Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No. 1 Hal: 31-38 e-ISSN: 2987-7016

Implementasi Random Forest Menggunakan Gridsearchcv Dalam Klasifikasi Penyakit Kanker Paru-Paru

Hendrawansyah Nasution[1], Evia Budianita[2], Novi Yanti[3], Jasril[4]
[1,2,3,4] Teknik Informatika, Sains dan Teknologi, Universitas Sultan Syarif Kasim II
[1]12150113914@students.uin-suska.ac.id*, [2]elvia.budianita@uin-suska.ac, [3]novi_yanti@uin-suska.ac.id, [4]jasril@uin-suska.ac.id

Abstract

Lung cancer was one of the leading causes of mortality worldwide due to the difficulty of early detection, as its initial symptoms were often non-specific. Delayed diagnosis significantly decreased patient survival rates. This study aimed to develop an accurate and efficient lung cancer classification model to support medical decision-making. The study utilized 50,000 patient records obtained from a secondary dataset, which were clinically validated by a pulmonologist at Pekanbaru General Hospital. The selected attributes were categorized into non-modifiable risk factors (age, gender, and family history) and modifiable risk factors (smoking exposure, radon exposure, asbestos exposure, COPD diagnosis, and alcohol consumption). The classification model was developed using the Random Forest algorithm and optimized using GridSearchCV to obtain the best hyperparameter configuration. Data preprocessing included label encoding, one-hot encoding, and Min-Max normalization. Experimental testing was conducted using three train-test split ratios: 90:10, 80:20, and 70:30. The results showed that the 90:10 split ratio achieved the best overall performance. The optimal model, using 300 estimators with no depth limitation (max_depth=None), achieved an accuracy of 70.3%, a Recall of 82.68%, and an F1-Score of 79.28%. The high Recall value indicated that the model was highly effective in detecting positive lung cancer cases and minimizing false negatives, making it suitable as a decision-support tool for early detection and medical intervention.

Keywords: *lung cancer, random forest, GridSearchCV, classification, machine learning*

Abstrak

Kanker paru-paru merupakan salah satu penyebab utama kematian di seluruh dunia karena sulitnya deteksi dini, karena gejala awalnya seringkali tidak spesifik. Diagnosis yang terlambat secara signifikan menurunkan angka harapan hidup pasien. Studi ini bertujuan untuk mengembangkan model klasifikasi kanker paru-paru yang akurat dan efisien untuk mendukung pengambilan keputusan medis. Studi ini menggunakan 50.000 rekam medis pasien yang diperoleh dari dataset sekunder, yang telah divalidasi secara klinis oleh seorang ahli paru di Rumah Sakit Umum Pekanbaru. Atribut yang dipilih dikategorikan menjadi faktor risiko yang tidak dapat dimodifikasi (usia, jenis kelamin, dan riwayat keluarga) dan faktor risiko yang dapat dimodifikasi (paparan merokok, paparan radon, paparan asbes, diagnosis PPOK, dan konsumsi alkohol). Model klasifikasi dikembangkan menggunakan algoritma *Random Forest* dan dioptimalkan menggunakan *GridSearchCV* untuk mendapatkan konfigurasi *hyperparameter* terbaik. Pra-pemrosesan data meliputi pengkodean label, pengkodean *one-hot*, dan normalisasi *Min-Max*. Pengujian eksperimental dilakukan menggunakan tiga rasio pembagian data latih-uji: 90:10, 80:20, dan 70:30. Hasil menunjukkan bahwa rasio pembagian 90:10 mencapai kinerja keseluruhan terbaik. Model optimal, menggunakan 300 estimator tanpa batasan kedalaman (*max_depth=None*), mencapai akurasi 70,3%, *Recall* 82,68%, dan *F1-Score* 79,28%. Nilai *Recall* yang tinggi menunjukkan bahwa model tersebut sangat efektif dalam mendeteksi kasus kanker paru-paru *positif* dan meminimalkan *false negative*, sehingga cocok sebagai alat pendukung keputusan untuk deteksi dini dan intervensi medis.

Kata kunci: kanker paru-paru, random forest, GridSearchCV, klasifikasi, machine learning

1. Pendahuluan

Paru-paru adalah organ vital dalam sistem pernapasan yang berfungsi untuk menyerap oksigen dan mengeluarkan karbon dioksida sebagai hasil metabolisme tubuh [1]. Kanker adalah suatu penyakit yang berasal dari adanya pertumbuhan sel tubuh yang progresif dan abnormal. Sampai saat ini penyakit tersebut masih menjadi masalah kesehatan karena merupakan salah satu penyebab utama tingginya angka kematian yang berkisar 8,2 juta orang, sedangkan pada tingkat nasional angka kematian kanker berkisar 5,7 % dari keseluruhan kasus kematian. Kanker dan berbagai pengobatan yang dilakukan memberikan efek samping yang berhubungan dengan permasalahan fisik & psikologi[2]. Kanker paru-paru didefinisikan sebagai kondisi di mana zat karsinogen memicu pertumbuhan dan pembelahan sel yang tidak terkendali di dalam paru-paru. Strategi esensial untuk meminimalisir angka kematian akibat kanker ini adalah dengan berfokus pada pencegahan dan deteksi dini [3]. Situasi di Indonesia menunjukkan kanker paru-paru merupakan jenis kanker terbanyak pada laki-laki dan terbanyak kelima untuk semua jenis kanker pada perempuan sedangkan di Amerika Serikat kanker paru adalah

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No. 1 Hal: 31-38 e-ISSN: 2987-7016

kanker paling umum kedua yang didiagnosis pada perempuan dan mencakup 26% dari perkiraan kematian akibat kanker pada tahun 2012 yang lebih besar dari jumlah kematian akibat kanker payudara dan usus besar atau rectum [4].

Tantangan utama dalam penanganan penyakit ini adalah gejala awal kanker paru-paru seringkali tidak spesifik yang menyebabkan sebagian besar penderita mengabaikannya sebagai gangguan pernapasan umum yang berujung pada penundaan diagnosis dan pengobatan yang tepat[5]. Kurangnya kesadaran di kalangan profesional kesehatan untuk melakukan pemeriksaan lanjutan juga memperburuk situasi ini serta meningkatkan tingkat keparahan dan mortalitas penyakit. Gangguan pada paru-paru ini dapat mengurangi efisiensi organ dalam menyerap oksigen dari udara. Oleh karena itu deteksi dini dan intervensi cepat sangat krusial dalam mengurangi angka kematian akibat kanker paru-paru[6]. Adapun faktor yang dapat menjadi penyebab kanker paru pada orang tidak merokok diantaranya asbestos radon dan karna polusi udara[4]. Munculnya penyakit kanker paru-paru di tubuh manusia ada beberapa gejala yang dapat dirasakan oleh kebanyakan pasien kanker paru-paru. Namun gejala ini seringkali diabaikan dan tidak terdeteksi oleh praktisi medis sehingga hanya 14% dari pasien yang didiagnosa kanker paru-paru sembuh dari penyakitnya[7]. Melihat kompleksitas faktor risiko dan gejala kanker paru-paru diperlukan pendekatan analisis berbasis data untuk mendukung deteksi dini dan prediksi penyakit ini salah satunya melalui penerapan metode data mining. Data mining yang biasa disebut sebagai *knowledge discovery in database* atau KDD merupakan kegiatan pengumpulan dan penggunaan data historis untuk menemukan keteraturan dan pola hubungan dalam kumpulan data yang sangat besar[8] . Dalam pengolahan data mining diperlukan tahapan pemilihan dan *preprocessing* data yang baik, baik itu untuk klasifikasi ataupun untuk fungsi yang lainnya [9]. Beberapa penelitian terdahulu telah berupaya memecahkan masalah ini. Pada penelitian yang berjudul prediksi penyakit kanker paru-paru dengan algoritma regresi linear hasil dari penelitian ini menunjukkan bahwa algoritma regresi linier dapat digunakan untuk memprediksi kemungkinan terkena kanker paru-paru dengan akurasi sekitar 90% dan mampu memberikan hasil yang baik dengan nilai Root Mean Squared Error sebesar 0.686 dan Squared Error senilai 0.471 [10]. Sedangkan pada penelitian lain yang menerapkan metode klasifikasi decision tree dalam memprediksi kanker paru-paru menggunakan algoritma C4.5 hasil klasifikasi yang kuat dari model ini menunjukkan bahwa analisis data menggunakan teknik Data Mining dan algoritma C4.5 dapat menjadi solusi efektif dalam deteksi dini dan pencegahan penyakit kanker paru-paru. Model ini dapat memberikan kontribusi yang signifikan dalam upaya penyelamatan nyawa dan perbaikan kualitas hidup bagi individu yang berisiko terkena penyakit ini [11]

Selain metode di atas optimasi pada algoritma Random Forest terbukti efektif di berbagai domain. Pada penelitian lainnya Model klasifikasi malware menggunakan algoritma random forest telah dibangun di mana model yang optimal didapatkan dengan hyperparameter tuning masing-masing parameter criterion berupa *entropy* nilai *max_depth* sebesar 128 nilai *max_features* berupa *log2* nilai *max_samples_spl*it sebesar 2 dan *n_estimators* sebanyak 400. *Hyperparameter* tersebut telah meningkatkan performa model terbaik dengan 99,23% pada akurasi 99,7% pada *presisi* 99,44% pada *TPR* dan 99,26% pada F1-Score [12]. Selanjutnya pada penelitian tahun 2023 yang dilakukan oleh Cinta Azzaria Erna Daniati dan Aidina Ristyawan dalam peningkatan akurasi deteksi *liver disease* melalui hyperparameter turning pada algoritma *random forest* hasil menunjukkan bahwa akurasi model meningkat dari 74% menjadi 75% dengan perbaikan pada *precision* dan *recall* khususnya untuk kelas pasien yang terdiagnosis *liver disease* [13]. Hal serupa juga ditemukan pada penelitian Pradistiani dalam optimasi *hyperparameter Random Forest* untuk memprediksi daya beli mobil menggunakan *GridSearch* di mana hasil penelitian menunjukkan bahwa model *Random Forest* dengan parameter terbaik mampu mencapai akurasi sebesar 93,33% [14].

Penerapan optimasi juga terlihat pada algoritma lain. Pada penelitian yang dilakukan oleh [15] berdasarkan penelitian dan pembahasan penerapan optimasi nilai k pada algoritma *k-nearest neighbor* menggunakan metode *gridsearchcv* dapat disimpulkan bahwa nilai k paling optimal yaitu 3 dengan perbandingan data latih 80% dan data uji 20%. Terdapat juga penelitian lain yang meningkatkan kemampuan model dalam memprediksi penyakit jantung dengan algoritma *NCL* dalam *gridsearchcv* di mana hasil pengujian model menggunakan metrik evaluasi Akurasi *Recall* dan Area *Under Curve* atau *AUC* menunjukkan peningkatan kemampuan model dengan skor *recall* meningkat dari 0.10 menjadi 0.93 dan skor *AUC* meningkat dari 0.83 menjadi 0.98 [16]. Berdasarkan latar belakang dan keberhasilan studi terdahulu penelitian ini menerapkan optimasi akurasi Random Forest menggunakan GridSearchCV dalam klasifikasi penyakit kanker paru-paru. Penelitian ini bertujuan untuk menghasilkan model klasifikasi yang akurat efisien dan cepat dalam mengklasifikasikan data risiko kanker paru-paru.

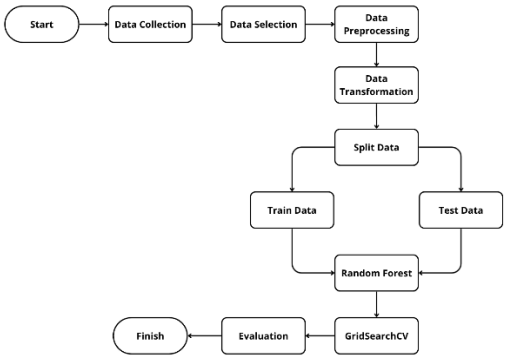
2. Metode Penelitian

Penelitian diawali dengan pengumpulan data sekunder terkait kanker paru-paru. Data kemudian melalui tahap seleksi atribut (*Selection Data*) dan pembersihan (*Preprocessing Data*) untuk menangani inkonsistensi nilai. Tahap selanjutnya adalah *Transformasi Data* (encoding dan scaling), diikuti dengan pembagian data

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No. 1 Hal: 31-38 e-ISSN: 2987-7016

uji dan latih (*Split Data*). Proses inti pemodelan menggunakan *Random Forest* sebagai *baseline*, dilanjutkan dengan optimasi parameter menggunakan *GridSearchCV*, dan diakhiri dengan evaluasi performa model. Dapat ditunjukkan pada gambar 1.



Gambar 1 Alur Penelitian Impelementasi Random Forest Menggunakan GridsearchCV Dalam Klasifikasi Penyakit Kanker Paru-Paru

2.1. Pengumpulan Data

Penelitian ini menggunakan metode pengumpulan data sekunder, di mana data diperoleh tanpa melakukan pengambilan sampel langsung di lapangan, melainkan memanfaatkan sumber data yang telah tersedia secara publik. Dataset yang digunakan dalam penelitian ini adalah "Lung Cancer Dataset" dari repositori Kaggle [https://www.kaggle.com/datasets/mikeytracegod/lung-cancer-risk-dataset/data].

2.2. Seleksi Data

Pada tahapan ini data yang relevan akan dipilih. Dimana data atau fitur yang tidak relevan akan dihapus.

2.3. Prosesing Data

Tahap selanjutnya adalah *Pre-processing* sebelum melakukan proses *data mining*, perlu dilakukan pembersihan data atau *cleaning* pada data hasil seleksi. Proses pembersihan tersebut antara lain meliputi penghapusan data ganda, pengecekan data yang tidak konsisten, dan perbaikan kesalahan pada data, seperti kesalahan ketik atau pencetakan .

2.4. Transfomasi Data

Transformasi data adalah proses penyesuaian format data agar kompatibel dengan model informasi atau jenis pola yang akan dicari dalam data mining . Penelitian ini menerapkan tiga teknik transformasi utama yaitu *Label Encoding*, *One-Hot Encoding*, dan Normalisasi (*Min-Max Scaler*).
Label encoding dan *one-hot encoding* adalah Teknik ini mengubah data bertipe kategori menjadi format numerik agar dapat diproses oleh algoritma.
Normalisasi data dilakukan menggunakan metode *Min-Max Scaler* untuk menyelaraskan skala atribut numerik ke dalam rentang seragam (0 hingga 1). Langkah ini krusial agar atribut dengan nilai besar tidak mendominasi atribut lain dalam perhitungan algoritma. Persamaan yang digunakan adalah:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \textbf{(2.1)}$$

2.5. *Random Forest*

Metode *Random Forest*, yang diperkenalkan oleh Breiman pada tahun 2001, merupakan teknik klasifikasi dan prediksi berbasis kumpulan pohon keputusan (*ensemble*), di mana proses keputusan berjalan dari akar menuju daun[17]. Keunggulan utamanya meliputi ketahanan terhadap *outliers*, kemampuan menangani data yang hilang, serta efisiensi pada *big* data melalui seleksi fitur otomatis untuk meningkatkan performa model .
Secara teknis, algoritma ini dibangun menggunakan pendekatan *bagging* dengan metode *CART (Classification and Regression Tree*), di mana setiap pohon dibiarkan tumbuh maksimal tanpa pemangkasan hingga membentuk sebuah "hutan" [18]. Dalam penentuan percabangan, algoritma mencari titik pemisah (*split*) terbaik untuk meminimalkan *impurity* menggunakan ukuran *Gini Index* atau *Entropy* agar menghasilkan klasifikasi yang optimal[19] . Dalam penelitian ini, kriteria yang digunakan adalah *Entropy* untuk mengukur *Information Gain*, dengan persamaan sebagai berikut:

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2 p_i \quad \textbf{(2.2)}$$

Dimana:

- c : Jumlah kelas target (dalam kasus ini: 2 kelas, yaitu Normal dan Kanker).
- p_i : Proporsi sampel yang termasuk dalam kelas i.

Setelah nilai Entropy dihitung, Information Gain (IG) dari atribut A diperoleh dengan rumus:

$$IG(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|s_v|}{|s|} H(s_v) \quad \textbf{(2.3)}$$

Dimana $Values(A)$ adalah kemungkinan nilai pada atribut A, dan s_v adalah subset data di mana atribut A memiliki nilai v. Model *Random Forest* yang dibangun pada tahap ini berfungsi sebagai model dasar (*baseline*) dengan parameter standar sebelum dilakukan optimasi lebih lanjut.

Meskipun *Random Forest* memiliki ketahanan *inheren* terhadap data yang tidak seimbang , efektivitasnya sangat bergantung pada pemilihan *hyperparameter* yang tepat. Penelitian sebelumnya [20] menunjukkan bahwa optimasi parameter sangat krusial untuk memastikan model tidak hanya mengandalkan pengaturan *default* yang suboptimal.

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

JUPTIK

2026 Vol. 4 No. 1 Hal: 31-38 e-ISSN: 2987-7016

2.6. GridSearchCV

Penelitian ini menerapkan *GridSearchCV* sebagai metode *hyperparameter tuning* untuk menyeleksi konfigurasi parameter yang paling optimal. Metode ini bekerja dengan mengevaluasi sejumlah kombinasi *hyperparameter* terpilih menggunakan teknik *cross-validation*, di mana kombinasi dengan performa validasi tertinggi akan dipilih sebagai parameter terbaik untuk membangun model final [12].

2.7. Hasil

Kinerja model diukur *menggunakan Confusion Matrix* untuk menghitung metrik utama seperti *Accuracy*, *Precision*, *Recall*, dan *F1-Score*

3. Hasil dan Pembahasan

3.1. Seleksi Data

Penelitian ini menggunakan "Lung Cancer Dataset" dari repositori Kaggle [https://www.kaggle.com/datasets/mikeytracegod/lung-cancer-risk-dataset/data] yang terdiri dari 50.000 data rekam medis dan 11 atribut atau kelas. Berdasarkan validasi pakar (dokter spesialis paru di RSUD Pekanbaru), diagnosis difokuskan pada risiko kanker dengan mengelompokkan fitur menjadi dua kategori klinis: faktor yang tidak bisa dimodifikasi (umur, jenis kelamin, riwayat keluarga) dan faktor yang bisa dimodifikasi (paparan rokok, asbes, gas radon, diagnosis *PPOK*, dan konsumsi alkohol). Dapat ditunjukkan pada table 1.

Tabel 1 Rincian Atribut Dataset

Nama Atribut	Tipe Data	Deskripsi Atribut	Nilai
patient_id	Integer	Identitas unik untuk setiap pasien.	100000 - 149999
age	Integer	Usia pasien (dalam tahun).	18 - 100
gender	String	Jenis kelamin pasien.	Male, Female
pack_years	Float	Tingkat paparan rokok (tahun merokok × bungkus per hari).	0 - 100
radon_exposure	String	Tingkat paparan gas radon di tempat tinggal.	Low, Medium, High
asbestos_exposure	String	Riwayat paparan kerja terhadap asbes.	Yes, No
secondhand_smoke_exposure	String	Paparan asap rokok pasif (perokok pasif).	Yes, No
copd_diagnosis	String	Diagnosis Penyakit Paru Obstruktif Kronis (PPOK).	Yes, No
alcohol_consumption	String	Pola konsumsi alkohol.	None, Moderate, Heavy
family_history	String	Riwayat keluarga dengan kanker paru-paru.	Yes, No
lung_cancer	String	(Target) Diagnosis kanker paru-paru.	Yes, No

3.2. Seleksi Data

Pada dataset penelitian ini, seleksi dilakukan terhadap atribut *patient_id.* Atribut ini merupakan identitas unik yang diberikan kepada setiap pasien secara berurutan (*indeks).* Secara statistik dan medis, *patient_id* tidak memiliki korelasi dengan diagnosis kanker paru-paru karena hanya berfungsi sebagai penanda administratif. Oleh karena itu, atribut *patient_id* dihapus dari dataset, sehingga 10 atribut utama (9 fitur prediktor dan 1 label target) yang akan digunakan untuk proses selanjutnya yang ditujunkan pada tabel 2.

Tabel 2 Setelah Seleksi Data

age	gender	pack_years	…	family_history	lung_cancer
69	Male	66,02524418	…	No	No
32	Female	12,78080002	…	Yes	Yes
89	Female	0,408278099	…	No	Yes
….	…	…	…	…	…
90	Male	14,34972227	…	Yes	Yes
33	Female	87,01255504	…	No	No
31	Male	37,59685057	…	Yes	Yes

3.3. Prosessing Data

Setelah tahap pemilihan data dilakukan dengan menghapus atribut patient_id, langkah selanjutnya adalah melakukan prosedur Pra-pemrosesan Data guna memastikan integritas seluruh baris data . Dalam penelitian ini, fokus utama pembersihan data diarahkan pada pemeriksaan catatan data ganda (*duplicate records*) serta verifikasi konsistensi format fitur numerik utama . Berdasarkan pemeriksaan lengkap terhadap 50.000 rekam medis pasien, dikonfirmasi bahwa dataset sudah bersih dan seluruh baris data tidak mengandung nilai yang hilang (*missing values*) . Atribut numerik seperti usia (age) dan pack_years juga dikonfirmasi memiliki format angka desimal yang konsisten dan bebas dari kesalahan ketik karakter non-

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No. 1 Hal: 31-38 e-ISSN: 2987-7016

numerik, sehingga siap dilanjutkan ke tahap transformasi .

3.4. Transformasi Data

Penelitian ini menerapkan teknik transformasi data meliputi *Label Encoding*, *One-Hot Encoding*, dan Normalisasi *Min-Max Scaler* . *Label Encoding* digunakan untuk mengubah variabel kategorikal biner yang memiliki dua kondisi menjadi format bilangan bulat tunggal 0 dan 1 . Aturan konversi ini diterapkan untuk atribut gender (Male dipetakan menjadi 1 dan Female menjadi 0) serta seluruh variabel biner dengan opsi jawaban *Yes* (dipetakan menjadi 1) dan *No* (dipetakan menjadi 0), sebagaimana disajikan dalam Tabel 3.

Tabel 3 Aturan Label Encoding

Kategori Data awal	Setelah Label Encoding
Yes	
Male	1
No	
Female	0

Teknik One-Hot Encoding secara khusus diterapkan pada dua atribut multinomial yang tidak memiliki tingkatan ordinal, yaitu atribut radon_exposure (kategori: *High, Medium, Low*) dan atribut alcohol_consumption (kategori: *Heavy, Moderate, None*) . Pendekatan ini memecah masing-masing kolom kategorikal tersebut menjadi tiga fitur biner baru yang bersifat independen . Hal ini sangat penting untuk mencegah algoritma *Random Forest* berasumsi bahwa terdapat urutan nilai matematika tertentu pada pola konsumsi alkohol atau paparan gas pasien, sehingga pembobotan fitur pada tiap cabang pohon keputusan berjalan objektif . Hasil transformasi biner ini masing-masing disajikan secara terperinci pada Tabel 4 dan Tabel 5 .

Tabel 4 Hasil *One-Hot Encoding* pada Atribut radon_exposure

radon_exposure_High	radon_exposure_Medium	radon_exposure_Low
1	0	0
0	1	0
0	0	1

Tabel 5 Hasil *One-Hot Encoding* pada Atribut alcohol_consumption

alcohol_consumption_none	alcohol_consumption_heavy	alcohol_consumption_moderate
1	0	0
0	1	0
0	0	1

Normalisasi (*Min-Max Scaler*) pada penelitian dilakukan berdasarkan analisis statistik pada dataset, atribut *age* memiliki rentang nilai minimum 18 dan maksimum 100. Sementara itu, atribut *pack_years* memiliki variasi nilai yang sangat lebar, mulai dari minimum 0,002752991 hingga maksimum 99,99920423. Sebagai ilustrasi, perhitungan normalisasi untuk data pasien dengan usia (age) 69 tahun adalah sebagai berikut:

$$X_{norm} = \frac{69 - 18}{100 - 18} = 0{,}62195121951 \quad \textbf{(3-1)}$$

3.5. *Random Forest*

Pada pelatihan menggunakan algoritma *random forest* dilakuka pembagian data dilakukan terhadap 50.000 data dengan tiga variasi rasio latih dan uji. Variasi tersebut meliputi rasio 90:10 (45.000 latih : 5.000 uji), rasio 80:20 (40.000 latih : 10.000 uji), dan rasio 70:30 (35.000 latih : 15.000 uji) guna menguji konsistensi performa model *Random Forest* terhadap jumlah data pelatihan.

3.6. *GridSearchCV*

GridSearchCV bekerja dengan cara menguji berbagai kombinasi parameter yang sudah ditentukan. Parameter yang digunakan pada penelitian ini meliputi *n_estimators*, *max_depth*, *min_samples_split*, dan *max_features*. Setiap kombinasi parameter diuji menggunakan teknik *cross-validation* dengan nilai cv=5. Melalui proses ini, dataset dibagi menjadi 5 bagian, di mana model dilatih menggunakan 4 bagian data dan diuji pada 1 bagian data secara bergantian. Proses ini akan dilakukan secara berulang sehingga seluruh kombinasi parameter selesai dievaluasi. Nilai evaluasi yang digunakan adalah akurasi, sehingga kombinasi parameter dengan nilai tertinggi akan dipilih sebagai parameter terbaik.

3.7. Hasil dan Pengujian

Tahap evaluasi bertujuan untuk mengukur performa model klasifikasi *Random Forest* dalam memprediksi diagnosis kanker paru-paru. Kinerja model diukur menggunakan *Confusion Matrix* untuk menghitung metrik utama seperti *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Evaluasi ini dilakukan secara komprehensif berdasarkan skenario eksperimen yang telah ditetapkan. Rincian desain eksperimen, mulai dari kombinasi *hyperparameter* yang diuji, variasi pembagian data, hingga tujuan akhir pengujian dirangkum dalam Tabel 6.

Tabel 6 Evaluatation

Aspek Evaluasi	Detail Spesifikasi
Metode Klasifikasi	Random Forest

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No. 1 Hal: 31-38 e-ISSN: 2987-7016

Normalisasi Data	Min-Max Scaler (Rentang 0 - 1)
Teknik Validasi (Split Data)	• Rasio 90 : 10 • Rasio 80 : 20 • Rasio 70 : 30
Parameter GridSearchCV	• n_estimators: 100, 200, 300 • max_depth: None, 10, 20 • min_samples_split: 2, 5 • max_features: sqrt, log2
Metrik Evaluasi	Accuracy, precision, recall, and F1 score

Berdasarkan hasil pengujian model *baseline* menggunakan parameter standar bawaan (*default*), terlihat bahwa algoritma *Random Forest* memberikan performa yang relatif stabil di seluruh skenario pembagian data dengan nilai akurasi berada pada rentang 69,07 hingga 69,44. Skenario *split* data 90:10 mencatatkan nilai akurasi tertinggi sebesar 69,44 dan nilai *F1-Score* sebesar 75,97. Meskipun model *baseline* mampu menghasilkan tingkat presisi yang cukup tinggi dan konsisten di angka 82,61% hingga 82,77, nilai *recall* pada seluruh skenario pengujian awal ini cenderung rendah dan tidak optimal, yaitu hanya berkisar antara 69,44 hingga 70,29. Hal ini menunjukkan bahwa tanpa adanya optimasi parameter, model *baseline* memiliki keterbatasan besar dalam mendeteksi kelas *positif* (pasien kanker) secara maksimal, di mana nilai *recall* terendah ditemukan pada skenario 70:30 yaitu hanya sebesar 69,44. Tingginya angka kesalahan prediksi terhadap pasien sakit yang tidak terdeteksi (*False Negative*) pada model awal ini memperkuat urgensi penerapan teknik pencarian parameter otomatis seperti *GridSearchCV* guna meningkatkan sensitivitas (*recall*) model secara signifikan dalam domain klasifikasi medis. Dapat ditunjukkan pada table 7.

Tabel 7 Baseline Results

Skenario	Metode	Akurasi	Presisi	Recall	F1-Score
Split 90:10	Baseline	69,44	82,65	70,29	75,97
Split 80:20	Baseline	69,35	82,61	70,17	75,89
Split 70:30	Baseline	69,07	82,77	69,44	75,52

Berdasarkan hasil pengujian optimasi menggunakan *GridSearchCV* pada tiga skenario pembagian data, dapat disimpulkan bahwa penggunaan metode pencarian parameter otomatis berhasil meningkatkan performa model secara signifikan dibandingkan model *baseline*. Performa terbaik dari keseluruhan eksperimen dicapai pada Skenario 90:10 lewat pengujian final parameter optimal, dengan konfigurasi parameter *max_depth*: *None*, *max_features*: *sqrt*, *min_samples_split*: 2, dan *n_estimators*: 200, yang menghasilkan nilai Akurasi sebesar 70,3 dan *Recall* mencapai 82,68. Sebagai pembanding, kandidat parameter terbaik lainnya pada skenario yang sama dengan kedalaman pohon terbatas (*max_depth*: 20 dan *n_estimators*: 200) mencatatkan Akurasi 70,25 dan *Recall* 82,54. Meskipun seluruh variasi kombinasi parameter memberikan hasil yang cukup kompetitif di kisaran angka 69 hingga 70, skenario dengan konfigurasi *max_depth*: *None* tetap dipilih sebagai hasil terbaik karena memberikan performa akumulatif paling tinggi. Hal ini sangat krusial dalam konteks diagnosa medis guna meminimalisir risiko pasien kanker yang tidak terdeteksi (*False Negative*), sehingga model akhir ini menjadi lebih sensitif dan handal dalam memprediksi risiko penyakit kanker paru-paru. Dapat ditunjukkan pada table 8.

Table 8 Gridsearch results for the 90:10 scenario

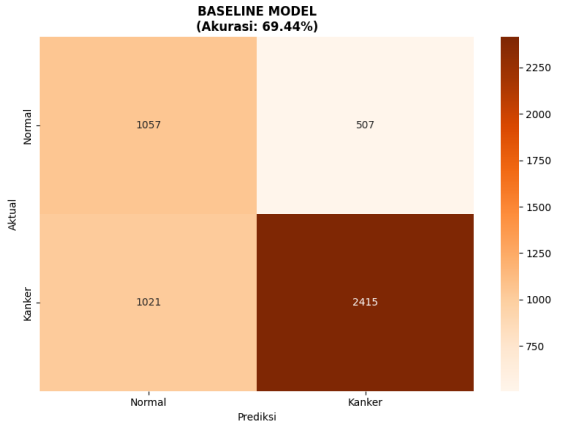
Skenario	max_depth	max_features	min_samples_split	n_estimators	Akurasi	Presisi	Recall	F1-Score
Split 90:10	None	sqrt	2	100	70,03	76,01	82,4	79,08
				200	70,23	76,05	82,74	79,26
				300	70,25	76,03	82,83	79,28
			5	100	69,96	77,37	79,56	78,45
				200	70,05	77,38	79,73	78,54
				300	70,02	77,29	79,84	78,55
		log2	2	100	70,03	76,01	82,4	79,08
				200	70,23	76,05	82,74	79,26
				300	70,25	76,03	82,83	79,28
			5	100	69,96	77,37	79,56	78,45
				200	70,05	77,38	79,73	78,54
				300	70,02	77,29	79,84	78,55
	10	sqrt	2	100	69,34	82,03	70,92	76,07
				200	69,37	82,03	70,98	76,11
				300	69,45	82,09	71,04	76,17
			5	100	69,48	82,07	71,13	76,21
				200	69,48	82,07	71,13	76,21
				300	69,48	82,07	71,13	76,21

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

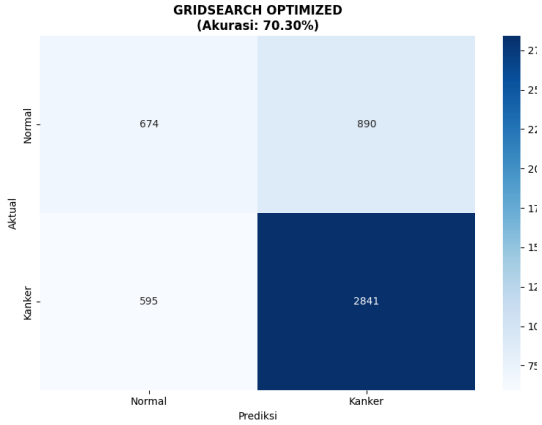
2026 Vol. 4 No. 1 Hal: 31-38 e-ISSN: 2987-7016

				200	69,5	82,12	71,09	76,21
				300	69,42	82,08	71	76,14
				100	69,34	82,03	70,92	76,07
			2	200	69,37	82,03	70,98	76,11
				300	69,45	82,09	71,04	76,17
		log2		100	69,48	82,07	71,03	76,2
			5	200	69,5	82,12	71,09	76,21
				300	69,42	82,08	71	76,14
				100	70,13	76,12	82,38	79,13
			2	200	70,25	76,17	82,54	79,22
				300	70,21	76,11	82,57	79,2
		sqrt		100	70,05	77,46	79,58	78,51
			5	200	70,05	77,46	79,58	78,51
				300	70,12	77,47	79,7	78,57
	20			100	70,13	76,12	82,38	79,13
			2	200	70,25	76,17	82,54	79,22
				300	70,21	76,11	82,57	79,2
		log2		100	70,05	77,46	79,58	78,51
			5	200	70,05	77,46	79,58	78,51
				300	70,12	77,47	79,7	78,57
			2	200	70,3	76,15	82,68	79,28
	None	sqrt						

Perbandingan antara model *baseline* pada figure 2 dan hasil optimasi *GridSearchCV* pada figure 3 menunjukkan peningkatan performa yang signifikan, terutama pada aspek sensitivitas model. Pada skenario terbaik (90:10), model *baseline* hanya mampu mencatatkan akurasi sebesar 69,44 dengan nilai *recall* 70,29. Namun, setelah dilakukan optimasi *hyperparameter*, akurasi model meningkat menjadi 70,3 dan nilai *recall* melonjak drastis hingga mencapai 82,68. Kenaikan *recall* sebesar 12,39 ini menjadi poin krusial dalam klasifikasi penyakit kanker paru-paru, karena menunjukkan bahwa model yang telah dioptimasi jauh lebih handal dalam mendeteksi pasien yang benar-benar sakit dan berhasil meminimalisir risiko kesalahan diagnosis (*False Negative*) yang sangat berbahaya dalam dunia medis. Dapat di tunjukkan di gambar 2 dan 3.



Gambar 2 Confusion Matrix Baseline



Gambar 3 Confusion Matrix GridsearchCV

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, model klasifikasi risiko kanker paru-paru menggunakan algoritma *Random Forest* yang dioptimasi dengan *GridSearchCV* berhasil meningkatkan performa klasifikasi dibandingkan model tanpa optimasi. Penelitian ini menggunakan dataset sebanyak 50.000 data pasien dengan tahapan pra-pemrosesan berupa verifikasi integritas data, transformasi data menggunakan kombinasi Label dan One-Hot

Encoding, serta normalisasi Min-Max Scaler untuk meningkatkan kualitas data sebelum proses pelatihan model. Hasil pengujian menunjukkan bahwa skenario pembagian data 90:10 memberikan performa terbaik dibandingkan skenario lainnya. Model *Random Forest* hasil optimasi *GridSearchCV* mampu mencapai akurasi sebesar 70,3%, Recall sebesar 82,68%, dan F1-Score sebesar 79,28%. Peningkatan nilai *Recall* menunjukkan bahwa model lebih efektif dalam mendeteksi kasus *positif* kanker paru-paru dan mampu mengurangi kesalahan *False Negative*, sehingga berpotensi mendukung proses deteksi dini dan pengambilan keputusan medis secara lebih baik. Meskipun model yang dihasilkan menunjukkan performa yang cukup baik, penelitian ini masih memiliki keterbatasan karena hanya menggunakan satu sumber dataset publik. Oleh karena itu, penelitian selanjutnya disarankan untuk menggunakan dataset klinis yang lebih beragam dari berbagai institusi kesehatan serta mengeksplorasi metode yang lebih kompleks seperti Deep Learning dan *Explainable Artificial Intelligence* (XAI) untuk meningkatkan kemampuan generalisasi dan interpretasi model.

Daftar Rujukan

[1] D. Aprilianto and E. Rizal, "Klasifikasi Penyakit Kanker Paru Menggunakan Algoritma Random Forest Berbasis Streamlit," vol. 9, p. 2025, 2025, doi: 10.47002/metik.v9i2.1076.

[2] J. P. Kesehatan *et al.*, "Kanker didefinisikan sebagai suatu penyakit yang berasal dari adanyapertumbuhan sel tubuh yang progresif dan abnormal . Kondisi ini disebabkanolehterjadinya perubahan pada deoxiribonucleid acid (DNA), sehingga sel kehilangan fungsinya secara normal . Pe," vol. 3, no. 2, pp. 141–149, 2020.

[3] A. Maulana, A. Pratama, D. Primanda, and N. Hariyanto, "Model Prediksi Kanker Paru-Paru dengan Random Forest Lung Cancer Prediction Model with Random Forest," vol. 15, no. 2, pp. 136–146, 2025. https://doi.org/10.30700/sisfotenika.v15i2.569

[4] I. Buana and D. A. Harahap, "Asbestos, Radon and Air Pollution as Risk Factors for Lung Cancer in Non-Smoking Women," *AVERROUS: Jurnal Kedokteran dan Kesehatan Malikussaleh*, vol. 8, no. 1, p. 1, 2022. https://doi.org/10.29103/averrous.v8i1.7088

[5] B. Bhinder, C. Gilvary, N. S. Madhukar, and O. Elemento, "Artificial Intelligence in Cancer Research and Precision Medicine," *Cancer Discovery*, vol. 11, no. 4, pp. 900–915, Apr. 2021, https://doi.org/10.1158/2159-8290.CD-21-0090

[6] H. Shimizu and K. I. Nakayama, "Artificial intelligence in oncology," *Cancer Science*, vol. 111, no. 5, pp. 1452–1460, 2020, https://doi.org/10.1111/cas.14377

[7] N. Nuraeni and P. Astuti, "Pendekatan Machine Learning untuk Deteksi Dini Kanker Paru-Paru: Mengoptimalkan Sensitivitas dan Akurasi," *Jurnal Informatika Polinema*, vol. 11, no. 3, pp. 339–346, 2025, https://doi.org/10.33795/jip.v11i3.7011.

[8] K. Aidi Saputra, J. Tata Hardinata, M. Ridwan Lubis, S. Retno Andani, and I. Syahputra Saragih, "KLIK: Kajian Ilmiah Informatika dan Komputer Klasifikasi Algoritma C4.5 Dalam Penerapan Tingkat Kepuasan Siswa Terhadap Media Pembelajaran Online," *Media Online)*, vol. 1, no. 3, pp. 113–118, 2020.

[9] M. Y. Zidane, B. Nurina Sari, I. Maulana, A. Primaya, and G. Garno, "Penerapan Data Mining Dalam Klasifikasi Data Transaksi Produk Koperasi Di Smk Pgri 2 Karawang," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 1, pp. 263–269, 2024, https://doi.org/10.36040/jati.v9i1.12196

[10] M. Abdul, R. Wahid, A. Nugroho, and A. H. Anshor, "Prediksi Penyakit Kanker Paru-Paru Dengan Algoritma Regresi Linier," *Bulletin of Information Technology (BIT)*, vol. 4, no. 1, pp. 63–74, 2023, https://doi.org/10.47065/bit.v3i1

[11] K. P. Menggunakan, C. Algoritma, J. Teknologi, S. Informasi, and P. N. Subang, "Penerapan Metode Klasifikasi Decision Tree dalam Prediksi," vol. 18, no. 1, pp. 126–139.

[12] I. Muhamad and M. Matin, "Hyperparameter Tuning menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware," vol. 9, no. 1, 2023. https://doi.org/10.32722/multinetics.v9i1.5578

[13] C. Azzaria, E. Daniati, and A. Ristyawan, "Peningkatan Akurasi Deteksi Liver Disease melalui Hyperparameter Tuning pada Algoritma Random Forest," vol. 4, no. 2, pp. 139–147, 2025. https://doi.org/10.59095/ijcsr.v4i2.198

[14] R. Pradistiani, "Jurnal JPILKOM (Jurnal Penelitian Ilmu Komputer) Optimasi Hyperparameter Random Forest dalam Memprediksi Daya Beli Mobil Menggunakan GridSearch," vol. 3, no. 1, 2025.

[15] K. Kunci, "Indonesian Journal of Computer Science," vol. 12, no. 1, pp. 2162–2171, 2023.

[16] Z. Ahmadi, A. Abdullah, and I. Fakhruzi, "Meningkatkan Kemampuan Model dalam Memprediksi Penyakit Jantung Meningkatkan Kemampuan Model dalam Memprediksi Penyakit Jantung dengan Algoritma NCL dan GridSearchCV," no. October, 2023, https://doi.org/10.30865/mib.v7i3.6142

[17] Suci Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, vol. 4, no. 3, pp. 121–127, 2022, https://doi.org/10.35580/variansiunm31

[18] S. Mahmuda, "Implementasi Metode Random Forest pada Kategori Konten Kanal Youtube," *Jurnal Jendela Matematika*, vol. 2, no. 01, pp. 21–31, 2024, https://doi.org/10.57008/jjm.v2i01.633

[19] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis : Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 13, no. 2, pp. 67–75, 2020, https://doi.org/10.51903/e-bisnis.v13i2.247

[20] M. C. E. all Rani, "Perbandingan Algoritma Random Forest, Naive Bayes, Dan Neural Network Dalam Klasifikasi Penyakit Jantung," *Jurnal Sains Informatika Terapan (JSIT)*, vol. 2, no. 1, pp. 16–20, 2023.