

Prediksi Risiko Penyakit Kanker Paru-Paru Berdasarkan Faktor-Faktor Yang Mempengaruhinya Menggunakan Metode *Decision Tree*

Mu'tashim Billah Rahman^{1*}, Hermanto², Jarot Dwi Prasetyo³

^{1,2,3} Teknologi Informasi, Universitas Ibrahimy, Situbondo, Indonesia

¹ tashimrahman2@gmail.com*, ² otnamreh54@gmail.com, ³ jdp170779@gmail.com

INFO ARTIKEL

Article history:

Received Juni 2026

Published Juli 2026

ABSTRAK

Kanker paru-paru merupakan salah satu penyakit yang memiliki tingkat kematian cukup tinggi dan sering terlambat terdeteksi karena gejalanya baru dirasakan ketika penyakit telah berkembang. Oleh karena itu, diperlukan suatu metode yang dapat membantu melakukan deteksi dini terhadap risiko kanker paru-paru berdasarkan faktor-faktor yang memengaruhinya. Penelitian ini bertujuan untuk membangun model prediksi risiko kanker paru-paru menggunakan algoritma Decision Tree ID3. Data yang digunakan berasal dari *Lung Cancer Dataset* yang diperoleh dari Kaggle dan terdiri dari 1.000 data pasien. Atribut yang digunakan dalam penelitian ini meliputi usia, jenis kelamin, polusi udara, bahaya oksipasional, risiko genetik, penyakit paru kronis, kebiasaan merokok, nyeri dada, dan penurunan berat badan, sedangkan variabel target berupa tingkat risiko kanker paru-paru yang terbagi menjadi tiga kategori, yaitu *Low*, *Medium*, dan *High*. Proses penelitian dilakukan dengan mengikuti tahapan *Knowledge Discovery in Databases* (KDD) yang meliputi seleksi data, *preprocessing*, transformasi data, pembentukan model menggunakan algoritma Decision Tree ID3, serta evaluasi model. Model yang telah dibangun kemudian diimplementasikan ke dalam aplikasi berbasis web menggunakan Streamlit sehingga dapat dimanfaatkan sebagai alat bantu untuk memberikan informasi awal mengenai tingkat risiko kanker paru-paru secara lebih mudah dan praktis.

Kata Kunci: Kanker Paru-Paru, Decision Tree ID3, Prediksi Risiko, Data Mining, Streamlit.

ABSTRACT

Lung cancer is one of the diseases with a high mortality rate and is often diagnosed at a late stage because its symptoms are usually not recognized until the disease has progressed. Therefore, a method is needed to assist in the early detection of lung cancer risk based on the factors that influence it. This study aims to develop a lung cancer risk prediction model using the Decision Tree ID3 algorithm. The data used in this study were obtained from the Lung Cancer Dataset available on Kaggle and consisted of 1,000 patient records. The attributes used include age, gender, air pollution, occupational hazards, genetic risk, chronic lung disease, smoking habits, chest pain, and weight loss, while the target variable is the level of lung cancer risk categorized into three classes: *Low*, *Medium*, and *High*. The research process follows the stages of Knowledge Discovery in Databases (KDD), which include data selection, preprocessing, data transformation, model development using the Decision Tree ID3 algorithm, and model evaluation. The resulting model is then implemented in a web-based application using Streamlit, enabling users to obtain preliminary information regarding lung cancer risk in a more convenient and practical manner.

Keywords: Lung Cancer, Decision Tree ID3, Risk Prediction, Data Mining, Streamlit.

1. Pendahuluan

Kesehatan merupakan aspek mendasar dalam kehidupan manusia yang menentukan kualitas hidup, produktivitas, serta kesejahteraan sosial. Salah satu ancaman kesehatan yang masih menjadi perhatian serius di dunia adalah penyakit kanker, termasuk kanker paru-paru. Berdasarkan laporan *World Health Organization* (WHO), kanker paru-paru menjadi penyebab kematian tertinggi di antara berbagai jenis kanker dengan angka kematian yang terus meningkat setiap tahunnya. Di Indonesia, kondisi ini juga tercermin dengan tingginya angka kejadian kanker paru-paru, yang sebagian besar disebabkan oleh kebiasaan merokok serta paparan polusi lingkungan (Marzuq et al., 2023).

Kanker paru dalam arti luas adalah semua penyakit keganasan di paru, men cakup keganasan yang berasal dari paru itu sendiri (primer) maupun keganasan dari luar paru (metastasis). Dalam pengertian klinis yang dimaksud dengan kanker paru primer adalah tumor ganas yang berasal dari epitel bronkus (karsinoma bronkus). Prevalensi kanker paru primer di negara maju sangat tinggi. Di USA tahun 2002 dilaporkan terdapat 169.400 kasus baru (merupakan 13% dari semua kanker baru yang terdiagnosis) dengan 154.900 kematian (merupakan 28% dari seluruh kematian akibat kanker). Indonesia menduduki peringkat ke-4 kanker paru terbanyak di dunia. Tahun 1998 di Rumah Sakit (RS) Kanker Dharmas Jakarta kanker paru menduduki urutan ke-3 sesudah kanker payudara dan leher rahim. Angka kematian akibat kanker paru di seluruh dunia mencapai kurang lebih satu juta penduduk setiap tahunnya. Sebagian besar kanker paru mengenai laki-laki (65%) life time risk 1:13 dan pada perempuan 1:20. Kelompok risiko tinggi ialah laki-laki dengan usia diatas 40 tahun (Joseph & Rotty, 2020).

Kanker paru-paru timbul akibat pertumbuhan sel abnormal pada jaringan paru yang tidak terkendali. Sel kanker dapat membentuk tumor, mengganggu fungsi paru, dan bahkan menyebar ke organ tubuh lainnya. Gejala yang muncul pada tahap awal sering kali tidak spesifik, seperti batuk kering, nyeri dada, atau sesak napas, sehingga banyak kasus baru terdeteksi saat penyakit sudah berada pada stadium lanjut. Hal ini mengakibatkan tingkat keberhasilan pengobatan menjadi lebih rendah serta biaya perawatan yang semakin besar. Berbagai penelitian menunjukkan bahwa faktor risiko utama kanker paru-paru adalah kebiasaan merokok. Sekitar 85–90% kasus berhubungan langsung dengan konsumsi tembakau. Selain itu, paparan polusi udara, bahaya pekerjaan (*occupational hazard*), riwayat penyakit paru kronis, serta faktor genetik juga memberikan kontribusi signifikan terhadap peningkatan risiko. Data dari Kementerian Kesehatan Indonesia memperlihatkan bahwa angka kematian akibat kanker paru-paru pada tahun-tahun terakhir terus meningkat, menjadikannya beban besar bagi sistem kesehatan nasional (Rahman Wahid et al., 2023).

Dalam dunia medis, tahapan awal dalam pendeteksian kanker menggunakan metode analisis manual dan sistem diagnosa dapat ditingkatkan menggunakan implementasi teknik pemrosesan gambar. Namun, banyak peneliti juga sudah meningkatkan metode pendeteksian kanker ini dengan mengaplikasikan teknik analisa gambar seperti Computed Tomography (CT), Sputum Cytology, Chest S-ray dan Magnetic Imaging Resonan (Teknologi, 2023). Untuk mendeteksi adanya kanker paru-paru, dapat dilakukan serangkaian pemeriksaan laboratorium, antara lain pemeriksaan pencitraan seperti rontgen dada dan CT-scan, analisis sitologi dahak, maupun biopsi jaringan paru (Fajarido & Erlina, 2024).

Namun, seiring perkembangan teknologi informasi, analisis data medis kini dapat dimanfaatkan untuk membantu proses deteksi dini. *Data mining* dan *machine learning* memungkinkan pengolahan data dalam jumlah besar guna menemukan pola tersembunyi yang sulit diidentifikasi secara manual oleh tenaga medis. Salah satu algoritma yang populer digunakan adalah *Decision Tree* karena kemampuannya menghasilkan model yang mudah dipahami, transparan, dan dapat divisualisasikan dalam bentuk pohon keputusan. Hal ini sangat bermanfaat dalam memberikan gambaran kepada masyarakat awam maupun tenaga medis mengenai faktor-faktor yang berkontribusi terhadap risiko kanker paru-paru (Jatnika Fahmi Idris et al., 2024). Dan pengembangan aplikasi berbasis web mulai banyak dimanfaatkan untuk mengimplementasikan metode berbasis data yang mampu mengintegrasikan proses analisis dan penyajian hasil secara interaktif. Streamlit merupakan salah satu framework yang banyak digunakan untuk membangun antarmuka aplikasi berbasis data secara interaktif dan mudah diimplementasikan (Maresti et al., 2026).

Data mining (penggalian data) adalah proses menemukan pola atau informasi yang berguna dari sejumlah besar data yang tersimpan dalam database, data warehouse, atau sumber data lainnya. Tujuan dari data mining adalah untuk mengekstraksi informasi yang berharga dari data, termasuk pola, keteraturan, tren, dan anomali yang dapat membantu dalam pengambilan keputusan. Data mining melibatkan penggunaan teknik-teknik analisis statistik, matematika, kecerdasan buatan, dan mesin learning untuk

mengidentifikasi pola dan tren dalam data. Proses ini dapat melibatkan beberapa tahapan, termasuk preprocessing data, eksplorasi data, pemodelan, dan evaluasi (Permana et al., 2023).

Beberapa penelitian sebelumnya telah memanfaatkan algoritma klasifikasi untuk prediksi kanker paru-paru, namun masih terdapat kekurangan dalam hal variasi dataset maupun metode yang digunakan. Algoritma *Decision Tree* varian ID3 memiliki potensi yang besar karena mengandalkan perhitungan *Information Gain* untuk memilih atribut terbaik dalam membentuk pohon keputusan. Metode ini relatif sederhana, cepat diproses, dan hasilnya mudah diinterpretasikan. Selain itu, penerapan ID3 pada kasus kanker paru-paru masih jarang dilakukan, sehingga penelitian ini diharapkan dapat memberikan kontribusi baru, baik dari sisi akademis maupun praktis. Tantangan nyata di lapangan diperkuat oleh data RSUD dr. Zainoel Abidin Banda Aceh yang menunjukkan rata-rata keterlambatan diagnosis pasien mencapai 98 hari akibat kemiripan gejala dengan TB paru, hambatan biaya, serta jarak fasilitas kesehatan, sementara secara nasional lebih dari 70% kasus baru terdeteksi pada stadium lanjut akibat rendahnya kesadaran skrining dini (Ferry Dwi et al., 2024).

Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk merancang dan mengimplementasikan model prediksi risiko kanker paru-paru dengan algoritma *Decision Tree* ID3 menggunakan data sekunder yang tersedia di Kaggle.com. Pemilihan algoritma ID3 didasarkan pada pertimbangan bahwa konsep *Information Gain* sangat sesuai untuk karakteristik data kategorikal pada dataset kesehatan.

2. Metode Penelitian

Penelitian ini dilaksanakan menggunakan pendekatan yang terstruktur untuk memastikan model klasifikasi yang dibangun dapat memprediksi risiko penyakit secara efektif. Metodologi penelitian dibagi menjadi dua fokus utama: pengumpulan data sekunder melalui platform Kaggle.com dan analisis pemodelan menggunakan algoritma *Decision Tree* ID3. Proses analisis dilakukan dengan mengolah data kategorikal faktor risiko kanker paru-paru berdasarkan perhitungan nilai *Information Gain* untuk menghasilkan aturan keputusan yang akurat dan mudah diinterpretasikan.

2.1 Teknik Pengumpulan Data

Teknik pengumpulan data dalam penelitian ini dilakukan melalui dua pendekatan utama, yaitu akuisisi data sekunder dan studi pustaka. Data sekunder diperoleh dari platform Kaggle.com menggunakan *Lung Cancer Dataset (Early Stage Symptoms and Risk Factors)*, yang mencakup variabel klinis seperti usia, jenis kelamin, tingkat polusi udara, kebiasaan merokok, konsumsi alkohol, alergi debu, bahaya okupasional, riwayat genetik, serta gejala fisik seperti sesak napas dan batuk kering. Untuk mendukung validitas analisis data tersebut, dilakukan studi pustaka melalui literatur jurnal ilmiah, artikel, dan buku guna mendalami faktor risiko kanker paru-paru serta landasan teori komputasi. Pemahaman literatur mengenai implementasi *data mining* di bidang kesehatan ini menjadi dasar kuat dalam perancangan model, proses implementasi algoritma *Decision Tree*, hingga acuan pada tahap evaluasi sistem klasifikasi yang dibangun.

2.2 Sumber Data dan Karakteristik Dataset

Dataset yang digunakan dalam penelitian ini adalah *Lung Cancer Dataset (Early Stage Symptoms and Risk Factors)* yang bersumber dari Kaggle. Dataset ini bertipe data terstruktur dengan format matriks yang memuat total 1.000 sampel data pasien. Dari keseluruhan sampel tersebut, sistem melakukan pembagian data (*data splitting*) secara otomatis melalui parameter kontrol aplikasi dengan proporsi *Test Size* sebesar 0.20. Konfigurasi ini memisahkan data secara acak menjadi 800 sampel data latih (*training data*) untuk pembentukan pohon keputusan dan 200 sampel data uji (*testing data*) untuk menguji performa final model. Tampilan potongan data mentah (*raw data*) dari *Lung Cancer Dataset* asli yang diunduh melalui platform Kaggle ditunjukkan pada Gambar 1.

indeks	Patient Id	Usia	Jenis Ke	Polusi Ud	alkohol	Alergi De	Bahaya Oksp	Risiko g	kronis Penyak	Diet seimbang	Obesitas	Merokok	Perokok pasi	Nyeri dada	Batuk darah
0	P1	33	1	2	4	5	4	3	2	2	4	3	2	2	4
1	P10	17	1	3	1	5	3	4	2	2	2	2	4	2	3
2	P100	35	1	4	5	6	5	5	4	6	7	2	3	4	8
3	P1000	37	1	7	7	7	7	6	7	7	7	7	7	7	8
4	P101	46	1	6	8	7	7	7	6	7	7	8	7	7	9
5	P102	35	1	4	5	6	5	5	4	6	7	2	3	4	8

Kelelahan	Penurunan berat b	Sesak napas	Mengi	Kesulitan menelan	Kuku dan jari bengkak	Sering pilek	Batuk Kering	Dengkur	Tingkat
3	4	2	2	3	1	2	3	4	Low
1	3	7	8	6	2	1	7	2	Medium
8	7	9	2	1	4	6	7	2	High
4	2	3	1	4	5	6	7	5	High
3	2	4	1	4	2	4	2	3	High

Gambar 1. Potongan Dataset Asli Kanker Paru-Paru dari Kaggle.com
Sumber: Penulis, 2026

Dataset ini memiliki karakteristik informasi yang padat yang terdiri dari 9 atribut klinis utama sebagai variabel prediktor (fitur input) yang nantinya diinputkan secara manual oleh pengguna melalui halaman *frontend*

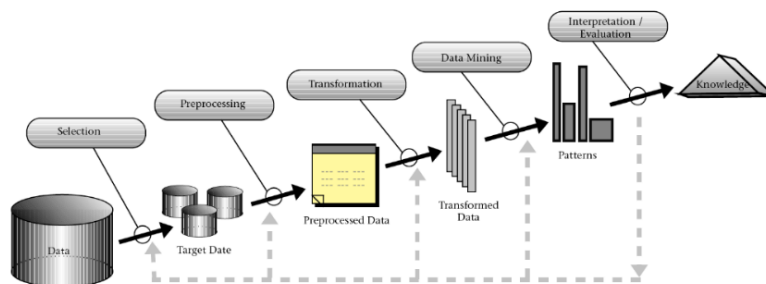
Streamlit. Sembilan atribut klinis tersebut meliputi karakteristik demografis, faktor lingkungan, kebiasaan hidup, serta gejala fisik spesifik pasien yang dirinci sebagai berikut:

- Karakteristik Demografis dan Klinis Utama: Mencakup variabel usia (*Age*) dan jenis kelamin (*Gender*).
- Faktor Risiko Lingkungan dan Medis: Mencakup tingkat paparan polusi udara (*Air Pollution*), bahaya lingkungan kerja (*Occupational Hazards*), dan faktor riwayat genetik (*Genetic Risk*).
- Kebiasaan Hidup (*Lifestyle*): Mencakup tingkat kebiasaan merokok (*Smoking*) dan intensitas konsumsi alkohol (*Alcohol Use*).
- Gejala Fisik Spesifik: Mencakup variabel riwayat alergi debu (*Dust Allergy*), serta gejala klinis yang dirasakan langsung oleh pasien seperti sesak napas (*Shortness of Breath*) dan batuk kering (*Dry Cough*).

Kesembilan atribut klinis di atas diekstraksi menggunakan metrik pembagi algoritma (*Criterion Entropy*) di area *backend* untuk mengklasifikasikan kondisi pasien ke dalam satu variabel target berupa tiga tingkatan kelas risiko kanker paru-paru, yaitu kelas *Low*, *Medium*, dan *High*.

2.3 Metode Pengembangan Sistem

Penelitian ini mengadopsi metode KDD (Knowledge Discovery in Databases) sebagai pendekatan metodologi yang di lakukan, Metode ini merupakan pendekatan sistematis yang diterapkan dalam data mining untuk menggali informasi yang bermanfaat dari data besar dan kompleks (Khairudin et al., 2025). Kerangka kerja KDD dipilih karena menyediakan tahapan sistematis untuk mengekstraksi pengetahuan dan pola tersembunyi dari dataset medis berdimensi besar. Implementasi kerangka kerja ini secara operasional digambarkan melalui diagram alir penelitian yang berfungsi sebagai acuan perancangan model dari tahap awal hingga evaluasi. Alur perancangan model prediksi risiko kanker paru-paru tersebut ditunjukkan pada Gambar 2



Gambar 2. Knowledge Discovery in Database

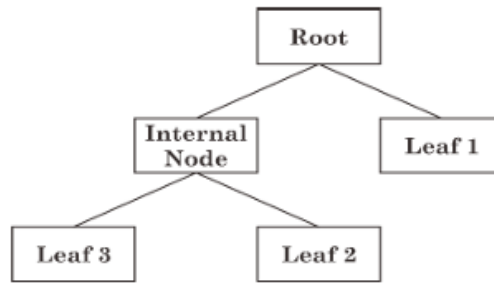
Sumber: Penulis, 2026

2.3 Perancangan Algoritma dan Pemodelan Sistem

2.3.1 Algoritma *Decision Tree* ID3

Metode ini menggunakan struktur pohon keputusan untuk memecah dataset besar menjadi bagian-bagian kecil berdasarkan aturan tertentu. Komponen utamanya terdiri dari akar (*root*) sebagai atribut paling berpengaruh di bagian teratas, simpul (*node*) yang mewakili pengujian atribut, cabang sebagai hasil pengujian, dan daun (*leaf*) yang menunjukkan kelas atau kategori keputusan akhir. Prinsip dasar dari metode ini adalah mengonversi data menjadi sebuah model pohon keputusan, kemudian menerjemahkan struktur pohon tersebut ke dalam bentuk aturan (*rule*), serta menyederhanakan aturan-aturan tersebut agar lebih mudah digunakan. Data yang dianalisis dalam pohon keputusan biasanya disajikan dalam bentuk tabel yang terdiri atas sejumlah atribut dan *record*.

Pembentukan pohon keputusan dilakukan secara rekursif dengan memilih atribut terbaik sebagai simpul akar (*root node*) maupun simpul cabang (*internal node*) berdasarkan derajat ketidakpastian data. Derajat ketidakmurnian atau ketidakpastian di dalam himpunan data awal dihitung terlebih dahulu menggunakan rumus *Entropy* yang disajikan pada Persamaan. Adapun contoh *Decision Tree* (pohon keputusan) ditunjukkan pada gambar 3.



Gambar 3. Contoh pohon keputusan

Sumber: Penulis, 2026

2.3.2 Matematika dalam Algoritma *Decision Tree*

Dalam metode *Decision Tree*, proses pemilihan terbaik untuk dijadikan cabang pohon keputusan dilakukan berdasarkan konsep matematis yaitu *Entropy* dan *Information Gain*.

- *Entropy* digunakan untuk mengukur tingkat ketidakpastian (*impurity*) dalam *Dataset*. *Entropy* didefinisikan dengan rumus sebagai berikut:

$$Entropy(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

- *Information Gain* digunakan untuk menentukan atribut terbaik yang akan dijadikan *Node* pada pohon keputusan. Rumus *Information Gain*:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Atribut-atribut medis yang dimasukkan ke dalam variabel perhitungan matematika ini disesuaikan dengan karakteristik variabel masukan dari skenario klinis pasien. Pola data masukan beserta target kelas klasifikasi yang digunakan sebagai dasar pembentukan pohon keputusan dirinci secara sistematis pada Tabel 1.

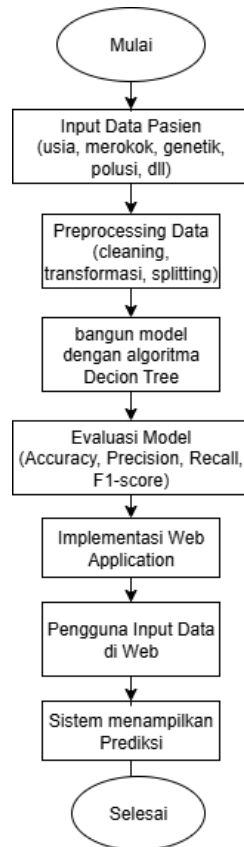
Tabel 1. Atribut-atribut prediksi manual

Nama Atribut	Jenis Data	Skala Nilai
Usia	Numerik	Angka (Tahun)
Jenis kelamin	Numerik	Laki-laki, Perempuan (1/2)
Paparan polusi udara	Numerik	1-8
Bahaya Oksipasional	Numerik	1-8
Riwayat genetik	Numerik	1-8
Penyakit paru kronis	Numerik	1-8
Kebiasaan merokok	Numerik	1-8
Nyeri dada	Numerik	1-8
Penurunan berat badan	Numerik	1-8
Tingkat Risiko (Target)	Numerik	Low, Medium, High

Setiap percabangan pohon yang terbentuk dari hasil komputasi rekursif algoritma ID3 akan menghasilkan sekumpulan aturan keputusan logis berupa fungsi implikasi (*IF-THEN rules*). Aturan-aturan keputusan inilah yang ditanamkan ke dalam logika program untuk mengeksekusi prediksi ketika pengguna menginputkan kondisi kesehatannya.

2.4 Perancangan Sistem

Perancangan sistem ini difokuskan pada pemodelan alur kerja aplikasi dan pemetaan aturan keputusan (*rules*) yang dihasilkan oleh algoritma ID3. Sistem dirancang menggunakan arsitektur berbasis web dengan memisahkan proses komputasi model (*backend*) dan interaksi pengguna (*frontend*). Aliran data pemrosesan prediksi risiko penyakit kanker paru-paru ini dirancang secara terstruktur melalui diagram alir yang tertera pada Gambar 4.



Gambar 4. Alur kerangka kerja sistem

Sumber: Penulis, 2026

Perancangan sistem ini terbagi menjadi subsistem *backend* untuk pemodelan data dan *frontend* untuk interaksi aplikasi web. Alur dimulai dengan input *Lung Cancer Dataset* yang kemudian melewati tahap *preprocessing* meliputi pembersihan, transformasi, dan pembagian data. Data siap pakai tersebut diproses menggunakan algoritma *Decision Tree* ID3 berbasis perhitungan *Entropy* dan *Information Gain* untuk membangun model prediksi. Model yang terbentuk dievaluasi menggunakan metrik akurasi (*confusion matrix*) sebelum akhirnya diekspor ke dalam format biner. Tahap akhir adalah integrasi model ke aplikasi web berbasis *Streamlit*, di mana pengguna menginput 9 data klinis untuk menampilkan hasil prediksi risiko secara *real-time*.

3. Hasil dan Pembahasan

Tahap ini memaparkan hasil implementasi dari perancangan model yang telah dilakukan serta pembahasan mengenai analisis performa algoritma dalam memprediksi tingkat risiko kesehatan pasien. Sistem prediksi risiko penyakit kanker paru-paru telah berhasil dikembangkan dengan mengintegrasikan model biner *Decision Tree* ID3 ke dalam aplikasi berbasis web menggunakan framework *Streamlit* dan bahasa pemrograman *Python*.

3.1 Hasil Implementasi Antarmuka

Implementasi antarmuka merupakan realisasi fisik dari cetak biru rancangan desain sistem yang telah dibuat sebelumnya menggunakan framework *Streamlit*. Aplikasi web ini dirancang dengan mengutamakan aspek kemudahan penggunaan agar masyarakat awam dapat melakukan deteksi risiko secara mandiri dan cepat. Halaman utama aplikasi menyediakan panel formulir interaktif di mana pengguna dapat memasukkan parameter kondisi klinis mereka, yang meliputi sembilan atribut gejala dan faktor risiko kanker paru-paru. Tampilan dari halaman utama aplikasi prediksi ini ditunjukkan pada Gambar 5.

PREDIKSI MANUAL (9 Atribut) ↔

Isikan skala 1–8 (lebih besar = lebih berisiko). Jenis kelamin otomatis menyesuaikan tipe data pada dataset.

Usia

20

Merokok (1-8)

7

Jenis Kelamin (1=Pria, 2=Wanita)

1

Polusi udara (1-8)

3

Bahaya oksipasional (1-8)

3

Risiko genetik (1-8)

3

Kronis penyakit paru-paru (1-8)

2

Penurunan berat badan (1-8)

5

Nyeri dada (1-8)

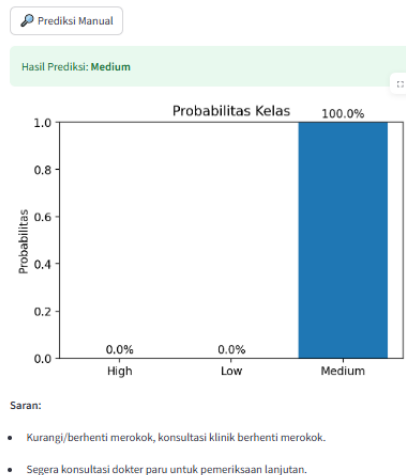
6

Prediksi Manual

Gambar 5. Tampilan halaman utama aplikasi prediksi

Sumber: Penulis, 2026

Ketika pengguna menekan tombol Prediksi Manual pada antarmuka, sistem secara *real-time* mengalirkan data input tersebut ke dalam model biner *Decision Tree* ID3 yang tertanam di *backend*. Hasil pencocokan data berdasarkan aturan pohon keputusan kemudian langsung ditampilkan pada halaman yang sama. Panel luaran (*output*) ini menampilkan klasifikasi tingkat risiko pasien dalam tiga kategori target, yaitu *Low*, *Medium*, atau *High*, disertai dengan visualisasi penjelas untuk memudahkan interpretasi medis pengguna. Hasil keluaran prediksi risiko penyakit pada antarmuka sistem ditunjukkan pada Gambar 6.



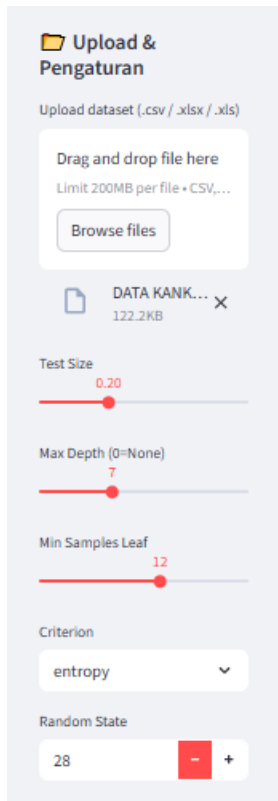
Gambar 6. Tampilan Panel Output Hasil Prediksi Risiko

Sumber: Penulis, 2026

3.2 Hasil Pengujian Model

Tahap pengujian model dilakukan untuk mengukur tingkat keandalan dan akurasi algoritma *Decision Tree* dalam mengklasifikasikan risiko kanker paru-paru. Berdasarkan konfigurasi *hyperparameter* dinamis yang diatur pada antarmuka sistem (seperti terlihat pada Gambar 7), dataset diuji menggunakan proporsi *Test Size* sebesar 0.20 (20% sebagai data uji). Pemisahan data dan pembentukan node dilakukan dengan metrik pembagi *Criterion Entropy*, pembatasan kedalaman pohon (*Max Depth*) sebesar 7, batasan minimum data pada daun (*Min Samples Leaf*) sebesar 12, serta dikunci dengan *Random State* bernilai 28. Performa dari kombinasi parameter ini dievaluasi secara mendalam berdasarkan metrik *Precision*, *Recall*,

dan *F1-Score* yang dirangkum pada Gambar 8.



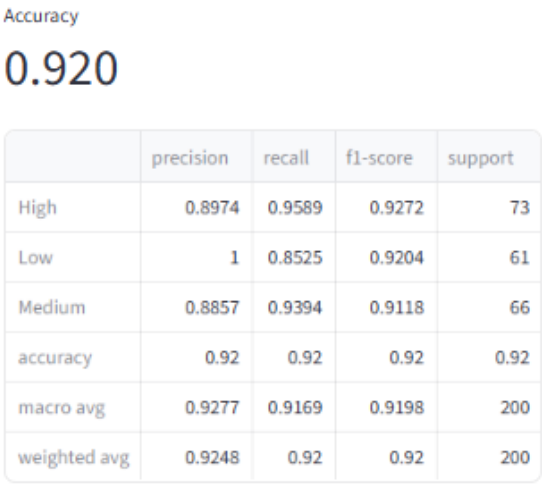
Gambar 7. Tahap pengujian model
Sumber: Penulis, 2026

3.3 Pembahasan

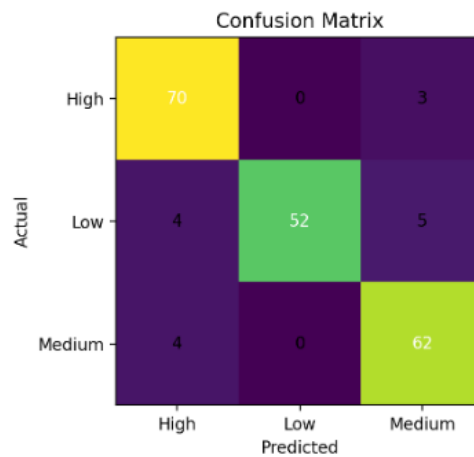
Berdasarkan visualisasi grafik *Confusion Matrix* pada Gambar 6 dan kalkulasi data pada Gambar 5, model *Decision Tree* berbasis *Entropy* ini terbukti memiliki keandalan yang sangat tinggi dalam mendeteksi tingkat risiko kanker paru-paru. Analisis mendalam terhadap grafik *Confusion Matrix* menunjukkan detail performa klasifikasi sistem sebagai berikut:

- Kelas High (Risiko Tinggi): Dari 73 data aktual pasien berisiko tinggi, sistem berhasil memprediksi secara tepat sebanyak 70 data, dan hanya 3 data yang salah diklasifikasikan ke dalam kelas *Medium*. Hal ini menghasilkan nilai *Recall* yang sangat tinggi yaitu 0.9589, menunjukkan bahwa sistem sangat sensitif dan aman dalam menjangkau pasien yang membutuhkan penanganan medis segera.
- Kelas Low (Risiko Rendah): Dari 61 data aktual pasien berisiko rendah, sistem berhasil menebak secara akurat sebanyak 52 data. Menariknya, nilai *Precision* untuk kelas ini mencapai angka sempurna yaitu 1.0000, yang membuktikan bahwa tidak ada satu pun pasien dari kelas risiko lain (*High* maupun *Medium*) yang salah deteksi ke dalam kategori risiko rendah.
- Kelas Medium (Risiko Sedang): Dari 66 data aktual, sistem berhasil memprediksi dengan benar sebanyak 62 data, sementara 4 data lainnya terprediksi masuk ke kelas *High*.

Kombinasi parameter *Max Depth* = 7 dan *Min Samples Leaf* = 12 terbukti efektif bertindak sebagai regulator untuk mencegah *overfitting*. Struktur aturan *IF-THEN* yang terbentuk dari nilai *Entropy* ini menghasilkan rata-rata akurasi keseluruhan sebesar 92,00%. Integrasi model ini ke dalam *frontend* Streamlit tidak hanya mempermudah visualisasi data evaluasi, melainkan memberikan instrumen skrining kesehatan yang valid, interaktif, dan dapat diandalkan secara *real-time* oleh pengguna.



Gambar 8. Matriks Evaluasi Performa Model Prediksi
Sumber: Penulis, 2026



Gambar 9. Grafik *Confusion Matrix* Hasil Pengujian Model Prediksi

Sumber: Penulis, 2026

2. Kesimpulan

Berdasarkan hasil penelitian dan implementasi yang telah dilakukan, dapat disimpulkan bahwa penerapan algoritma *Decision Tree* berhasil diintegrasikan ke dalam aplikasi berbasis web menggunakan *framework* Streamlit untuk memprediksi tingkat risiko penyakit kanker paru-paru. Proses pengujian model menggunakan kriteria pembagi *Entropy* dengan kombinasi parameter *Max Depth* sebesar 7 dan *Min Samples Leaf* bernilai 12 terbukti mampu menghasilkan performa klasifikasi yang optimal dan stabil. Sistem ini berhasil mencapai tingkat akurasi keseluruhan sebesar 92,00% dengan nilai *Precision* sempurna mencapai 1.0000 pada deteksi risiko rendah (*Low*) serta tingkat sensitivitas (*Recall*) tertinggi sebesar 0.9589 pada deteksi risiko tinggi (*High*). Hasil tersebut menegaskan bahwa pembatasan kedalaman struktur pohon keputusan yang diterapkan efektif dalam menekan terjadinya *overfitting*. Produk akhir berupa antarmuka web interaktif dengan fitur prediksi manual menggunakan sembilan atribut klinis ini sukses mentransformasikan komputasi *data mining* yang kompleks menjadi instrumen skrining kesehatan mandiri yang valid, praktis, dan dapat diakses secara *real-time* oleh masyarakat luas.

3. Ucapan Terima Kasih

Rasa syukur yang mendalam saya panjatkan atas selesainya penelitian ini. Terima kasih yang sebesar-besarnya saya sampaikan kepada Universitas Ibrahimy dan Fakultas Sains dan Teknologi atas segala dukungan dan bimbingan berharga yang diberikan selama ini. Apresiasi juga saya tujukan kepada komunitas Kaggle beserta para peneliti independen yang telah menyediakan *Lung Cancer Dataset* secara terbuka sehingga penelitian ini dapat berjalan dengan baik. Terakhir, terima kasih banyak untuk rekan-rekan mahasiswa Program Studi Teknologi Informasi atas segala masukan dan semangat kebersamaan selama masa pengerjaan jurnal ini.

4. Daftar Pustaka

- Fajarido, A., & Erlina, L. (2024). Metabolite Biomarker Discovery for Lung Cancer Using Machine Learning. *Indonesian Journal of Medical Chemistry and Bioinformatics*, 3(1). <https://doi.org/10.7454/ijmcb.v3i1.1028>
- Ferry Dwi, K., Dianova, S., & Rinaldy, R. (2024). Analisis Keterlambatan Diagnosis dan Terapi Kanker Paru Di RSUD Dr. Zainoel Abidin, Banda Aceh. *Journal of Medical Science*, 5(1), 1–7. <https://doi.org/10.55572/jms.v5i1.107>
- Jatnika Fahmi Idris, Rafid Ramadhani, & Muhammad Malik Mutoffar. (2024). Klasifikasi Penyakit Kanker Paru Menggunakan Perbandingan Algoritma Machine Learning. *Jurnal Media Akademik (JMA)*, 2(2). <https://doi.org/10.62281/v2i2.145>
- Joseph, J., & Rotty, L. W. A. (2020). *Kanker Paru : Laporan Kasus*. 2(1), 17–25.
- Khairudin, K., Machfud, S., & Cahyono, Y. (2025). Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Metode C4.5. *KERNEL: Jurnal Riset Inovasi Bidang Informatika Dan Pendidikan Informatika*, 5(2), 83–93. <https://doi.org/10.31284/j.kernel.2024.v5i2.7315>
- Maresti, F. A., Aulia, M. Z., Kolin, P. B., Mustaqim, K., Kunci, K., & Bantuan, D. (2026). *Implementasi Framework Streamlit Untuk Menentukan Besaran Dana Bantuan Berdasarkan Prioritas Dengan Metode Weighted Product*. 6(1), 47–58.
- Marzuq, R. D., Wicaksono, S. A., & Setiawan, N. Y. (2023). Prediksi Kanker Paru-Paru menggunakan Algoritme Random Forest Decision Tree. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 7(7), 3448–3456.
- Permana, A. Y., Fazri, H. N., & Athoilah, M. F. N. (2023). *Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest*. 3(2).

Rahman Wahid, M. A., Nugroho, A., & Halim Anshor, A. (2023). Prediksi Penyakit Kanker Paru-Paru Dengan Algoritma Regresi Linier. *Bulletin of Information Technology (BIT)*, 4(1), 63–74. <https://doi.org/10.47065/bit.v4i1.501>

Teknologi, J. (2023). *Scan*. 18.