JURNAL KHATULISTIWA INFORMATIKA VOL. 13 NO. 2 Desember 2025
p-ISSN: 2339-1928 e-ISSN: 2579-633X Halaman : 152 – 156

DETEKSI KANKER PARU-PARU MENGGUNAKAN TEKNIK ENSEMBLE LEARNING : STUDI KOMPARATIF KINERJA ADABOOST

F. Lia Dwi Cahyanti[1];Elly Firasari[2];Nurul Khasanah[3]

Program Studi Sistem Infroamsi[1,2], Program Studi Informatika[3], Fakultas Teknologi dan Informasi
Universitas Nusa Mandiri
flia.fdc@nusamandiri.ac.id

INFO ARTIKEL

Diajukan :
21 Agustus 2025

Diterima :
25 November 2025

Diterbitkan:
31 Desember 2025

Kata Kunci :
Adaboost, Ensemble Learning, Kanker Paru-Paru, Smote

INTISARI

Penyakit kanker paru-paru merupakan penyebab kematian di berbagai negara yang memerlukan deteksi dini untuk meningkatkan peluang keberhasilan medis, namun pengembangan model prediksinya sering terkendala oleh ketidakseimbangan dataset. Penelitian ini bertujuan untuk melakukan analisis komparatif sistematis terhadap empat algoritma ensemble learning, yaitu XGBoost, LightGBM, Bagging, dan AdaBoost, guna mengidentifikasi model terbaik dalam memprediksi risiko kanker paru-paru. Metode penelitian meliputi pra-pemrosesan data menggunakan *Lung Cancer Survey Dataset*, penerapan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk menangani ketidakseimbangan kelas, serta evaluasi kinerja melalui 10-fold cross-validation. Hasil penelitian menunjukkan bahwa algoritma AdaBoost memberikan performa paling superior dengan perolehan akurasi pengujian sebesar 90,32% dan nilai AUC mencapai 0,9340. Model AdaBoost juga menunjukkan stabilitas tinggi dengan nilai presisi dan recall yang seimbang pada angka 0,9444. Implementasi SMOTE terbukti krusial dalam menyeimbangkan distribusi kelas dari semula 31:216 menjadi 216:216, sehingga seluruh model mampu menghasilkan nilai AUC di atas 0,92. Penelitian ini menyimpulkan bahwa kombinasi penyeimbangan data dan algoritma berbasis boosting merupakan solusi handal yang dapat diintegrasikan ke dalam sistem pendukung keputusan klinis untuk mempercepat diagnosis kanker paru-paru.

I. PENDAHULUAN

Penyakit kanker merupakan ancaman Kesehatan secara global yang harus diperhatiakan, penyakit ini menduduki peringkat kedua penyebab kematian terbesar dengan jumlah 9,6 jt orang terjangkit kanker (Rejeki et al., 2020). Salah satu penyakit yang banyak diderita yaitu kanker paru-paru, penyakit yang muncul akibat pertumbuhan sel abnormal dan tidak terkendali di dalam paru-paru, dipicu oleh paparan zat karsinogen di dalam tubuh manusia Dari data global, penyakit ini menempati urutan pertama sebagai penyebab kematian akibat kanker pada pria, serta urutan kedua pada wanita (Asiyah et al., 2025). Berdasarkan data GLOBOCAN, jutaan orang meninggal setiap tahun karena penyakit ini. Angka ini menunjukkan betapa besarnya beban yang harus ditanggung rumah sakit dan tenaga medis dalam menghadapi masalah kanker (Sung et al., 2021). Banyaknya kematian terjadi karena dekteksi yang terlambat, jika terdeteksi sejak awal, peluang pasien untuk sembuh dan bertahan hidup jauh lebih besar (Siegel et al., 2023).

Semakin canggihnya teknologi saat ini penggabungan Big Data dengan kecerdasan buatan (AI) telah membawa perubahan besar dalam dunia medis, terutama untuk mendiagnosis pasien (Tarumingkeng, 2025). Penelitian terdahulu yang telah dilakukan bahwa Pembelajaran Mesin (*Machine Learning*) memiliki kemampuan superior dalam menganalisis pola kompleks dari data klinis, gaya hidup, dan demografis pasien yang tidak terjangkau oleh analisis statistik konvensional (Wardhana et al., 2023). Masalah utama dalam penerapan model prediksi dalam *machine learning* yaitu ketidakseimbangan dataset (class imbalance), dimana jumlah sampel kelas terjangkit kanker sering kali jauh lebih kecil dibandingkan dengan kelas mayoritas pasien tidak terjangkit kanker). Ketidakseimbangan ini sering kali menghasilkan model yang bias terhadap kelas mayoritas, sehingga menghasilkan metrik akurasi yang tinggi secara artifisial namun gagal dalam

mendeteksi kasus positif yang sebenarnya (Yulian & Susanto, 2025).

Untuk mengatasi kendala tersebut, peneliti menggunakan teknik oversampling seperti Synthetic Minority Over-sampling Technique (SMOTE) yang terbukti efektif dalam menyeimbangkan distribusi data dengan menciptakan sampel sintetis pada kelas minoritas (Syahwaluddin & Alita, 2024). Di sisi lain, perkembangan algoritma Ensemble Learning telah menunjukkan performa yang menjanjikan dalam meningkatkan generalisasi model. Teknik seperti XGBoost, LightGBM, dan AdaBoost dikenal mampu mengagregasi kekuatan beberapa model dasar (base learners) untuk mengurangi risiko overfitting serta meningkatkan presisi prediksi (Chen & Guestrin, 2016; Ke et al., 2017). Kajian literatur menunjukkan bahwa algoritma berbasis boosting secara konsisten mengungguli model tunggal dalam berbagai tugas klasifikasi medis yang kompleks.

Berdasarkan urgensi dan hasil kajian pustaka tersebut, penelitian ini bertujuan untuk melakukan analisis komparatif sistematis terhadap empat algoritma Ensemble Learning terkemuka, yaitu XGBoost, LightGBM, Bagging, dan AdaBoost, pada dataset risiko kanker paru-paru yang telah dioptimasi menggunakan SMOTE. Kebaruan penelitian ini terletak pada evaluasi kinerja yang komprehensif melalui validasi silang (10-fold cross-validation) dengan menggunakan metrik kritis seperti Akurasi, Presisi, Recall, dan AUC (Area Under the Curve). Hasil penelitian ini diharapkan dapat memberikan rekomendasi berbasis bukti mengenai model terbaik yang dapat diintegrasikan ke dalam sistem pendukung keputusan klinis guna mempercepat deteksi dini dan meningkatkan prognosis pasien kanker paru-paru.

II. BAHAN DAN METODE

Dalam penelitian ini menerapkan Metodologi penelitian yang menguraikan langkah-langkah sistematis untuk membangun, melatih, dan mengevaluasi model prediksi kanker paru-paru menggunakan teknik Ensemble Learning.
Ensemble Learning merupakan teknik yang melibatkan pelatihan dan pemaduan beberapa model dasar (weak learners) untuk meningkatkan kualitas hasil prediksi, metode ini didasarkan pada teori bahwa integrasi dari berbagai model yang memiliki keterbatasan individu dapat membentuk satu kesatuan model yang jauh lebih akurat dalam memecahkan suatu masalah (Cendani & Wibowo, 2022).

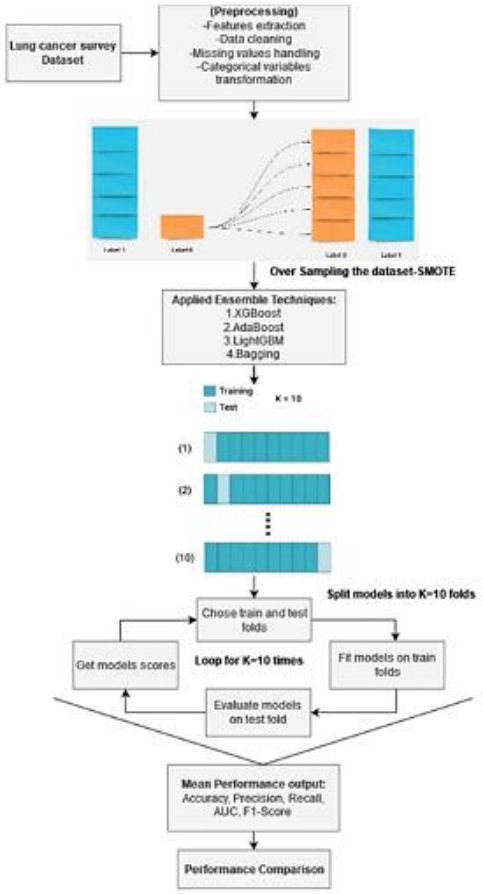
Sumber data yang digunakan dalam penelitian ini Adalah *Lung Cancer Survey Dataset* yang diperoleh dari data public (*Kaggle*), terdiri dari 309 sampel pasien dengan 16 atribut. Atribut tersebut mencakup seperti jenis kelamin dan usia, serta 13 fitur gejala klinis dan kebiasaan hidup (merokok, kecemasan, dan nyeri dada) yang direkam dalam format kategorikal biner. Variabel target dalam penelitian ini adalah diagnosis kanker paru-paru dengan klasifikasi "Yes" atau "No".

Gambar 1 menunjukkan dataset survei penyakit kanker paru-paru mencakup 16 kolom.

No.	Atribut	Tipe Data	Keterangan
1	GENDER	Kategorikal	Jenis Kelamin (Pria/Wanita)
2	AGE	Numerik	Usia pasien
3-15	Fitur Biner	Kategorikal Biner	13 fitur gejala dan kebiasaan (misalnya: SMOKING, ANXIETY, CHEST PAIN), dengan nilai **1** atau **2**
16	LUNG_CANCER	Target	Variabel target, diagnosis kanker paru-paru (**YES** atau **NO**)

Gambar 1. Dataset

Tahapan penelitian yang digunakan oleh peneliti terdiri dari berbagai tahapan untuk menggali, menganalisis, dan mengeksplorasi data



Gambar 2. Tahapan Penelitian

Implementasi *machine learning* yang digunakan dalam penelitian ini dengan urutan :

1. Lung Cancer Survey Dataset: Tahap awal dimulai dengan penggunaan **Lung Cancer Survey Dataset** yang diperoleh dari repositori publik.
2. Preprocessing: Tahap awal pembersihan dan transformasi data :
   - Features Extraction & Data Cleaning: Membersihkan data dari anomali atau redudansi.
   - Missing Values Handling: Menangani data yang hilang agar tidak mengganggu performa model.
   - Categorical Variables Transformation: Mengonversi variabel kategori menjadi format numerik agar dapat diproses oleh algoritma machine learning.
3. Pemisahan data untuk menjaga integritas pengujian, data dibagi menggunakan strategi Stratified Split dengan proporsi (80% Training, 20% Testing) dengan Stratified Split: Data dibagi menjadi set pelatihan dan pengujian dengan menjaga proporsi kelas target.
4. Over Sampling the Training dataset - SMOTE: Pada fase ini, diterapkan teknik SMOTE (Synthetic Minority Over-sampling Technique) khusus pada data pelatihan. Langkah ini bertujuan untuk mengatasi masalah ketidakseimbangan kelas (class imbalance) yang sangat ekstrem, di mana distribusi kelas diseimbangkan dari semula 31:216 menjadi 216:216 untuk memastikan model tidak bias terhadap kelas mayoritas.Data Pelatihan Seimbang: Hasil dari SMOTE.
5. Penerapan Teknik Ensemble Learning: Empat model ensemble yang akan digunakan (**XGBoost, LightGBM, AdaBoost, Bagging)**
6. Split models into K=10 folds: Menunjukkan proses pembagian data untuk k-fold cross-validation.
7. Loop for K=10 times: Proses iterasi validasi silang, meliputi : Choose train and test folds, Fit models on train folds, Evaluate models on test fold.
8. Get models scores: Mengumpulkan metrik kinerja dari setiap iterasi.
9. Mean Performance Output: Rata-rata metrik kinerja (Akurasi, Presisi, Rekal, AUC, dan F1-Score jika dihitung) dari 10 kali validasi silang.
10. Performance Comparison: Tahap perbandingan hasil metrik antar model.
11. Identifikasi Model Ensemble Terbaik: Kesimpulan akhir dari proses, mengidentifikasi model dengan kinerja superior.

III. HASIL DAN PEMBAHASAN

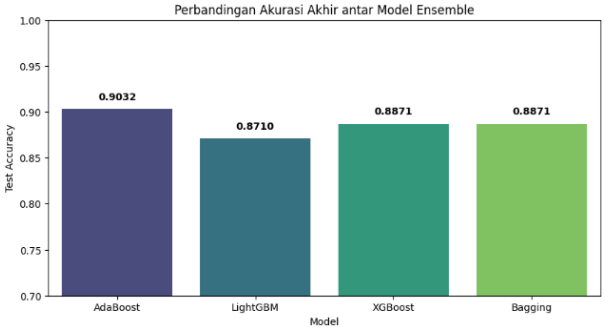
Penelitian yang sudah dilakukan akan menyajikan hasil dari analisis komparatif model Ensemble Learning dalam memprediksi risiko kanker paru-paru. Hasil pengujian kinerja model pada testing set yang tidak terlihat sebelumnya didiskusikan secara mendalam, dengan fokus pada metrik utama: Akurasi, Presisi, Rekal, dan AUC.

1. Hasil Validasi Silang dan Evaluasi Akhir

Pada table 1 terdapat hasil pengujian yang menunjukkan variasi kinerja di antara empat algoritma ensemble learning yang dieksplorasi.

Tabel 1. Hasil evaluasi 4 algoritma

Model	CV Akurasi ($\bar{x}$)	Test Akurasi	Test Presisi	Test Recal	Test AUC
AdaBoost	0.9628	**0.9032**	0.9444	0.9444	0.9340
LightGBM	0.9559	0.8709	0.9423	0.9074	0.9328
XGBoost	0.9559	0.8870	0.9433	0.9259	0.9282
Bagging	0.9467	0.8870	0.9795	0.9795	0.9259



Gambar 3. Perbandingan Hasil Akurasi

Berdasarkan hasil perbandingan model ensemble yang dilakukan, algoritma AdaBoost menunjukkan performa yang paling superior dibandingkan dengan model lainnya. Hal ini dibuktikan dengan perolehan nilai akurasi pengujian (Test Accuracy)

tertinggi sebesar 90,32% serta nilai Test AUC sebesar 0,9340, yang mengindikasikan kemampuan diskriminasi model yang sangat baik dalam mengklasifikasikan data.

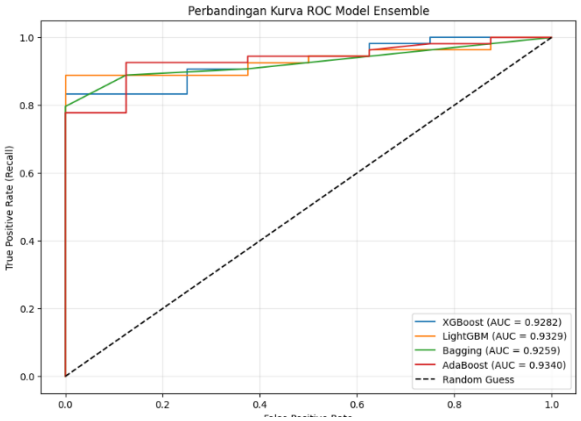
Selain itu, AdaBoost menunjukkan stabilitas yang tinggi dengan keseimbangan antara nilai Precision dan Recall yang masing-masing mencapai 0,9444. Di sisi lain, model XGBoost dan Bagging memberikan hasil yang kompetitif dengan nilai akurasi pengujian yang identik, yaitu sebesar 88,71%, di mana model Bagging menonjol dengan tingkat presisi tertinggi mencapai 0,9796. Sementara itu, model LightGBM mencatatkan akurasi terendah di angka 87,10%. Meskipun terdapat sedikit penurunan dari nilai akurasi validasi silang (CV Accuracy) ke akurasi pengujian pada seluruh model, secara keseluruhan performa model ensemble dalam penelitian ini tetap tergolong sangat kuat karena konsisten berada di atas ambang 87%, menunjukkan generalisasi model yang baik terhadap data baru.

2. Presisi dan Recal.

Dapat di lihat dari table.1 hasil penelitian menunjukkan bahwa AdaBoost memiliki keseimbangan performa yang optimal dengan nilai presisi dan recall. Presisi dan Recall adalah dua metrik evaluasi yang memberikan pemahaman lebih dalam tentang kualitas prediksi model dibandingkan sekadar akurasi (Cahyanti et al., 2020).
Nilai presisi dan recall yang identik sebesar 0,9444, membuktikan bahwa kemampuan model dalam meminimalisir kesalahan prediksi (false positive) sekaligus mencegah luputnya deteksi (false negative) secara merata.

3. Analisis ROC

Analisis visual melalui Kurva ROC (Receiver Operating Characteristic) adalah grafik yang digunakan untuk mengevaluasi kinerja model klasifikasi biner pada berbagai ambang batas (threshold) Keputusan (Surono & Pusparini, 2020). Kurva ini memberikan gambaran visual tentang seberapa baik model dapat membedakan antara kelas positif (terjangkit kanker) dan kelas negatif (tidak terjangkit kanker).



Gambar 4. Kurve ROC

Seperti yang diilustrasikan oleh Gambar 4.1, Evaluasi kinerja model ensemble menunjukkan bahwa algoritma AdaBoost memberikan performa paling superior dibandingkan model lainnya, dengan perolehan akurasi pengujian (Test Accuracy) tertinggi sebesar 90,32% dan skor AUC mencapai 0,9340. Kekuatan utama AdaBoost terletak pada keseimbangan performanya, di mana nilai presisi dan recall yang identik sebesar 0,9444 membuktikan kemampuan model yang sangat stabil dalam meminimalisir kesalahan klasifikasi (false positive) sekaligus mencegah luputnya deteksi (false negative).

Meskipun model Bagging mencatatkan nilai presisi tertinggi (0,9796), model tersebut memiliki tingkat recall terendah di angka 0,8888. Analisis melalui kurva ROC mempertegas kualitas seluruh model ensemble yang diuji, di mana setiap model memiliki kemampuan diskriminasi yang sangat baik dengan nilai AUC di atas 0,92. Optimalnya hasil ini juga dipengaruhi oleh implementasi teknik SMOTE yang berhasil menyeimbangkan distribusi kelas dari semula 31:216 menjadi 216:216, sehingga model dapat menggeneralisasi data dengan lebih adil dan akurat. Secara keseluruhan, hasil pengujian ini mengonfirmasi bahwa pendekatan ensemble, khususnya AdaBoost, sangat efektif untuk diimplementasikan pada dataset *lung cancer* ini.

IV. KESIMPULAN

Penelitian ini menyimpulkan bahwa penerapan metode ensemble learning, khususnya algoritma AdaBoost, memberikan performa yang paling optimal dalam tugas klasifikasi pada dataset yang digunakan. Hal ini dibuktikan dengan capaian akurasi pengujian tertinggi sebesar 90,32% serta nilai AUC sebesar 0,9340 yang menunjukkan

kemampuan diskriminasi kelas yang sangat kuat dibandingkan model lainnya. Keunggulan utama AdaBoost terletak pada stabilitas prediksinya yang ditunjukkan oleh nilai presisi dan recall yang identik pada angka 0,9444, menandakan bahwa model mampu memberikan hasil yang sangat akurat sekaligus menyeluruh dalam mendeteksi data positif. Selain itu, efektivitas seluruh model ensemble yang diuji juga didukung oleh penggunaan teknik SMOTE yang berhasil menyeimbangkan distribusi kelas dari 31:216 menjadi 216:216, sehingga semua model mampu mencapai nilai AUC di atas 0,92. Temuan ini membuktikan bahwa kombinasi penyeimbangan data dan algoritma berbasis boosting merupakan solusi yang handal untuk menghasilkan model klasifikasi dengan tingkat generalisasi yang tinggi.

V. REFERENSI

Asiyah, S. N., Rachman Al Syaiba, F., Nailun Nabilah, S., Marsha Finanda, Z., & Fadhilah Mudrik, P. Z. (2025). Peran Gizi Dalam Menangani Kanker Paru-Paru : Literature Review. *Media Gizi Pangan*, *32*(1), 1–8. https://doi.org/10.32382/mgp.v32i1.691

Cahyanti, D., Rahmayani, A., & Husniar, S. A. (2020). Biasanya kanker payudara jinak ditandai dengan berbentuk benjolan kecil bulat, dan lembut. Kanker payudara dalam tingkat jinak biasanya akan mempunyai keadaan dan pertumbuhan yang tidak bersifat kanker. Kanker ini bisa terdeteksi tetapi tidak akan menjala. *Indonesian Journal of Data and Science*, *1*(2), 39–43.

Cendani, L. M., & Wibowo, A. (2022). Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes. *Jurnal Masyarakat Informatika*, *13*(1), 33–44. https://doi.org/10.14710/jmasif.13.1.42912

Rejeki, M., Pratiwi, E. N., Administrasi, S., Sakit, R., Fakultas, /, Kesehatan, I., Kusuma, U., Surakarta, H., & Kebidanan, S. (2020). Diagnosis dan Prognosis Kanker Paru, Probabilitas Metastasis dan Upaya Prevensinya. *The 12th University Research Colloqium 2020*, 1–6.

Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, *73*(1), 17–48. https://doi.org/10.3322/caac.21763

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660

Surono, G., & Pusparini, N. N. (2020). Journal of technology information. *Jurnal Of Technology Information*, *5*(2), 99–104.

Syahwaluddin, R., & Alita, D. (2024). Penerapan Oversampling Pada Klasifikasi Ujaran Kebencian Menggunakan Bidirectional Encoder Representations from Transformers. *The Indonesian Journal of Computer Science*, *13*(4), 6615–6625. https://doi.org/10.33022/ijcs.v13i4.4295

Tarumingkeng, I. R. C. (2025). Big Data dan AI: Sinergi dalam Era Digital. *Rudyct.Com*. https://rudyct.com/ab/BigData.dan.AI-Sinergi.dalam.Era.Digital.pdf

Wardhana, R. G., Wang, G., & Farida Sibuea. (2023). PENERAPAN MACHINE LEARNING DALAM PREDIKSI TINGKAT KASUS PENYAKIT DI INDONESIA Master of Information Systems Management Bina Nusantara University Abstraksi Keywords : Pendahuluan Tinjauan Pustaka Metode Penelitian Proses implementasi machine learning dapat. *Journal of Information System Management (JOISM)*, *5*(1), 40–45.

Yulian, T., & Susanto, E. R. (2025). Analisis Perbandingan Teknik Oversampling dan SMOTEENN pada Algoritma Machine Learning untuk Prediksi Penyakit Kanker Payudara Comparative Analysis of Oversampling and SMOTEENN Techniques in Machine Learning Algorithms for Breast Cancer Prediction. *Sistemasi: Jurnal Sistem Informasi*, *14*, 1318–1331. http://sistemasi.ftik.unisi.ac.id