DOI: https://doi.org/10.37817/ikraith-informatika.v9i3
P-ISSN : 2580-4316
E-ISSN : 2654-8054

Deteksi Dini Penyakit Kanker Paru-Paru Menggunakan Perbandingan Algoritma Xgboost Dan Decision Tree

[1]Keysha Salma Hanun Anzani, [2]Wiwin Windihastuty
[1,2]Sistem Informasi, Universitas Budi Luhur, Jakarta

E-mail: [1]keyshasalma@gmail.com, [2] wiwin.windihastuty@budiluhur.ac.id

ABSTRAK

Kanker paru-paru merupakan salah satu penyebab utama kematian tertinggi di seluruh dunia, sehingga penting melakukan deteksi dini untuk meningkatkan kemungkinan sembuh bagi pasien. Namun, proses deteksi seringkali terhambat oleh berbagai faktor, termasuk kurangnya tenaga medis, ketepatan diagnosis, serta rendahnya kesadaran masyarakat yang seringkali menunda pemeriksaan atau merasa takut untuk konsultasi ke dokter saat munculnya gejala awal. Penelitian ini bertujuan untuk menciptakan model klasifikasi berbasis data mining yang dapat membantu dalam mendeteksi kanker paru-paru lebih awal. Algoritma yang digunakan adalah Decision Tree dan XGBoost, didukung oleh metode SMOTE untuk mengatasi ketidakseimbangan kelas dalam data. Dataset yang dianalisis diperoleh dari website kaggle dengan berbagai atribut klinis dan gaya hidup pasien. Model yang dikembangkan yaitu model dengan akurasi dan AUC yang tertinggi mencapai 93.89% dan 0.972 yaitu algoritma XGBoost. Hasil ini menunjukkan algoritma XGBoost dapat memberikan kinerja yang lebih baik dalam mendeteksi kanker paru-paru secara dini. Penelitian ini menunjukkan bahwa usia adalah faktor yang paling berpengaruh dalam mendeteksi kanker paru-paru. Selain itu, gejala mengi (wheezing) dan tekanan sosial (peer pressure) juga berkontribusi pada penyakit kanker paru- paru.

Kata kunci : ***Kanker paru-paru, Deteksi Dini, Decision Tree, XGBoost, SMOTE***

ABSTRACT

Lung cancer is one of the leading causes of death worldwide, making early detection crucial to improving the chances of recovery. However, the detection process is often hampered by various factors, including a shortage of medical personnel, inaccurate diagnoses, and low public awareness, often leading to delays in examinations or fear of consulting a doctor when initial symptoms appear. This study aims to create a data mining-based classification model that can help detect lung cancer early. The algorithms used are Decision Tree and XGBoost, supported by the SMOTE method to address class imbalance in the data. The analyzed dataset was obtained from the Kaggle website with various clinical and lifestyle attributes of patients. The developed model, namely the model with the highest accuracy and AUC reaching 93.89% and 0.972, is the XGBoost algorithm. These results indicate that the XGBoost algorithm can provide better performance in detecting lung cancer early. This study shows that age is the most influential factor in detecting lung cancer. In addition, wheezing symptoms and social pressure (peer pressure) also contribute to the development of lung cancer.

Keyword : ***Lung Cancer, Early Detection, Decision Tree, XGBoost, SMOTE***

P-ISSN : 2580-4316
E-ISSN : 2654-8054
DOI: https://doi.org/10.37817/ikraith-informatika.v9i3

1. PENDAHULUAN

Paru-paru adalah sistem respirasi pada manusia yang sangat penting untuk menyediakan oksigen yang diperlukan oleh tubuh. Selain itu, paruparu juga berfungsi sebagai tempat di mana oksigen dari udara ditukar dengan karbon dioksida dari darah. Dalam beberapa keadaan, paru-paru bisa mengalami masalah yang berdampak negatif pada fungsi sistem pernapasan. Jika paru-paru tidak berfungsi dengan optimal, hal ini dapat memicu munculnya berbagai penyakit **(Hasibuan, 2024)**. Beberapa jenis penyakit paru-paru yang paling umum dan berbahaya termasuk tuberkulosis, pneumonia, dan kanker paru-paru **(Khoirul Anwar, 2025)**.

Kanker paru-paru adalah salah satu tipe kanker yang paling mematikan di seluruh dunia. Data dari World Health Organization (WHO) menunjukkan bahwa kanker paru-paru menjadi penyebab terbesar kematian akibat kanker secara global. Menurut informasi dari Globocan 2020, diperkirakan ada sekitar 45.000 kematian disebabkan oleh kondisi tersebut. Statistik ini menekankan bahwa kanker paru adalah salah satu jenis kanker paling mematikan di Indonesia. Tingginya angka insiden kanker paru di negara ini menunjukkan perlunya langkah-langkah pencegahan dan deteksi awal yang serius. Sebagai langkah lanjutan, penting untuk melakukan edukasi dan kampanye mengenai bahaya merokok dan faktor risiko lainnya secara aktif guna menurunkan jumlah kasus kanker paru di Indonesia **(Abdi Mangun, 2023)**. Deteksi awal terhadap kanker paru- paru adalah Tindakan penting untuk meningkatkan kemungkinan pasien dapat sembuh. Namun, masih banyak orang yang ragu atau merasa takut untuk konsultasi dengan dokter, terutama jika gejala awal dianggap tidak serius. Proses identifikasi kanker paru-paru juga sering terkendala oleh keterbatasan tenaga medis dan tepetan diagnosis. Maka dari itu, penggunaan kecerdasan buatan dengan metode klasifikasi bisa menjadi solusi yang efektif untuk mendeteksi penyakit ini lebih awal **(Nuraeni, 2025)**.

Dalam penelitian ini, diterapkan dua algoritma yaitu XGBoost dan Decision Tree untuk mengklasifikasikan data pasien dalam mendeteksi kanker paru-paru. XGBoost (eXtreme Gradient Boosting) merupakan metode gradient boosting yang dirancang agar sangat efektif. Pendekatan ini memakai berbagai cara modern untuk memperkecil tingkat kesalahan dan mengoptimalkan performa model. Popularitasnya didasarkan pada keandalannya serta ketepatan prediksinya yang luar biasa, menjadikannya pilihan yang sangat baik untuk menyelesaikan persoalan klasifikasi yang kompleks **(Sen, 2024)**. Decision Tree adalah metode klasifikasi dan prediksi yang sangat terkenal dan efisien. Pada Decision Tree ini, data yang merupakan fakta diubah menjadi struktur pohon yang berisi aturan, yang pastinya lebih mudah dipahami dengan bahasa sehari-hari **(Jatnika Fahmi Idris, 2024)**.

Dengan menggunakan algoritma *XGBoost* dan *Decision Tree* pada data survey diharapkan dapat tercipta model prediksi yang mampu mengelompokkan kondisi kanker paru-paru menjadi kategori terdeteksi atau tidak terdeteksi, serta memberikan kemampuan untuk mengidentifikasi risiko kanker paru- paru dengan lebih cepat dan efisien. Penggunaan kedua algoritma ini juga bertujuan untuk menilai algoritma mana yang lebih tepat dalam melakukan prediksi.

Penelitian ini dilakukan dengan memanfaatkan dataset “survey lung cancer” yang berisi informasi terkait gejala dan kebiasaan pasien, seperti merokok, sesak nafas, dan riwayat kesehatan. Penulis tertarik mengangkat

DOI: https://doi.org/10.37817/ikraith-informatika.v9i3
P-ISSN : 2580-4316
E-ISSN : 2654-8054

topik ini karena kanker paru-paru masih menjadi masalah serius di bidang kesehatan dan teknologi data mining terbukti dapat memberikan solusi alternatif dalam proses diagnosis.

2. LANDASAN TEORI

Penelitian ini berfokus pada deteksi dini kanker paru-paru dengan memanfaatkan pendekatan data mining dan machine learning sebagai upaya pendukung diagnosis medis yang lebih akurat. Kerangka analisis yang digunakan mengacu pada model CRISP-DM (Goenawan Brotosaputro, 2025), yang terdiri dari tahap pemahaman data, persiapan, pemodelan, evaluasi, hingga implementasi. Dengan tahapan tersebut, penelitian diharapkan mampu menghasilkan model prediksi yang dapat membantu tenaga kesehatan dalam mengidentifikasi resiko penyakit sejak awal.

Secara umum, data mining didefinisikan sebagai proses pengolahan dan analisis data dalam jumlah besar untuk menemukan pola tersembunyi yang bermanfaat. Metode ini telah banyak dimanfaatkan pada berbagai bidang, termasuk kesehatan, karena mampu mengekstraksi informasi penting dari data medis. Teknik yang lazim digunakan dalam data mining antara lain klasifikasi, klastering, regresi, asosiasi, serta deteksi anomali **(Muttaqin, 2023)**. Salah satu pendekatan yang paling berkembang adalah machine learning, yang terbukti efektif dalam mengolah big data klinis dan memberikan hasil prediksi dengan tingkat ketepatan tinggi.

a) Algoritma Decision Tree

Decision Tree merupakan sebuah struktur yang menggunakan proses berurutan. Proses ini dimulai dari titik awal, kemudian dilakukan penilaian terhadap suatu fitur dan memilih salah ini berlanjut terus sampai mencapai cabang terakhir, yang dikenal sebagai daun, yang biasanya mencerminkan tujuan akhir yang ingin dicapai. Pemilihan atribut yang menjadi akar dilakukan berdasarkan atribut dengan nilai gain tertinggi di antara seluruh atribut yang tersedia. Nilai gain dihitung menggunakan rumus yang ditunjukkan sebagai berikut **(Sri Indra Maiyanti, 2023)**.

$$Gain(S,A) = Entropy(S) \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$

Entropi digunakan untuk mengukur seberapa informatif suatu atribut dalam menghasilkan sebuah keputusan. Adapun rumus Entropi adalah sebagai berikut:

$$Entropy(S) = \sum_{i=1}^{n} -(pi) * log_2 pi$$

b) Algoritma XGBoost

Extreme Gradient Boosting yang dikenal sebagai XGBoost adalah suatu metode pembelajaran supervised learning yang dapat digunakan untuk klasifikasi maupun regresi. Algoritma ini adalah sebuah sistem machine learning yang berbasis pada tree boosting dan dapat dioptimalkan untuk membangun dengan skala lebih besar. Sebagai suatu perkembangan dari Gradient Boosting, XGBoost merupakan metode ensemble yang menggunakan Decision Tree dan dirancang untuk mempercepat proses, bahkan ketika berhadapan dengan dataset yang besar. XGBoost beroperasi dengan menggabungkan sejumlah pengklasifikasi lemah menjadi sebuah model yang lebih tangguh, melalui pelatihan bertahap berdasarkan hasil klasifikasi sebelumnya, yang dikenal sebagai residuals atau error [9]. Rumus XGBoost menambahkan regulasi dalam fungsi tujuan untuk menghindari overfitting yang dijelaskan sebagai berikut **(Nurfansepta, 2025)**.

Optimasi Bobot Daun (Optimal Leaf Weights) :

P-ISSN : 2580-4316
E-ISSN : 2654-8054
DOI: https://doi.org/10.37817/ikraith-informatika.v9i3

$$w_j^* = -\frac{\sum_{i \epsilon Ij} g_i}{\sum_{i \epsilon Ij} h_i + \lambda}$$

Keterangan:

I_j = Set data yang jatuh ke daun j

$\sum g_i$ = Jumlah turunan pertama (gradient) untuk semua contoh yang jatuh ke daun tersebut

$\sum h_i$ = Jumlah turunan kedua (hessian)

λ = Parameter regularisasi yang mengontrol seberapa besar penalty diberikan pada pohon yang terlalu besar atau bobot yang terlalu besar

Split Finding:

$I_L dan\ I_R$ = Set contoh yang jatuh ke node kiri dan kanan setelah split

Kanker paru-paru sendiri merupakan salah satu masalah kesehatan global dengan tingkat mortalitas tinggi. Menurut Global Cancer Observatory WHO, kanker paru-paru adalah penyebab utama kematian akibat kanker di berbagai usia, baik pria maupun wanita. Ini ditandai oleh pertumbuhan sel tidak terkendali di jaringan paru-paru yang dapat meyebar ke organ lain melalui metastasis **(Jamil, 2024)**.

Berbagai penelitian sebelumnya telah menunjukkan hasil yang beragam dalam penerapan algoritma machine learning untuk kasus kanker paru-paru. Dinova dan Prasetiyo (2024) melaporkan akurasi sebesar 78% menggunakan Random Forest. Ishak dan Lauro (2025) menunjukkan kinerja tinggi dengan Decision Tree yang mencapai 98%. Penelitian Wulandari (2022) dengan Naïve Bayes mencatat akurasi 94,62%, sementara Lovita et al. (2022) menggunakan Logistic Regression dengan akurasi 90%. Pendekatan XGBoost bahkan menghasilkan nilai AUC 0,945 (Nurfansepta et al., 2025), sedangkan Rahmaeda et al. (2025) melaporkan kombinasi SVM dan Logistic Regression dengan akurasi rata-rata 96,77%.

Hasil-hasil ini menguatkan bahwa metode berbasis machine learning sangat potensial untuk mendukung deteksi dini kanker paru-paru (Guo et al., 2024)[12].

3. METODOLOGI

Penelitian ini menggunakan pendekatan data mining dengan kerangka CRISP-DM (Cross-Industry Standard Process for Data Mining). Model ini dipilih karena menyediakan tahapan yang sistematis, mulai dari pemahaman masalah, persiapan data, pemodelan, hingga evaluasi. Tujuan utama penelitian adalah membandingkan performa algoritma Decision Tree dan XGBoost dalam mengklasifikasikan potensi kanker paru-paru secara dini.

Dataset yang digunakan bersifat publik dan diperoleh dari situs Kaggle dengan judul Survey Lung Cancer. Data mencakup 309 baris dengan 16 atribut, yang terdiri atas informasi demografis (jenis kelamin, usia), kebiasaan hidup (merokok, konsumsi alkohol), serta kondisi kesehatan (penyakit kronis, batuk, sesak napas, dan lainnya). Variabel target adalah diagnosis pasien dengan label lung cancer (yes/no).

a) Tahapan Penelitian

Proses penelitian dilaksanakan secara bertahap mengikuti model CRISP-DM:

1. Pemahaman Masalah
2. Pemahaman Data
3. Persiapan Dat
4. Pemodelan
5. Evaluasi
6. Penerapan (Deployment)

b) Teknik Pengujian

1. Akurasi dipakai untuk melihat ketepatan keseluruhan prediksi.
2. Presisi digunakan untuk menilai keakuratan prediksi pada kelas positif.
3. Recall menekankan pada

DOI: https://doi.org/10.37817/ikraith-informatika.v9i3
P-ISSN : 2580-4316
E-ISSN : 2654-8054

kemampuan model mendeteksi kasus positif secara benar.

4. F1-Score menghitung keseimbangan antara presisi dan recall.
5. AUC digunakan untuk mengukur kekuatan model dalam membedakan kelas positif dan negatif.

c) Pengembangan Model

Hasil pengujian kedua algoritma dibandingkan untuk menentukan metode yang paling efektif. Algoritma dengan performa terbaik selanjutnya diusulkan sebagai alternatif solusi dalam sistem pendukung keputusan di bidang kesehatan, khususnya dalam deteksi dini kanker paru-paru.

4. HASIL DAN PEMBAHASAN

a) Business Understanding

Tujuan utama penelitian ini adalah mengembangkan model klasifikasi yang mampu mendeteksi risiko kanker paru-paru secara dini. Permasalahan yang diangkat berangkat dari tingginya angka kematian akibat kanker paru-paru serta keterbatasan tenaga medis dan biaya pemeriksaan. Dengan memanfaatkan data mining dan algoritma machine learning, diharapkan dapat dihasilkan model prediksi yang cepat, akurat, dan dapat digunakan sebagai sistem pendukung keputusan di bidang kesehatan.

b) Data Understanding

Dataset yang digunakan berasal dari laman Kaggle dengan judul Survey Lung Cancer. Data berjumlah 309 record dengan 16 atribut. Lima belas atribut digunakan sebagai variabel independen, sementara satu atribut digunakan sebagai label klasifikasi (lung cancer: Yes/No).

Table 1 Jumlah Record

No.	Nama Atribut	Tipe Atribut
1	*Gender*	Kategorikal
2	*Age*	Numerik
3	*Smoking*	Numerik
4	*Yellow Fingers*	Numerik
5	*Anxiety*	Numerik
6	*Peer Pressure*	Numerik
7	*Chronic Disease*	Numerik
8	*Fatigue*	Numerik
9	*Allergy*	Numerik
10	*Wheezing*	Numerik
11	*Alcohol Consuming*	Numerik
12	*Coughing*	Numerik
13	*Shortness of Breath*	Numerik
14	*Swallowing Difficulty*	Numerik
15	*Chestpain*	Numerik
16	*Lung cancer*	Kategorikal

c) Data Preparation

Proses pra-pengolahan meliputi:

1. Pemeriksaan data → memastikan tidak ada missing value.
2. Encoding → mengubah atribut kategorikal menjadi numerik.
3. SMOTE (Synthetic Minority Oversampling Technique) → digunakan untuk Menyeimbangkan distribusi kelas.
4. Cross Validation → data dibagi menggunakan 10-fold cross validation, sehingga hasil evaluasi lebih reliabel

d) Modeling

Penelitian ini membandingkan dua algoritma klasifikasi:

1. Decision Tree → sederhana, interpretatif, dan mampu menampilkan struktur

P-ISSN : 2580-4316
E-ISSN : 2654-8054
DOI: https://doi.org/10.37817/ikraith-informatika.v9i3

keputusan dalam bentuk pohon.

2. XGBoost → algoritma berbasis ensemble boosting yang lebih kompleks dan biasanya memiliki performa lebih tinggi dalam klasifikasi.

e) Evaluation

Model dievaluasi menggunakan Confusion Matrix dengan lima metrik utama: Accuracy, Precision, Recall, F1- Score, dan AUC.

Table 2 Perbandingan Algoritma Decision Tree dan XGBoost

Dataset	*Decision Tree*				
	Akurasi	*Precision*	*Recall*	AUC	F-1 *Score*
Cross Validation	91.48 %	92.05 %	91.48 %	0.921	91.51 %
Dataset	*XGBoost*				
	Akurasi	*Precision*	*Recall*	AUC	F-1 *Score*
Cross Validation	93.89 %	93.02 %	95.19 %	0.972	93.94 %

f) Deployment

Berdasarkan hasil pengujian, model XGBoost dipilih sebagai algoritma terbaik. Model ini berpotensi untuk diintegrasikan dalam sistem pendukung keputusan medis, khususnya sebagai alat bantu skrining awal pasien yang berisiko kanker paru-paru. Dengan penerapan model ini, tenaga medis dapat memperoleh gambaran awal yang lebih cepat dan akurat sebelum melakukan pemeriksaan lanjutan.

5. KESIMPULAN

Penelitian ini mengembangkan model klasifikasi untuk deteksi dini kanker paru-paru dengan membandingkan performa algoritma XGBoost dan Decision Tree. Pada tahap preprocessing, diterapkan metode SMOTE untuk mengatasi ketidakseimbangan kelas dalam data. Model yang dihasilkan yaitu model dengan akurasi tertinggi, yaitu algoritma XGBoost dengan nilai akurasi sebesar 93.89% dan AUC 0.972. Hasil ini menunjukkan bahwa XGBoost efektif dalam mengenali pola pada data dan menunjukkan bahwa usia adalah faktor yang paling berpengaruh dalam mendeteksi kanker paru- paru. Selain itu, gejala mengi (wheezing) dan tekanan sosial (peer pressure) juga berkontribusi pada deteksi dini kanker paru- paru, karena keduanya dapat menjadi indikator awal yang mendorong individu untuk melakukan pemeriksaan lebih lanjut.

6. UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada Universitas Budi Luhur yang telah memberikan dukungan fasilitas penelitian, serta kepada dosen pembimbing atas arahan dan bimbingan yang sangat berharga. Ucapan terima kasih juga ditujukan kepada keluarga dan rekan-rekan yang senantiasa memberikan dukungan moral selama penelitian ini berlangsung.

.

DAFTAR PUSTAKA

Abdi Mangun, N. I. (2023). Lung Cancer Analysis Using K-Nearst Neighbor Algorithm. *Jurnal Teknik Industri Sistem Informasi dan Teknik Informatika* , 68-61.

Goenawan Brotosaputro, J. S. (2025). Prediction of Claim Fund Reserves in Insurance Companies Using the ARIMA Method. *Jurnal Atmaluhur Sisfokom*, 1-7.

Hasibuan. (2024). Prediksi Penyakit Paru-Paru Menggunakan Algoritma Naive Bayes Dan Adaboost. *Jurnal Darmajaya*, 34-38.

Jamil, M. R. (2024). Komparasi Kinerja Algortima Machine Learning untuk Mendeteksi Penyakit Infeksi Saluran PErnafasan . *Jurnal Rekayasa Teknologi Informasi (JURTI)*, 84-92.

P-ISSN : 2580-4316
E-ISSN : 2654-8054

Jatnika Fahmi Idris, R. R. (2024). Klasifikasi Penyakit Kanker Paru Menggunakan Perbandingan Algoritma Machine Learning. *Jurnal Media Akademik* .

Khoirul Anwar, R. M. (2025). Analisis Performa Deteksi Penyakit Paru-Paru Dengan Model Klasifikasi Gambar Menggunakan LSTM Deep Learning. *Jurnal Ilmiah Universitas Batang Hari Jambi*, 972-979.

Muttaqin, W. W. (2023). *Pengenalan Data Mining.* Jakarta: Yayasan Kita Menulis.

Nuraeni, A. (2025). Pendekatan Machine Learning Untuk Deteksi Dini Kanker Paru-Paru: Mengoptimalkan Sensitivitas Dan Akurasi. *Jurnal Informatika Polinema* , 339-346.

Nurfansepta, A. G. (2025). Deteksi Mutasi Epidermal Growth Factor Receptor Pada Kanker Paru Menggunakan Extreme Gradient Boosting. *JPTIIK*, 1-9.

Sen, B. K. (2024). Comparative Analysis of Machine Learning Techniques for Arrhythmia Detection. *Journal of Electrical and Electronic Engineering*, 92-103.

Sri Indra Maiyanti, D. A. (2023). Perbandingan Klasifikasi Penyakit Kanker Paru-paru menggunakan Support Vector Machine dan K-Nearest Neighbor. *Jurnal Processor*, 54-62.