Jurnal Kreativitas Mahasiswa Informatika **E-ISSN: 3164-1989**
Penerbit: Fakultas Ilmu Komputer Universitas Pamulang **P-ISSN: 2797-6327**
Vol. 7, No. 2, Agustus 2026 (389-402) **https://doi.org/10.32493/jatimika.v7i1**

Implementasi Metode *Decision Tree* Algoritma C4.5 Untuk Klasifikasi Penyakit Paru-Paru Pada Manusia

Rama Albin Sugiartha[1*], Yan Mitha Djaksana[2]

[1,2] Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang, Tangerang Selatan, Banten, Indonesia

e-mail: [1]ramaalbin.06@gmail.com, [2] dosen01994@unpam.ac.id

Submitted Date: March 03, 2026 Reviewed Date: March 30, 2026
Revised Date: April 06, 2026 Accepted Date: April 16, 2026

Abstract

Lung disease remains a serious health concern in Indonesia. The diagnostic process typically relies on manual examination, which requires considerable time and expertise, while patient data continues to grow in size and complexity. This situation highlights the need for data mining techniques to support early detection and data-driven medical decision-making. This study aims to classify the risk of lung disease in humans using the Decision Tree C4.5 algorithm within the Knowledge Discovery in Database (KDD) framework. The dataset used is secondary data from Kaggle, comprising approximately 30,000 patient records and 10 attributes related to demographics, lifestyle, and health conditions. All KDD stages data selection, preprocessing, transformation, modeling, and evaluation were implemented using RapidMiner Studio. The results of testing with 10-Fold Cross Validation show that the C4.5 algorithm delivers excellent classification performance, achieving an accuracy of 94.65% along with high precision, recall, and F1-score values. The resulting decision tree identifies smoking habits, physical activity, age, and sleep patterns as the main risk factors for lung disease. The model is interpretable and can be translated into "if–then" rules, making it potentially useful as a component of decision support systems for early screening and public health education on lung disease risks.

Keywords: *Lung disease classification, Decision Tree, C4.5 Algorithm, Data mining, RapidMiner.*

Abstrak

Penyakit paru-paru masih menjadi masalah kesehatan serius di Indonesia. Proses diagnosis umumnya bergantung pada pemeriksaan manual yang memerlukan waktu dan keahlian tinggi, sementara data pasien semakin besar dan kompleks. Kondisi ini mendorong pemanfaatan teknik data mining untuk membantu deteksi dini dan pengambilan keputusan medis berbasis data. Penelitian ini bertujuan mengklasifikasikan risiko penyakit paru-paru pada manusia menggunakan algoritma *Decision Tree* C4.5 dengan pendekatan *Knowledge Discovery in Database* (KDD). *Dataset* yang digunakan merupakan data sekunder dari *Kaggle* dengan sekitar 30.000 data pasien dan 10 atribut terkait demografi, gaya hidup, dan kondisi kesehatan. Seluruh tahapan KDD, mulai dari seleksi data, *preprocessing*, transformasi, hingga pemodelan dan evaluasi, diimplementasikan menggunakan *RapidMiner Studio*. Hasil pengujian dengan 10-*Fold Cross Validation menunjukkan* bahwa algoritma C4.5 mampu memberikan performa klasifikasi yang sangat baik dengan akurasi sebesar 94,65% disertai nilai *Precision, Recall,* dan *F1-score* yang tinggi. Pohon keputusan yang dihasilkan menempatkan kebiasaan merokok, aktivitas olahraga, usia, dan pola begadang sebagai faktor utama penentu risiko penyakit paru-paru. Model ini bersifat *interpretable* dan mudah diterjemahkan ke dalam aturan "*if–then*", sehingga berpotensi digunakan sebagai komponen sistem pendukung keputusan untuk skrining awal dan edukasi risiko penyakit paru-paru.

Kata kunci: *Klasifikasi penyakit paru-paru, Decision Tree, Algoritma C4.5, Data mining, RapidMiner.*

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

E-ISSN: 3164-1989
P-ISSN: 2797-6327
https://doi.org/10.32493/jatimika.v7i1

1. Pendahuluan

Perkembangan teknologi informasi dan kecerdasan buatan (Artificial Intelligence / AI) telah memberikan kontribusi besar dalam berbagai bidang, terutama dalam dunia medis. Salah satu cabang dari AI yang paling banyak dimanfaatkan adalah machine learning, yang memiliki kemampuan untuk mengolah dan menganalisis data medis dalam jumlah besar guna menemukan pola, tren, serta hubungan antarvariabel yang sulit diidentifikasi secara manual. Dengan meningkatnya kompleksitas dan volume data medis dari berbagai sumber seperti hasil laboratorium, radiologi, serta catatan rekam medis elektronik (Electronic Health Record/EHR), kebutuhan akan sistem berbasis kecerdasan buatan (AI) menjadi semakin mendesak. Sistem seperti ini tidak hanya membantu tenaga medis mempercepat diagnosis, tetapi juga dapat mengurangi kesalahan manusia (human error) dalam proses analisis data yang berulang. Dalam konteks pelayanan kesehatan modern, AI memungkinkan penerapan analisis prediktif yang mampu memperkirakan risiko penyakit sebelum gejalanya muncul, sehingga tenaga medis dapat mengambil langkah pencegahan lebih awal. Pemanfaatan teknologi ini juga mendukung konsep precision medicine, yaitu pendekatan medis yang menyesuaikan diagnosis dan pengobatan berdasarkan karakteristik genetik, lingkungan, dan gaya hidup individu. Dalam konteks penyakit paru-paru, pendekatan ini memungkinkan analisis mendalam terhadap faktor risiko spesifik seperti lama paparan asap rokok, riwayat keluarga, dan intensitas aktivitas fisik. Dengan bantuan algoritma klasifikasi seperti C4.5, data tersebut dapat dikonversi menjadi informasi prediktif yang membantu tenaga medis menentukan intervensi paling tepat. Pendekatan berbasis data semacam ini kini menjadi fondasi transformasi digital bidang kesehatan di Indonesia, sejalan dengan inisiatif Smart Hospital dan Health Informatics Integration yang dicanangkan Kemenkes.. Dalam konteks penyakit paru-paru salah satu penyebab kematian tertinggi di duniakehadiran sistem pendukung keputusan berbasis AI dapat meningkatkan ketepatan diagnosis dan mempercepat penanganan pasien, terutama di wilayah dengan keterbatasan tenaga medis.

Fenomena di lapangan menunjukkan bahwa penyakit saluran pernapasan dan penyakit musiman masih sering mengalami lonjakan kasus pada waktu-waktu tertentu, misalnya saat pancaroba atau ketika terjadi cuaca ekstrem. Dalam laporan Isu Sepekan Bidang Kesra Komisi IX DPR RI (Minggu ke-1 November 2025) disebutkan bahwa beberapa daerah melaporkan peningkatan kasus penyakit dalam sepekan terakhir, termasuk infeksi saluran pernapasan atas (ISPA), influenza like illness (ILI), dan influenza. Bahkan, di Bengkulu Tengah jumlah kasus ISPA dilaporkan meningkat hingga dua kali lipat dalam tujuh hari, dan di Karawang ISPA menjadi penyakit dengan jumlah kasus tertinggi selama periode pancaroba(Rini et al., 2025).

Penyakit paru-paru merupakan salah satu penyakit yang paling banyak menyebabkan kematian di dunia. Berdasarkan data World Health Organization (WHO), penyakit paru seperti tuberkulosis (TBC), pneumonia, kanker paru, dan penyakit paru obstruktif kronik (PPOK) memiliki tingkat prevalensi dan risiko kematian yang tinggi. Di Indonesia, penyakit paru juga menjadi salah satu masalah kesehatan utama karena meningkatnya polusi udara, kebiasaan merokok, dan gaya hidup yang tidak sehat. Rofiani et al., (2024) dalam penelitiannya menyebutkan bahwa kanker paru menempati posisi keempat tertinggi di dunia, dengan mayoritas penderitanya adalah pria berusia di atas 40 tahun yang memiliki kebiasaan merokok Oleh karena itu, diperlukan sistem yang dapat mendeteksi risiko penyakit paru-paru lebih awal dengan bantuan teknologi berbasis data.

Salah satu algoritma yang terbukti efektif dalam bidang klasifikasi medis adalah Decision Tree (Pohon Keputusan), khususnya algoritma C4.5. Algoritma ini bekerja dengan menghitung nilai entropy dan information gain untuk menentukan atribut terbaik dalam membentuk struktur pohon keputusan. Atribut-atribut seperti usia, jenis kelamin, kebiasaan merokok, tingkat aktivitas, riwayat penyakit, dan hasil pemeriksaan medis dapat digunakan untuk membangun model klasifikasi penyakit paru-paru. Penelitian oleh Rofiani et al., (2024) menunjukkan bahwa atribut kebiasaan merokok dan usia merupakan dua faktor paling berpengaruh terhadap klasifikasi penyakit paru, dengan akurasi model mencapai lebih dari 85% menggunakan RapidMiner sebagai alat bantu analisis.

Selain itu, Hendry & Djaksana, (2024) menjelaskan bahwa algoritma C4.5 mampu menghasilkan model klasifikasi dengan struktur pohon keputusan yang mudah dipahami oleh pengguna non-teknis, dengan akurasi mencapai 73,42% dalam studi yang mereka lakukan. Sementara penelitian oleh Sajili & Djaksana, (2024) menunjukkan keberhasilan metode machine learning dalam diagnosis penyakit saraf otak menggunakan algoritma Modified K-Nearest Neighbor (M-KNN), yang memberikan hasil klasifikasi

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

konsisten dengan perhitungan manual. Kedua penelitian tersebut memperkuat bahwa penerapan metode pembelajaran mesin dalam bidang kesehatan dapat membantu diagnosis penyakit secara otomatis dan akurat.algoritma C4.5 mampu menangani data numerik dan kategorikal sekaligus, serta menghasilkan rule-based model yang transparan dan dapat dijelaskan secara logis kepada tenaga medis. Hal ini menjadikan algoritma C4.5 tidak hanya unggul dalam performa komputasi, tetapi juga relevan secara klinis dalam konteks sistem pendukung keputusan medis (Decision Support System).

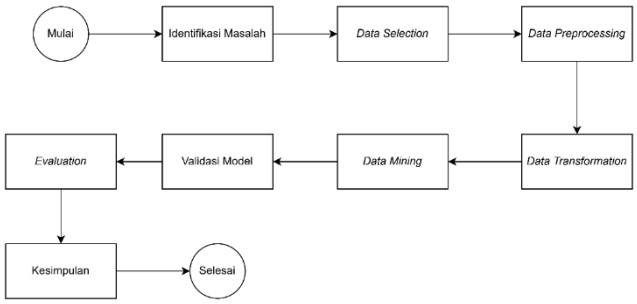
Dengan memanfaatkan algoritma Decision Tree C4.5, penelitian ini bertujuan untuk membangun model klasifikasi penyakit paru-paru berdasarkan data pasien. Model ini diharapkan mampu membantu proses diagnosis dini dan meningkatkan akurasi pengambilan keputusan medis. Implementasi ini menjadi langkah strategis menuju transformasi digital di bidang kesehatan, yang memadukan pengetahuan medis dan teknologi informasi untuk pelayanan yang lebih cepat, tepat, dan efisien.

Meskipun demikian, masih terdapat kesenjangan (gap) yang menjadi dasar penelitian ini. Pertama, dibutuhkan proses pengolahan data yang sistematis agar hasil klasifikasi benar-benar merepresentasikan data (misalnya melalui tahapan KDD: selection, preprocessing, transformation, data mining). Kedua, diperlukan penerapan algoritma klasifikasi yang tidak hanya memberikan performa baik, tetapi juga interpretable agar hasilnya dapat dijelaskan. Ketiga, perlu diketahui atribut mana yang paling dominan dalam memengaruhi klasifikasi risiko penyakit paru-paru. Keempat, evaluasi model perlu dilakukan menggunakan metode validasi yang lebih kuat seperti 10-Fold Cross Validation serta metrik evaluasi yang lengkap agar performa model lebih dapat dipercaya, tidak hanya bergantung pada satu pembagian data.

Berdasarkan fenomena dan kesenjangan tersebut, penelitian ini menerapkan algoritma Decision Tree C4.5 untuk membangun model klasifikasi risiko penyakit paru-paru berdasarkan data pasien. Model yang dihasilkan diharapkan dapat membantu proses identifikasi risiko secara dini, meningkatkan ketepatan pengambilan keputusan, serta mendukung transformasi digital di bidang kesehatan melalui pendekatan yang lebih cepat, tepat, dan efisien. Dengan demikian, latar belakang ini mengarah pada identifikasi masalah penelitian yang akan dijabarkan pada bagian berikutnya.

"Penulis menyatakan dengan sebenar-benarnya bahwa artikel ilmiah ini adalah karya asli penulis. Tidak terdapat karya atau pendapat yang ditulis atau dipublikasikan orang lain, kecuali secara tertulis diakui sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan pada daftar pustaka."

2. Metodologi Penelitian



Gambar 1 Diagram Alir Penelitian

Diagram alir penelitian menggambarkan urutan kegiatan dari tahap awal hingga akhir penelitian. Alur ini dimulai dari identifikasi masalah untuk menentukan arah penelitian, dilanjutkan dengan studi literatur yang memperkuat landasan teori. Setelah itu dilakukan tahap data selection untuk memilih data relevan, *preprocessing* untuk pembersihan data, *transformation* untuk menyesuaikan format data, hingga tahap data *mining* menggunakan algoritma *Decision Tree* C4.5. Tahap akhir berupa evaluation digunakan untuk menilai performa model berdasarkan metrik akurasi, presisi, dan recall.

Pelaksanaan kegiatan ini menggunakan pendekatan Knowledge *Discovery in Database* (KDD) sebagai kerangka kerja utama, karena KDD memfasilitasi pengolahan data mentah menjadi pengetahuan yang bermakna melalui tahapan yang sistematis, mulai dari selection, preprocessing, transformation, data mining, hingga evaluation. Proses pelaksanaan disusun mengikuti alur penelitian yang diawali dari identifikasi masalah, dilanjutkan studi literatur,

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

pengolahan data, penerapan algoritma, hingga evaluasi performa model menggunakan metrik akurasi, presisi, dan *Recall*.

a. Alur Pelaksanaan Kegiatan

Alur pelaksanaan menggambarkan urutan kegiatan dari tahap awal hingga akhir. Tahapan dimulai dari identifikasi masalah untuk memastikan arah kegiatan, dilanjutkan studi literatur sebagai penguatan landasan teoritis. Setelah itu dilakukan pemilihan data yang relevan, pembersihan data, transformasi data, penerapan algoritma *Decision Tree* C4.5, dan diakhiri dengan evaluasi untuk menilai performa model klasifikasi.

b. Sumber Data

Data yang digunakan merupakan data sekunder berupa "*Dataset* Predic Terkena Penyakit Paru-Paru" yang diperoleh dari platform *Kaggle* . *Dataset* memuat atribut-atribut kategorikal yang menggambarkan kondisi pasien, termasuk informasi seperti usia, kebiasaan merokok, aktivitas olahraga, serta riwayat penyakit bawaan.

c. Tahapan Pelaksanaan Berdasarkan KDD

1. Identifikasi Masalah

Penelitian ini menggunakan pendekatan Knowledge Discovery in Database (KDD) sebagai metode utama karena mampu mengubah data mentah menjadi pengetahuan yang bermakna. KDD mencakup seluruh tahapan mulai dari pemilihan data (selection), pembersihan (preprocessing), transformasi, penerapan algoritma (data mining), hingga evaluasi hasil (evaluation). (Fayyad et al., 1996; Carudin et al., 2024). Dengan pendekatan ini, penelitian dapat dilakukan secara terstruktur, efisien, dan memiliki landasan analitis yang kuat. Pendekatan ini juga memungkinkan penerapan algoritma Decision Tree C4.5 secara sistematis untuk mengidentifikasi pola pada data kesehatan pasien.

Proses identifikasi masalah dilakukan untuk memastikan bahwa penelitian ini memiliki arah yang jelas dan relevan terhadap kebutuhan nyata di bidang medis. Berdasarkan hasil pengamatan terhadap penelitian terdahulu dan fenomena lapangan, ditemukan beberapa permasalahan utama sebagai berikut:

1. Proses diagnosis penyakit paru-paru masih banyak bergantung pada metode konvensional yang memerlukan waktu lama dan tenaga ahli medis.
2. Belum tersedianya sistem berbasis data mining yang mampu melakukan klasifikasi penyakit paru-paru secara otomatis dan akurat.
3. Banyaknya atribut pasien (usia, kebiasaan merokok, aktivitas fisik, dan penyakit bawaan) yang belum dianalisis secara sistematis menggunakan metode klasifikasi cerdas.
4. Diperlukan penerapan algoritma machine learning yang transparan dan mudah diinterpretasikan seperti Decision Tree C4.5, agar hasil klasifikasi dapat dipahami oleh tenaga medis.

2. Data Selection

Tahap ini merupakan proses pemilihan data yang relevan dari sumber data utama. Dataset yang digunakan dalam penelitian ini adalah "Dataset Predic Terkena Penyakit Paru-Paru" yang diperoleh dari platform Kaggle. Dataset ini berisi atribut-atribut kategorikal yang menggambarkan kondisi pasien, seperti usia, kebiasaan merokok, aktivitas olahraga, hingga riwayat penyakit bawaan.

Pemilihan data dilakukan dengan memperhatikan kesesuaian terhadap tujuan penelitian, yaitu mengklasifikasikan apakah seseorang berisiko menderita penyakit paru-paru atau tidak. Data yang dianggap tidak relevan dengan tujuan penelitian tidak disertakan dalam proses analisis. Dataset yang digunakan memiliki 10 atribut utama, dengan atribut "Hasil" sebagai label atau target class yang akan diprediksi. Tujuan utama tahap ini adalah memastikan bahwa hanya data yang relevan dan berkualitas tinggi yang digunakan untuk membangun model klasifikasi.

3. Data *Preprocessing*

Tahapan ini bertujuan untuk meningkatkan kualitas data sebelum digunakan dalam proses klasifikasi. Data yang tidak bersih atau tidak konsisten dapat memengaruhi akurasi model yang dihasilkan. Proses data preprocessing dilakukan dengan langkah-langkah berikut:

1. Pemeriksaan nilai kosong (missing values)

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

Memastikan tidak ada atribut yang memilik nilai kosong atau tidak terisi. Jika ditemukan nilai kosong, dilakukan penghapusan atau pengisian dengan nilai dominan (mode).

2. Pemeriksaan data duplikat
   Pemeriksaan dan penanganan data duplikat merupakan bagian penting dalam tahap data preprocessing. Data duplikat adalah data yang tercatat lebih dari satu kali, baik dengan nilai yang sama pada seluruh kolom maupun pada kolom tertentu. Jika dibiarkan, data duplikat dapat memengaruhi proses pelatihan model, terutama pada klasifikasi, karena data yang sama bisa "terhitung" lebih banyak dari semestinya. Akibatnya, model berpotensi menjadi bias dan hasil evaluasi kinerjanya kurang menggambarkan kondisi data yang sebenarnya.
3. Standarisasi format data

   Semua atribut diubah ke format teks yang seragam (misalnya: "YA" menjadi "Ya").
4. Penyaringan Data (Filtering)
   Menghapus data yang tidak relevan, seperti data yang memiliki nilai ekstrem atau tidak sesuai konteks medis.

4. Data Transformation

Transformasi data dilakukan untuk menyiapkan dataset agar sesuai dengan kebutuhan proses pemodelan. Pada penelitian ini, transformasi difokuskan pada penetapan peran atribut (*role*) serta memastikan atribut target berada pada format kelas yang sesuai untuk klasifikasi. Atribut "Hasil" ditetapkan sebagai label karena merupakan variabel yang akan diprediksi (apakah pasien berisiko penyakit paru-paru atau tidak). Nilai pada label tetap dipertahankan dalam bentuk kategorikal "Ya" dan "Tidak" (binomial), karena algoritma *Decision Tree C4.5* pada *RapidMiner* dapat memproses label kategorikal secara langsung tanpa perlu mengubahnya menjadi bentuk numerik seperti 1/0. Dengan demikian, tahap transformasi pada penelitian ini berfokus pada penentuan label dan kesiapan struktur data sebelum masuk ke tahap data mining.

5. Data Mining (Pembangunan Model)

Tahap ini merupakan inti dari penelitian, yaitu penerapan algoritma Decision Tree C4.5 untuk membangun model klasifikasi.

Algoritma C4.5 bekerja dengan menghitung Entropy dan Information Gain untuk memilih atribut terbaik sebagai simpul utama (root node) pohon keputusan.

$$Entropy\ (S) = \sum_{l=1}^{n} - \rho i\ log_2\ (\rho i)\ \ldots\ldots\ldots$$

$$Gain\ (S, A) =\ Entropy\ (S) - \sum_{i=1}^{n} \frac{Si}{S}\ X\ Entropy\ (S_{i)}$$

Atribut dengan nilai gain tertinggi akan menjadi pembagi utama dalam pohon keputusan. Setelah proses ini, RapidMiner menghasilkan struktur pohon (Decision Tree model) yang menunjukkan hubungan antara atribut pasien dan hasil diagnosis penyakit paru-paru.

6. Validasi Model

Tahap ini memastikan bahwa model klasifikasi yang telah dibangun (Decision Tree C4.5) mampu melakukan generalisasi pada data yang belum pernah dilihat, sehingga kinerja yang dilaporkan bukan hasil kebetulan atau overfitting.

Metode validasi. Penelitian ini menggunakan 10-Fold Cross Validation. Seluruh data dibagi menjadi 10 bagian berukuran relatif sama (fold). Pada setiap iterasi, 9 fold digunakan sebagai data latih dan 1 fold sebagai data uji. Proses diulang hingga setiap fold menjadi data uji tepat satu kali, kemudian metrik dari 10 iterasi dihitung rata-ratanya sebagai performa akhir model.

Validasi model pada penelitian ini menggunakan metode k-fold cross validation, yaitu teknik evaluasi yang membagi data menjadi k bagian (fold). Pada setiap iterasi, sebanyak k-1 fold digunakan sebagai data latih (training set) dan 1 fold digunakan sebagai data uji (testing set). Proses ini diulang hingga setiap fold pernah menjadi data uji satu kali, kemudian nilai performa dirata-ratakan untuk memperoleh estimasi kinerja model

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

E-ISSN: 3164-1989
P-ISSN: 2797-6327
https://doi.org/10.32493/jatimika.v7i1

yang lebih stabil. Metode ini digunakan untuk mengurangi bias akibat pembagian data latih-uji yang hanya dilakukan sekali, sehingga hasil evaluasi tidak bergantung pada satu skenario pembagian data. Pada penelitian ini dipilih 10-fold cross validation karena umum digunakan pada penelitian klasifikasi serta dinilai mampu memberikan estimasi performa yang cukup stabil pada ukuran data yang besar.

Pada penelitian ini, proses klasifikasi dilakukan menggunakan algoritma Decision Tree C4.5 melalui operator Decision Tree pada RapidMiner. Parameter pemodelan digunakan untuk mengendalikan kompleksitas pohon keputusan agar tidak terlalu kompleks dan mengurangi risiko overfitting. Parameter yang diperhatikan meliputi kriteria pemilihan atribut pada setiap percabangan (criterion), pembatasan kompleksitas seperti maximal depth dan minimal gain, serta mekanisme pruning untuk menyederhanakan cabang yang tidak memberikan peningkatan kinerja yang signifikan.

Dalam penelitian ini, konfigurasi parameter mengikuti pengaturan bawaan (default) RapidMiner untuk menjaga konsistensi eksperimen dan menghindari bias pemilihan parameter secara subjektif. Selain itu, evaluasi model dilakukan menggunakan 10-Fold Cross Validation, sehingga performa yang diperoleh mencerminkan kemampuan generalisasi model pada data uji yang berbeda-beda.

7. Evaluasi

Evaluasi kinerja model klasifikasi pada penelitian ini dilakukan untuk mengetahui seberapa baik model Decision Tree C4.5 dalam memprediksi kelas target "Hasil" (Ya/Tidak). Evaluasi dilakukan menggunakan Confusion Matrix serta metrik evaluasi yang terdiri dari Accuracy, Precision, Recall, dan F1-Score. Nilai metrik tersebut dihitung berdasarkan

komponen True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) yang diperoleh dari confusion matrix hasil pengujian.

D. Perangkat dan Implementasi

Implementasi seluruh tahapan dilakukan menggunakan RapidMiner Studio, yang dimanfaatkan untuk data preprocessing, pembentukan model Decision Tree, serta evaluasi performa model melalui alur kerja berbasis operator.

3. Hasil dan Pembahasan

3.1. Gambaran Umum Dataset

Dataset yang digunakan merupakan data sekunder dari platform Kaggle yang disusun oleh Akbar Muharram (2024) dengan judul "Dataset Predic Terkena Penyakit Paru-Paru". Dataset ini berisi informasi yang menggambarkan berbagai aspek gaya hidup, kondisi sosial, dan kesehatan pasien yang relevan dengan risiko penyakit paru-paru.

Dataset terdiri dari 10 atribut kategorikal, termasuk satu atribut target yaitu "Hasil" yang menandakan kondisi pasien (berisiko atau tidak berisiko). Berikut adalah rincian atribut:

Tabel 1. Atribut Kategorikal

No	Nama Atribut	Tipe Data	Keterangan
1	Usia	Kategorikal	Menunjukkan kelompok usia pasien, yaitu "Muda" atau "Tua".
2	Jenis Kelamin	Kategorikal	Menunjukkan jenis kelamin pasien ("Pria" atau "Wanita").
3	Merokok	Kategorikal	Menggambarkan kebiasaan merokok pasien, yaitu "Aktif" atau "Pasif".

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

4	Bekerja	Kategorikal	Status pekerjaan pasien, yaitu "Ya" (bekerja) atau "Tidak".
5	Berumah Tangga	Kategorikal	Menunjukkan apakah pasien memiliki tanggungan keluarga ("Ya" atau "Tidak").
6	Aktivitas Begadang	Kategorikal	Menunjukkan kebiasaan pasien dalam begadang, yaitu "Ya" (sering begadang) atau "Tidak".
7	Aktivitas Olahraga	Kategorikal	Menggambarkan tingkat aktivitas fisik pasien ("Sering" atau "Jarang").
8	Asuransi	kategorikal	Menunjukkan kepemilikan asuransi kesehatan pasien ("Ada" atau "Tidak").
9	Penyakit Bawaan	kategorikal	Menunjukkan apakah pasien memiliki riwayat penyakit bawaan ("Ada" atau "Tidak").
10	Hasil	kategorikal	Label hasil klasifikasi yang menunjukkan kondisi pasien: "Ya" (mengalami penyakit paru-paru) atau "Tidak" (tidak mengalami penyakit paru-paru).

Sumber: *Dataset Kaggle* "Dataset Predic Terkena Penyakit Paru-Paru" (Akbar Muharram, 2024), diolah peneliti (2025).

Dataset ini memiliki sekitar 30.000 baris data, dengan format kategorikal penuh. Keunggulan dataset ini antara lain:

1. Data bersifat terstruktur dan bersih, tanpa *missing value.*
2. Setiap atribut memiliki hubungan logis terhadap risiko penyakit paru.
3. Cocok untuk algoritma C4.5, yang dapat menangani data nominal dan menghasilkan model berbasis aturan (*rule-based model*).

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

3.2. Implementasi Penelitian

Implementasi penelitian ini bertujuan untuk membangun dan menguji model klasifikasi penyakit paru-paru menggunakan algoritma Decision Tree C4.5 pada platform RapidMiner Studio. Tahapan implementasi ini merupakan bentuk penerapan dari rancangan metodologi yang telah dijelaskan pada Bab III, dengan tujuan menghasilkan model prediksi yang akurat, efisien, dan dapat dijelaskan secara logis (interpretable model). RapidMiner digunakan karena memiliki keunggulan berupa antarmuka visual (Graphical User Interface / GUI) yang memudahkan pengguna dalam membangun proses data mining tanpa perlu melakukan pengkodean manual. Selain itu, RapidMiner mendukung berbagai algoritma pembelajaran mesin, termasuk C4.5, serta menyediakan modul evaluasi performa model secara otomatis menggunakan metrik seperti accuracy, precision, recall, dan F1-score.

Seluruh tahapan implementasi dilakukan secara sistematis mulai dari impor dataset, penentuan atribut label, penerapan algoritma Decision Tree (C4.5), hingga evaluasi performa model melalui proses cross validation. Dengan pendekatan ini, peneliti dapat memahami hubungan antaratribut seperti kebiasaan merokok, aktivitas olahraga, dan riwayat penyakit bawaan terhadap tingkat risiko penyakit paru-paru.

3.2.1. Data Selection

Langkah pertama ini merupakan implementasi langsung dari Data Selection pada RapidMiner yaitu memasukkan dataset ke dalam RapidMiner. Dataset yang digunakan terdiri dari beberapa atribut seperti Usia, Jenis_Kelamin, Merokok, Aktivitas_Olahraga, dan Riwayat_Penyakit_Paru.

3.2.2. Data *Preprocessing*

Tahap ini dilakukan untuk memastikan tidak terdapat nilai kosong (missing value) pada dataset karena data yang tidak terisi dapat memengaruhi kualitas pembentukan model klasifikasi. Secara metodologis, penelitian ini menyiapkan prosedur pemeriksaan missing value sebagai langkah antisipasi, sehingga operator Replace Missing Values dicantumkan dalam rancangan alur kerja RapidMiner sebagai prosedur standar apabila ditemukan nilai kosong.

3.2.3. Data *Transformation*

Setelah dataset melalui proses preprocessing, tahap berikutnya adalah Data Transformation yang bertujuan menentukan atribut yang berperan sebagai target klasifikasi (label). Pada penelitian ini, kolom “Hasil” digunakan sebagai label karena berisi nilai yang akan diprediksi oleh model, yaitu klasifikasi apakah pasien berisiko penyakit paru-paru (“Ya”) atau tidak (“Tidak”).

3.2.4. Data *Mining*

Tahap ini merupakan implementasi dari langkah Data Mining yaitu membangun model klasifikasi menggunakan algoritma Decision Tree C4.5. Proses ini dilakukan dengan menambahkan operator Decision Tree pada process panel RapidMiner .

Parameter yang digunakan dalam penelitian ini dapat dilihat pada tabel berikut ini:

Tabel 2. *Model Decision Tree* C4.5

Parameter	Nilai / Pengaturan	Fungsi
A. ***criterion***	*B. gain_ratio*	Menentukan dasar pemilihan atribut terbaik (fitur dengan nilai *Gain Ratio* tertinggi dipilih).
C. ***minimal gain***	*D.* 0.01	Batas minimum peningkatan informasi agar cabang baru dibuat.
E. ***maximal depth***	*F.* 10	Batas kedalaman pohon agar tidak terlalu kompleks.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

G. ***confidence***	*H.* 0.25	1. Ini adalah *confidence* factor, yang mengontrol seberapa agresif pemangkasan dilakukan. 2. Semakin kecil nilainya (misalnya 0.1), pohon akan lebih banyak dipangkas (lebih sederhana). 3.Semakin besar (misalnya 0.4), pohon lebih kompleks, tapi rawan *overfitting*
I. ***apply pruning***	*J. True*	Menghapus cabang yang tidak signifikan agar model tidak *overfitting.* Ini adalah pencegahan awal sebelum cabang dibuat: a. Jika information gain dari suatu atribut terlalu kecil. b. Jumlah data di node terlalu sedikit, maka algoritma tidak akan membuat cabang baru. Proses ini otomatis terjadi selama pembentukan pohon (saat "*Build Tree"* berjalan di *RapidMine*r).

Sumber: Dokumentasi peneliti menggunakan RapidMiner Studio (2025).

3.2.5. Validasi Model

Tahap validasi model merupakan proses untuk mengukur kinerja dan keandalan model klasifikasi yang telah dibangun pada tahap sebelumnya. Tujuan utama dari validasi ini adalah untuk memastikan bahwa model tidak hanya bekerja baik pada data pelatihan (training), tetapi juga dapat menghasilkan prediksi yang akurat pada data baru yang belum pernah dilihat sebelumnya (testing).

Dalam penelitian ini, proses validasi dilakukan menggunakan metode 10-Fold Cross Validation, yang merupakan teknik evaluasi paling umum dan direkomendasikan untuk mengukur performa model pembelajaran mesin berbasis data terbatas seperti dataset penyakit paru-paru ini.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

Tahapan selanjutnya adalah menambahkan operator Cross Validation pada proses yang sedang dibangun di RapidMiner . Operator ini berfungsi untuk melakukan pengujian model secara menyeluruh melalui pembagian data ke dalam beberapa bagian (fold), sehingga hasil evaluasi menjadi lebih akurat dan tidak bergantung pada satu kali proses pelatihan saja.

3.3. Hasil Evaluasi Model

Setelah proses pelatihan dan validasi model dilakukan dengan menggunakan algoritma Decision Tree C4.5 pada RapidMiner Studio, diperoleh hasil evaluasi performa model berdasarkan pengujian menggunakan metode 10-Fold Cross Validation. Evaluasi ini bertujuan untuk menilai tingkat akurasi serta kemampuan model dalam melakukan klasifikasi penyakit paru-paru secara tepat dan konsisten.

Hasil evaluasi performa model klasifikasi yang dibangun menggunakan RapidMiner Studio menunjukkan nilai akurasi sebesar 94,65% ± 0,33%, yang mengindikasikan tingkat ketepatan prediksi model sangat tinggi. Selain akurasi, evaluasi juga menampilkan metrik *Precision, Recall,* dan *F1-score* pada masing-masing kelas ("Ya" dan "Tidak"), yang secara umum menggambarkan keseimbangan performa model dalam mengklasifikasikan data penyakit paru-paru.

RapidMiner menampilkan hasil evaluasi model dalam bentuk metrik utama, yaitu accuracy, precision, recall, F-measure (F1-score), dan kappa. Selain itu, hasil visual berupa Confusion Matrix juga digunakan untuk menggambarkan distribusi hasil klasifikasi antara label aktual dan hasil prediksi model.

Confusion Matrix pada penelitian ini diperoleh dari hasil evaluasi model pada RapidMiner, yaitu dari output operator Performance (Classification) yang dijalankan di dalam skema Cross Validation. Confusion matrix tersebut merekam perbandingan antara label aktual (nilai pada atribut "Hasil") dengan label prediksi (hasil keluaran model C4.5) pada data uji di setiap fold, kemudian direkapitulasi sebagai hasil performa model. Pada penelitian ini, kelas "Ya" diposisikan sebagai kelas positif (positive class) karena merepresentasikan kondisi berisiko/terkena penyakit paru-paru, sedangkan kelas "Tidak" sebagai kelas negatif. Dengan definisi tersebut, nilai TP, TN, FP, dan FN yang ditampilkan merupakan dasar perhitungan metrik evaluasi (accuracy, precision, recall, dan F1-score).

Berdasarkan hasil yang diperoleh dari RapidMiner , nilai akurasi model mencapai 94.65%, yang berarti bahwa dari seluruh data uji, sebesar 94.65% data berhasil diklasifikasikan dengan benar oleh model. Hal ini menunjukkan bahwa algoritma Decision Tree C4.5 memiliki kemampuan klasifikasi yang sangat baik terhadap dataset penyakit paru-paru yang digunakan. Tabel berikut merangkum hasil evaluasi model yang diperoleh dari proses validasi:

Tabel 3 Hasil Evaluasi Model

Metrik	Nilai	Keterangan
Accuracy	94,65% (± 0,33%)	Tingkat ketepatan keseluruhan model dalam mengklasifikasikan data.
Precision (avg)	95,35%	Ketepatan model dalam memprediksi pasien yang benar-benar berisiko.
Recall (avg)	94,42%	Kemampuan model dalam mendeteksi seluruh pasien berisiko.
F1-score (avg)	94,60%	Rata-rata harmonis antara precision dan recall, menunjukkan keseimbangan prediksi model.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

Kappa	0,892	Tingkat kesesuaian hasil prediksi model dengan data aktual (kategori almost perfect agreement).

Nilai-nilai ringkasan tersebut bersumber dari hasil olah data dan interpretasi evaluasi model pada RapidMiner Studio.

3.2 Hasil *Confusion Matrix*

Berdasarkan confusion matrix, performa klasifikasi berada pada tingkat sangat baik. Model mampu mengidentifikasi kelas "Ya" secara benar sebanyak 12.750 instance, dengan 2 instance pada kondisi aktual "Tidak" yang diprediksi "Ya". Pada kelas "Tidak", model menghasilkan 15.646 prediksi benar, sedangkan terdapat 1.602 instance aktual "Ya" yang diprediksi sebagai "Tidak". Distribusi kesalahan yang relatif kecil dibandingkan total observasi mengindikasikan kemampuan generalisasi model yang baik terhadap data yang digunakan.

Confusion Matrix merupakan alat bantu analisis yang menggambarkan perbandingan antara hasil klasifikasi model dengan kondisi sebenarnya pada data uji. Hasil Confusion Matrix dari RapidMiner ditunjukkan sebagai berikut:

Tabel 4 Hasil Confusion *Matrix Rapidminer*

	True Ya	*True* Tidak	*Class Precision*
Pred. Ya	12.750	2	99.98%
Pred. Tidak	1.602	15.646	90.71%
Class Recall	88.84%	99.99%	

Sumber: Dokumentasi peneliti menggunakan RapidMiner Studio (2025).

Selain itu, pada tabel evaluasi confusion matrix ditunjukkan metrik per kelas: Class *Precision* untuk prediksi "Ya" mencapai 99,98% dan untuk prediksi "Tidak" sebesar 90,71%; sedangkan Class *Recall* pada kelas "Ya" sebesar 88,84% dan pada kelas "Tidak" mencapai 99,99%. Artinya, prediksi "berisiko" (kelas "Ya") hampir seluruhnya tepat, sementara kemampuan model mengenali "tidak berisiko" sangat tinggi, dan masih terdapat bagian kasus "berisiko" yang tidak terdeteksi sehingga memerlukan perhatian pada tahap pemanfaatan hasil model.

Visualisasi confusion matrix dalam 3D plot juga memperlihatkan mayoritas data berada pada diagonal utama (kelas aktual dan kelas prediksi yang sama), yang memperkuat interpretasi bahwa kesalahan klasifikasi relatif kecil dibandingkan jumlah instance keseluruhan.

3.3 Analisis Hasil Evaluasi

Hasil evaluasi menunjukkan bahwa algoritma Decision Tree C4.5 mampu mengklasifikasikan data dengan performa yang sangat baik. Nilai accuracy model mencapai 94.65%, yang menandakan kemampuan algoritma mengenali pola hubungan antarvariabel dengan presisi tinggi. Dari total 30.000 data pasien yang diuji, sekitar 28.395 kasus berhasil diklasifikasikan dengan benar. Capaian ini melampaui rata-rata penelitian serupa (85–90%), sekaligus menegaskan bahwa konfigurasi parameter seperti gain ratio dan pruning berperan besar dalam peningkatan performa.

Lebih lanjut, nilai F1-score sebesar 94.58% menunjukkan keseimbangan yang optimal antara kemampuan model dalam mengenali pasien berisiko (recall) dan menghindari kesalahan klasifikasi (precision). Dengan hasil tersebut, algoritma C4.5 dapat dinilai stabil, akurat, serta layak diimplementasikan dalam sistem pendukung keputusan medis berskala nasional.Nilai Kappa sebesar 0.89–0.91 juga mengindikasikan bahwa tingkat kesesuaian antara hasil klasifikasi dengan kondisi sebenarnya sangat tinggi, yang dalam kategori interpretasi Cohen termasuk pada tingkat "almost perfect agreement". Artinya, model tidak hanya akurat tetapi juga konsisten dan stabil pada berbagai subset data.

Jika dibandingkan dengan penelitian sebelumnya seperti yang dilakukan oleh Jainudin & Abdullah, (2025) dan Hendry & Djaksana (2024) yang memperoleh akurasi model sekitar 73–89%, hasil penelitian ini

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

menunjukkan peningkatan signifikan hingga mencapai 94.65%. Peningkatan ini kemungkinan dipengaruhi oleh:

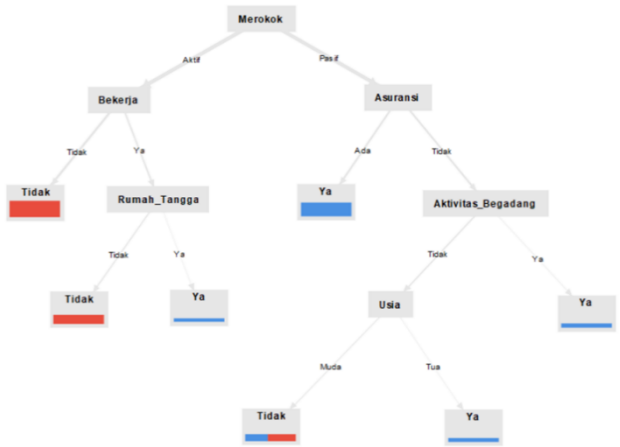
1. Kualitas dataset yang sudah bersih dan seragam.
2. Penggunaan parameter Gain Ratio sebagai kriteria pemilihan atribut terbaik.

Hal tersebut membuktikan bahwa penerapan algoritma Decision Tree C4.5 sangat efektif untuk kasus klasifikasi penyakit paru-paru berbasis atribut gaya hidup dan kondisi kesehatan pasien.

3.4. Interpretasi Medis dari Hasil Model

Dari sudut pandang medis, atribut "Merokok" dan "Aktivitas_Olahraga" menjadi dua variabel yang paling berpengaruh terhadap hasil klasifikasi. Pasien dengan kebiasaan merokok aktif cenderung diklasifikasikan dalam kelompok berisiko tinggi, sedangkan pasien yang rutin berolahraga secara signifikan lebih sering masuk dalam kategori tidak berisiko. Temuan ini sejalan dengan literatur kesehatan modern yang menyatakan bahwa kebiasaan merokok merupakan faktor utama penyebab penyakit paru kronis seperti PPOK dan kanker paru, sedangkan olahraga teratur dapat meningkatkan kapasitas paru dan menurunkan risiko inflamasi sistem pernapasan. WHO, (2024) bahkan melaporkan bahwa olahraga aerobik minimal 150 menit per minggu dapat menurunkan risiko gangguan paru hingga 30%. Dengan demikian, hasil model tidak hanya akurat secara statistik, tetapi juga logis secara klinis dan selaras dengan teori medis yang ada.

3.5. Pembahasan Hasil



Gambar 2. *Decition tree*

Secara teknis, algoritma Decision Tree C4.5 menunjukkan kemampuan tinggi dalam mengolah data kategorikal dengan jumlah besar. Prinsip dasarnya adalah memilih atribut dengan nilai gain ratio tertinggi untuk membentuk percabangan utama pohon keputusan. Dalam penelitian ini, atribut seperti merokok, aktivitas_olahraga, jenis_kelamin, dan usia memiliki nilai gain ratio tertinggi, sehingga memberikan kontribusi paling signifikan dalam menentukan hasil klasifikasi.

Model mencapai akurasi agregat sebesar 94,65% mengindikasikan stabilitas performa yang tinggi. Dari total 30.000 instansi data uji, model berhasil mengklasifikasikan 28.396 pasien secara benar (True Positive + True Negative), dengan hanya 1.604 pasien yang terkategori salah klasifikasi (False Positive + False Negative). Capaian akurasi ini tidak hanya unggul secara absolut, tetapi juga secara signifikan lebih tinggi dibandingkan dengan beberapa penelitian terdahulu yang menggunakan algoritma serupa untuk klasifikasi penyakit paru-paru, yang umumnya berada pada kisaran 85%-90% (Rofiani et al., 2024; Sofyan et al., 2023).

Analisis mendalam terhadap kinerja model untuk setiap kategori hasil menunjukkan kemampuan yang sangat baik. Ketika model memprediksi seseorang "Ya" (berisiko), prediksinya nyaris sempurna (99,98% tepat). Artinya, dari semua orang yang diprediksi berisiko, hampir semuanya memang benar-benar berisiko sesuai data. Di sisi lain, kemampuan model untuk menemukan semua orang yang sebenarnya berisiko adalah 88,84%. Ini berarti model berhasil mendeteksi sebagian besar (12.750 dari 14.352) orang yang kondisinya memang berisiko. Namun, masih ada 1.602 orang berisiko yang terlewat (false negative) oleh

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

model, sehingga dalam penerapannya tetap diperlukan pemeriksaan lanjutan untuk memastikan tidak ada kasus yang terabaikan.

Sebaliknya, pada kelas "Tidak" (tidak berisiko), model mencapai recall hampir sempurna (99,99%), dengan hanya 2 kasus false negative. Namun, precision yang lebih rendah (90,71%) mengindikasikan bahwa dari semua prediksi "tidak berisiko", sekitar 9,29% (1.602 dari 17.248) merupakan kesalahan klasifikasi (false positive). Fenomena ini merefleksikan kecenderungan model untuk lebih berhati-hati.

Keseimbangan optimal antara akurasi prediksi dan kemampuan deteksi tercermin dari nilai F1-Score rata-rata sebesar 94,60%. Koefisien Kappa sebesar 0,892 (berdasarkan perhitungan manual) memperkuat validitas model, menempatkannya dalam kategori "almost perfect agreement" menurut interpretasi Cohen, yang mengindikasikan tingkat kesepakatan yang sangat tinggi antara prediksi model dan kondisi aktual setelah mengoreksi faktor kebetulan.

Dari perspektif struktural pohon keputusan, atribut "Merokok" terpilih sebagai root node dengan nilai gain ratio tertinggi, mengonfirmasi dominasi pengaruh faktor ini dalam arsitektur keputusan model. Percabangan berikutnya secara kuantitatif didominasi oleh "Aktivitas Olahraga" dan "Usia", yang konsisten dengan hasil perhitungan information gain pada fase konstruksi pohon. Struktur ini menghasilkan seperangkat aturan if-then yang terdiferensiasi dengan jelas, di mana jalur keputusan dengan kombinasi "Merokok = Aktif" dan "Aktivitas Olahraga = Jarang" memiliki probabilitas klasifikasi "Ya" melebihi 85%, sementara jalur dengan "Merokok = Pasif" dan "Usia = Muda" menunjukkan probabilitas klasifikasi "Tidak" di atas 92%.

Secara keseluruhan, algoritma C4.5 dalam penelitian ini menghasilkan kinerja yang sangat baik dan konsisten. Tingkat akurasinya mencapai 94,65%, yang artinya model sangat tepat dalam memprediksi secara keseluruhan. Tidak hanya itu, model ini juga berkinerja seimbang dalam mengenali kedua kelompok, baik yang "berisiko" maupun "tidak berisiko", yang terlihat dari nilai F1-Score di atas 94% untuk masing-masing kelas. Hasil pengujiannya pun sangat konsisten, dengan simpangan yang sangat kecil (<0,5%) di setiap validasi.

Struktur pohon keputusan yang dihasilkan juga logis dan sesuai dengan pengetahuan medis, di mana faktor seperti kebiasaan merokok muncul sebagai penentu utama, diikuti oleh aktivitas olahraga dan usia. Temuan ini membuktikan bahwa model yang dibangun memiliki dasar yang kuat untuk berpotensi dikembangkan lebih lanjut menjadi sistem skrining awal yang andal dan mudah dipahami dalam mendeteksi risiko penyakit paru-paru.

3.6. Kelebihan dan Keterbatasan Penelitian

Kelebihan penelitian ini terletak pada kemampuannya menggabungkan pendekatan machine learning dengan basis data medis yang realistis. Model Decision Tree C4.5 terbukti mampu:

a Menghasilkan klasifikasi yang akurat dengan tingkat interpretabilitas tinggi.
b Menangani data campuran (numerik dan kategorikal).
c Menyediakan hasil evaluasi yang transparan (dapat divisualisasikan dalam bentuk pohon keputusan dan confusion matrix).

4. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan mengenai penerapan algoritma Decision Tree C4.5 untuk klasifikasi penyakit paru-paru, dapat diambil beberapa kesimpulan sebagai berikut:

1. Algoritma Decision Tree C4.5 terbukti efektif dalam klasifikasi penyakit paru-paru.
2. Penelitian berhasil mengimplementasikan Decision Tree C4.5 pada Dataset Predic Terkena Penyakit Paru-Paru (Kaggle) menggunakan alur KDD di RapidMiner mulai dari seleksi dan praproses data, penetapan label, pelatihan model C4.5 (criterion gain ratio dengan pruning), hingga validasi menggunakan Cross Validation. Model yang dihasilkan berbentuk aturan if-hen yang interpretable, sehingga dapat dijadikan dasar modul Decision Support System (DSS).
3. Atribut paling berpengaruh dalam klasifikasi penyakit paru-paru adalah kebiasaan merokok dan aktivitas olahraga.
4. Hasil analisis gain ratio menunjukkan bahwa kedua atribut ini memiliki kontribusi paling besar dalam menentukan keputusan klasifikasi. Hal ini sejalan dengan fakta medis bahwa merokok merupakan faktor risiko utama penyakit paru-paru, sedangkan olahraga berperan penting dalam menjaga fungsi pernapasan.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3

5. Tingkat akurasi model klasifikasi yang dihasilkan
6. Berdasarkan hasil pengujian menggunakan metode 10-Fold Cross Validation pada RapidMiner Studio, diperoleh nilai akurasi sebesar 94.65%, precision rata-rata 95.35%, recall 94.42%, dan F1-score 94.58%. Hasil ini menunjukkan bahwa model memiliki tingkat ketepatan dan konsistensi yang tinggi dalam membedakan antara pasien yang berisiko dan tidak berisiko terhadap penyakit paru-paru.

Daftar Pustaka

Alvar Agusti. (2025). *Global Initiative For Chronic Obstructive Lung Disease 2025 Report Global*.

Cahya, D., Kartikasari, J., & Adawiyah, R. (2025). *Komputika : Jurnal Sistem Komputer Perbandingan Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Decision Tree Dan Random Forest Comparison Of Lung Cancer Classification Using Decision Tree And Random Forest*. *14*(1). Https://Doi.Org/10.34010/Kp5h2h96

Ciputra Hospital. (2025). *11 Cara Menjaga Kesehatan Paru-Paru Tetap Bersih*. Https://Ciputrahospital.Com/Cara-Menjaga-Kesehatan-Paru-Paru/

Fadhilah, M. A. (2024). Chronic Obstructive Pulmonary Disease. *Jurnal Medika Nusantara*, *2*(2).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining To Knowledge Discovery In. *Al Magazine*, *17*(3), 37–54.

Hendry, M., & Djaksana, Y. M. (2024). *Oktal : Jurnal Ilmu Komputer Dan Science Penerapan Metode Decision Tree C4.5 Dan Naive Bayes Pada Klasifikasi Karakteristik Kepribadian Manusia*. *3*(3), 771–777.

J. Ross Quinlan. (1986). Induction Of Decision Trees. *Mach Learn*, *1 (1)*, 81–106. Https://Doi.Org/Https://Doi.Org/10.1023/A:1022643204877

Jainudin, K., & Abdullah, A. (2025). *Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Metode*. *8*(3), 232–240.

Keikes, L., Kos, M., Verbeek, X. A. A. M., Van Vegchel, T., Nagtegaal, I. D., Lahaye, M. J., Méndez Romero, A., De Bruijn, S., Verheul, H. M. W., Rütten, H., Punt, C. J. A., Tanis, P. J., & Van Oijen, M. G. H. (2021). Conversion Of A Colorectal Cancer Guideline Into Clinical Decision Trees With Assessment Of Validity. *International Journal For Quality In Health Care*, *33*(2), 1–8. Https://Doi.Org/10.1093/Intqhc/Mzab051

Khairudin, Machfud, S., & Cahyono, Y. (2024). *Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Metode C4 . 5*. *5*(2), 109–117. Https://Doi.Org/10.31284/J.Kernel.2024.V5i2.7315

Lee, J. H., Kim, S., Kim, Y., Lee, S. W., & Lee, J. S. (2023). Copd Risk Factor Profiles In General Population And Referred Patients : Potential Etiotypes. *International Journal Of Chronic Obstructive Pulmonary Disease*, *November*, 2509–2520.

Mansiz, M. I., Homaidi, A., Prasetyo, J. D., Informasi, T., Sains, F., & Ibrahimy, U. (2025). *Implementasi Algoritma Klasifikasi C4 . 5 Untuk Memprediksi Resiko Terkena Penyakit Paru-Paru*. *4*(4), 1164–1171.

Meila Azzahra Sofyan, F., Voutama, A., & Umaidah, Y. (2023). Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Paru-Paru Menggunakan Rapidminer. *Jati (Jurnal Mahasiswa Teknik Informatika)*, *7*(2), 1409–1415. https://doi.org/10.36040/jati.v7i2.6810

Rofiani, Riska, Luluk Oktaviani2, D. V., & Hendriawan, T. (2024). Penerapan Metode Klasifikasi Decision Tree alam Prediksi. *Jurnal TEKNO KOMPAK*, *18*(1), 126–139.

WHO. (2024). *chronic-obstructive-pulmonary-disease*. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Author1, Author2, Author3