikasikan tingkat keparahan kanker menjadi dua kategori, diikuti oleh SVM, sedangkan K-NN kurang mampu menangkap pola pemisahan kelas secara optimal.

Hasil Validasi Silang

Untuk menilai kestabilan kinerja model, dilakukan validasi silang *(k-fold cross validation)* dengan nilai k = 5 dan skema *stratified* sehingga proporsi kelas pada tiap lipatan tetap seimbang. Nilai rata-rata akurasi dan simpangan baku dari setiap model ditunjukkan pada Tabel 2.

Tabel 2. Hasil 5-fold cross validation

Model	Rata-rata Akurasi	Simpangan Baku
Logistic Regression	0,9984	0,0002
K-NN	0,7924	0,0049
SVM (RBF)	0,9838	0,0006

Tabel 2 menunjukkan bahwa *Logistic regression* kembali memperoleh rata-rata akurasi tertinggi, yaitu 99,84%, dengan simpangan baku yang sangat kecil. Hal ini mengindikasikan bahwa performa model stabil pada berbagai pembagian data. SVM juga menunjukkan performa tinggi dengan rata-rata akurasi 98,38% dan variasi yang relatif kecil. Sebaliknya, K-NN memiliki rata-rata akurasi sekitar 79,24% dengan simpangan baku yang lebih besar, yang menandakan kinerja yang kurang konsisten dibandingkan dua model lainnya.

Hasil validasi silang ini memperkuat temuan pada evaluasi data uji, yaitu bahwa *Logistic regression* dan SVM memiliki kemampuan generalisasi yang baik, sedangkan K-NN kurang cocok digunakan untuk karakteristik data yang digunakan pada penelitian ini.

Pembahasan

Secara keseluruhan, hasil penelitian menunjukkan bahwa model linear *Logistic regression* justru mampu mengungguli model non-linear SVM dalam memprediksi tingkat keparahan kanker. Beberapa faktor yang dapat menjelaskan fenomena ini adalah sebagai berikut.

Pertama, tingginya performa *Logistic regression* hingga mendekati 100% pada dataset ini dapat dijelaskan oleh beberapa faktor. Variabel-variabel prediktor yang digunakan relatif kaya dan informatif, mencakup faktor demografis, gaya hidup, lingkungan, klinis, dan ekonomi yang secara teoritis berkaitan erat dengan tingkat keparahan kanker. Selain itu, pembentukan variabel target melalui kategorisasi berdasarkan nilai median *Target Severity Score* menghasilkan dua kelas yang relatif seimbang, sehingga model tidak mengalami masalah ketidakseimbangan kelas yang sering menurunkan kinerja. Pra-proses data yang sistematis—meliputi pembersihan data, transformasi variabel kategorikal dengan *one-hot encoding*, serta standardisasi fitur numerik—membantu algoritma menemukan batas pemisah yang jelas antara kelas keparahan rendah dan tinggi. Dikombinasikan dengan ukuran sampel yang besar (50.000 pasien), kondisi ini membuat model linear seperti *Logistic regression* mampu mempelajari parameter secara akurat tanpa menunjukkan gejala *overfitting*.

Kedua, dimensi fitur yang cukup tinggi akibat proses *one-hot encoding* membuat metode berbasis jarak seperti K-NN kurang efektif. Pada ruang berdimensi tinggi, perbedaan jarak antar titik cenderung menjadi kurang informatif sehingga K-NN sulit menemukan tetangga yang benar-

benar representatif untuk menentukan kelas. Hal ini tercermin dari tingginya jumlah *false positive* dan *false negative* pada model K-NN.

Ketiga, dalam konteks medis, fokus utama bukan hanya akurasi keseluruhan, tetapi terutama kemampuan model dalam mengurangi kesalahan klasifikasi pada pasien dengan keparahan tinggi. Dari sisi ini, *Logistic regression* menghasilkan jumlah *false negative* paling sedikit (11 kasus) dibandingkan SVM (98 kasus) maupun K-NN (1.030 kasus). Dengan kata lain, *Logistic regression* lebih baik dalam mendeteksi pasien yang benar-benar berada pada kondisi berat, sehingga lebih sesuai untuk digunakan sebagai dasar sistem pendukung keputusan klinis.

Temuan ini sejalan dengan berbagai penelitian sebelumnya yang menunjukkan bahwa model yang relatif sederhana namun terkalibrasi dengan baik sering kali memberikan kinerja yang kompetitif pada data klinis *real-world*. Delen et al. (2005) dan Panda et al. (2022) melaporkan bahwa *Logistic regression* mampu bersaing dengan metode yang lebih kompleks dalam memprediksi luaran klinis, terutama ketika jumlah fitur cukup besar dan hubungan antar variabel tidak sepenuhnya non-linier. Hasil penelitian di Indonesia juga menunjukkan bahwa regresi logistik dan tree-based methods sering kali memberikan kombinasi terbaik antara akurasi dan interpretabilitas pada klasifikasi kanker payudara (Cahyana & Nurlayli, 2023; Desiani et al., 2025; Nurnawati, 2022; Oktafiani et al., 2023; Wardhana et al., 2023; Warnilah et al., 2024). Dengan demikian, dominasi *Logistic regression* pada penelitian ini memperkuat bukti bahwa model linier terkalibrasi tetap relevan untuk digunakan dalam sistem pendukung keputusan klinis, khususnya ketika interpretabilitas menjadi faktor penting.

Dengan demikian, dapat disimpulkan bahwa kombinasi pra-proses data yang tepat, pemilihan fitur yang relevan, serta penggunaan algoritma klasifikasi yang sesuai mampu menghasilkan model prediksi tingkat keparahan kanker dengan performa yang sangat baik. Berdasarkan hasil evaluasi dan pertimbangan konteks klinis, *Logistic regression* direkomendasikan sebagai model utama dalam penelitian ini, sementara SVM dapat dipertimbangkan sebagai alternatif dan K-NN kurang disarankan untuk digunakan pada dataset sejenis. Namun demikian, hasil ini memiliki sejumlah keterbatasan: model hanya dievaluasi pada satu dataset sekunder dari Kaggle sehingga sangat bergantung pada karakteristik data tersebut, variabel keparahan direduksi menjadi dua kategori biner sehingga berpotensi menyederhanakan kondisi klinis yang kompleks, dan belum dilakukan validasi eksternal maupun penalaan hiperparameter yang lebih ekstensif. Keterbatasan-keterbatasan ini perlu diperhatikan ketika menginterpretasikan hasil dan mendorong perlunya penelitian lanjutan pada dataset dan pengaturan yang berbeda.

Kesimpulan

Penelitian ini bertujuan mengembangkan dan membandingkan tiga model klasifikasi, yaitu *Logistic Regression, K-Nearest Neighbors* (K-NN), dan *Support Vector Machine* (SVM), dalam memprediksi tingkat keparahan kanker berdasarkan data pasien kanker global tahun 2015–2024 yang diperoleh dari *platform* data terbuka Kaggle (www.kaggle.com). Berdasarkan hasil evaluasi pada data uji dan validasi silang, diperoleh bahwa *Logistic regression* merupakan model dengan kinerja terbaik dengan akurasi mencapai 99,82%, nilai presisi, *recall*, dan F1-score yang sangat tinggi serta jumlah kesalahan klasifikasi yang minimal. SVM menunjukkan performa yang juga tinggi dengan akurasi 98,22%, namun masih menghasilkan *false negative* lebih besar

dibanding Logistic Regression. Sementara itu, K-NN menghasilkan akurasi yang jauh lebih rendah, yaitu sekitar 79%, dan menunjukkan tingkat kesalahan klasifikasi yang tinggi sehingga kurang sesuai digunakan pada dataset berdimensi tinggi seperti pada penelitian ini. Secara keseluruhan, *Logistic regression* dinilai paling efektif, stabil, dan mudah diinterpretasikan sehingga direkomendasikan sebagai pendekatan utama dalam memprediksi tingkat keparahan kanker. Hasil penelitian ini menunjukkan bahwa kombinasi pra-proses data yang tepat dan pemilihan algoritma yang sesuai dapat menghasilkan model prediksi yang sangat akurat, dan temuan ini berpotensi digunakan sebagai dasar pengembangan sistem pendukung keputusan untuk analisis risiko keparahan kanker di masa mendatang. Namun demikian, penelitian ini memiliki beberapa keterbatasan, antara lain penggunaan satu sumber data sekunder dari Kaggle yang membuat hasil sangat bergantung pada karakteristik dataset tersebut, keterbatasan variabel klinis yang belum mencakup informasi rinci seperti biomarker dan hasil pencitraan, serta penalaan hiperparameter yang masih dilakukan pada tingkat dasar. Keterbatasan-keterbatasan ini diharapkan dapat menjadi acuan bagi penelitian lanjutan untuk menghasilkan model prediksi keparahan kanker yang lebih robust dan aplikatif di lingkungan klinis.

Daftar Pustaka

Adiningrum, N. T. R., Rianti, R., & Priyanto, C. (2023). Rancang bangun aplikasi prediksi kanker payudara dengan pendekatan machine learning. Jurnal Informatika dan Teknik Elektro Terapan, 11(3s1). https://doi.org/10.23960/jitet.v11i3s1.3351

Cahyana, C. W., & Nurlayli, A. (2023). Analisis performa logistic regression, naïve Bayes, dan random forest sebagai algoritma pendeteksi kanker payudara. INSERT: Information System and Emerging Technology Journal, 4(1), 51–64. https://doi.org/10.23887/insert.v4i1.62362

Chazar, C., & Widhiaputra, B. E. (2020). *Machine learning* diagnosis kanker payudara menggunakan algoritma Support Vector Machine. INFORMASI (Jurnal Informatika dan Sistem Informasi), 12(1), 67–80. https://doi.org/10.37424/informasi.v12i1.48

Desiani, A., Zayanti, D. A., Ramayanti, I., Ramadhan, F. F., & Giovillando. (2025). Perbandingan algoritma Support Vector Machine (SVM) dan *Logistic regression* dalam klasifikasi kanker payudara. Jurnal Kecerdasan Buatan dan Teknologi Informasi, 4(1), 33–42. https://doi.org/10.69916/jkbti.v4i1.191

Hero, S. K. (2021). Faktor resiko kanker payudara. Jurnal Medika Hutama, 3(1), 1533–1537.

Juliani, D. (2024). Implementasi *machine learning* untuk klasifikasi penyakit kanker paru menggunakan metode naïve Bayes dengan tambahan fitur chatbot. Jurnal Ilmu Pengetahuan dan Teknologi (IPTEK), 8(2). https://doi.org/10.31543/jii.v8i2.351

Kusumawaty, J., Noviati, E., Sukmawati, I., Srinayanti, Y., & Rahayu, Y. (2021). Efektivitas edukasi SADARI (pemeriksaan payudara sendiri) untuk deteksi dini kanker payudara. ABDIMAS: Jurnal Pengabdian Masyarakat, 4(1), 496–501.

Maulani, R. N., & Fatah, Z. (2025). Klasifikasi data kanker payudara menggunakan algoritma Decision Tree berbasis RapidMiner. JAMASTIKA: Jurnal Mahasiswa Teknik Informatika, 4(2). https://doi.org/10.35473/jamastika.v4i2.4504

Marfianti, E. (2021). Peningkatan pengetahuan kanker payudara dan keterampilan periksa payudara sendiri (SADARI) untuk deteksi dini kanker payudara di Semutan Jatimulyo Dlingo. Jurnal Abdimas Madani dan Lestari (JAMALI), 3(1), 25–31. https://doi.org/10.20885/jamali.vol3.iss1.art4

Mubarog, I., Setyanto, A., & Sismoro, H. (2019). Sistem klasifikasi pada penyakit breast cancer dengan menggunakan metode naïve Bayes. Creative Information Technology Journal, 6(2), 109–118.

Nurnawati, E. K. (2022). Penerapan algoritma Decision Tree untuk memprediksi kanker payudara menggunakan data mining dan machine learning. Jurnal Dinamika Informatika, 11(2), 103–112.

Oktafiani, R., Hermawan, A., & Avianto, D. (2023). Pengaruh komposisi split data terhadap performa klasifikasi penyakit kanker payudara menggunakan algoritma machine learning. Jurnal Sains dan Informasi, 9(1), 19–28. https://doi.org/10.34128/jsi.v9i1.622

Septiany, E. S., Handayani, H. H., Al Mudzakir, T., & Masruriyah, A. F. N. (2024). Optimasi metode Support Vector Machine menggunakan seleksi fitur recursive feature elimination dan forward selection untuk klasifikasi kanker payudara. TIN: Terapan Informatika Nusantara, 5(2), 144–154.

Wardhana, A., Yuliana, T., & Putri, M. (2023). Penerapan algoritma C4.5 untuk prediksi diagnosis kanker payudara. Jurnal Sains Komputer dan Informatika, 9(1), 78–87.

Warnilah, A. I., Sutisna, H., Ratningsih, R., Christian, V., & Maharani, R. (2024). Implementasi *machine learning* untuk prediksi kanker payudara menggunakan model regresi logistik. EVOLUSI: Jurnal Sains dan Manajemen, 12(2), 76–84. https://doi.org/10.31294/evolusi.v12i2.23315

Cruz, J. A., & Wishart, D. S. (2006). Applications of *machine learning* in cancer prediction and prognosis. Cancer Informatics, 2, 59–77. https://doi.org/10.1177/117693510600200030

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine, 34(2), 113–127. https://doi.org/10.1016/j.artmed.2004.07.002

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). *Machine learning* applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Panda, N. R., Pati, J. K., Mohanty, J. N., & Bhuyan, R. (2022). A review on *Logistic regression* in medical research. National Journal of Community Medicine, 13(4), 265–270. https://doi.org/10.55489/njcm.134202222

Kaggle. (2024). Global cancer patients 2015–2024 (global_cancer_patients_2015_2024.csv) [Data set]. Kaggle. https://www.kaggle.com/