

Analisis Kinerja Decision Tree dan Random Forest Menggunakan Dataset Breast Cancer

Uus Firdaus¹, Ananda Alfiah², Lorina Mohdo³

^{1,2,3}Universitas Djuanda, Indonesia

¹Uus.firdaus@unida.ac.id, ²i.2210378@unida.ac.id, ³lorin.mohdo@gmail.com



***Uus Firdaus**

Histori Artikel:

Submit: 2026-01-24

Diterima: 2026-02-06

Dipublikasikan: 2026-02-08

Kata Kunci:

Machine Learning; Data Mining;
Decision Tree; Random Forest;
Kanker Payudara

ABSTRAK

Kanker payudara merupakan salah satu penyakit dengan tingkat mortalitas yang tinggi pada perempuan di seluruh dunia, sehingga deteksi dini menjadi faktor krusial dalam meningkatkan peluang keberhasilan pengobatan dan harapan hidup pasien. Perkembangan teknologi komputasi, khususnya machine learning, membuka peluang pemanfaatan data medis sebagai dasar pengambilan keputusan dalam proses diagnosis. Penelitian ini bertujuan untuk menganalisis dan membandingkan kinerja algoritma Decision Tree dan Random Forest dalam melakukan klasifikasi kanker payudara menggunakan Wisconsin Breast Cancer Dataset. Dataset yang digunakan terdiri dari 569 data dengan 30 atribut numerik yang merepresentasikan karakteristik sel kanker, serta satu label kelas yang menunjukkan kategori kanker jinak (benign) atau ganas (malignant). Tahapan penelitian meliputi prapemrosesan data, pembagian data menjadi data latih dan data uji dengan rasio 80:20, penerapan algoritma klasifikasi, serta evaluasi kinerja model menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil pengujian menunjukkan bahwa algoritma Random Forest menghasilkan kinerja yang lebih unggul dibandingkan Decision Tree. Random Forest memperoleh akurasi sebesar 98,68% pada data latih dan 95,61% pada data uji, sedangkan Decision Tree menghasilkan akurasi sebesar 96,92% pada data latih dan 91,23% pada data uji. Perbedaan ini menunjukkan bahwa Random Forest memiliki kemampuan generalisasi yang lebih baik serta lebih tahan terhadap overfitting. Temuan penelitian ini mengindikasikan bahwa Random Forest lebih efektif digunakan untuk klasifikasi kanker payudara berbasis data dibandingkan Decision Tree. Oleh karena itu, algoritma Random Forest direkomendasikan sebagai metode yang lebih andal untuk mendukung sistem pendukung keputusan dalam deteksi dini kanker payudara.

Jurnal Pendidikan Sains dan Komputer is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

LATAR BELAKANG

Kanker payudara merupakan salah satu penyakit kanker yang paling banyak menyerang perempuan dan masih menjadi penyebab utama kematian akibat kanker di berbagai belahan dunia. Tingginya angka kejadian dan mortalitas menunjukkan bahwa penanganan kanker payudara masih memerlukan perhatian serius, khususnya dalam aspek deteksi dini. Deteksi pada tahap awal sangat berpengaruh terhadap keberhasilan terapi dan peningkatan harapan hidup pasien.

Perkembangan teknologi informasi dan komputasi telah membuka peluang baru dalam bidang kesehatan, terutama dalam pemanfaatan data medis untuk mendukung proses diagnosis. Salah satu pendekatan yang banyak digunakan adalah machine learning, yaitu metode komputasi yang memungkinkan sistem mempelajari pola dari data historis dan menghasilkan model prediksi secara otomatis. Dalam konteks kanker payudara, machine learning dimanfaatkan untuk membantu mengklasifikasikan kondisi pasien berdasarkan karakteristik sel kanker yang diamati.

Penelitian terdahulu menunjukkan bahwa penerapan algoritma machine learning pada data kanker payudara mampu memberikan hasil klasifikasi yang cukup baik. (Hiba, Mousannif, Moatassime, & Noel, 2016) membuktikan bahwa penggunaan algoritma machine learning pada dataset Wisconsin Breast Cancer menghasilkan tingkat akurasi yang tinggi dalam membedakan kanker jinak dan ganas. Namun demikian, pemilihan algoritma yang tepat masih menjadi tantangan karena setiap metode memiliki karakteristik dan keterbatasan masing-masing.

Algoritma Decision Tree dikenal memiliki kelebihan dalam hal kemudahan interpretasi model, sehingga hasil klasifikasinya relatif mudah dipahami. Akan tetapi, algoritma ini cenderung mengalami overfitting ketika dihadapkan pada data dengan kompleksitas tinggi apabila tidak dilakukan pengaturan parameter yang optimal (Liu, Wang, Xiao, & Liang, 2016). Untuk mengatasi permasalahan tersebut, (Breiman, 2001) memperkenalkan Random Forest sebagai metode ensemble yang menggabungkan banyak pohon keputusan sehingga mampu meningkatkan akurasi dan kestabilan model.

Meskipun sejumlah penelitian telah membahas penerapan machine learning dalam klasifikasi kanker payudara, sebagian besar penelitian masih berfokus pada penggunaan satu algoritma atau belum menyajikan analisis perbandingan secara mendalam. Oleh karena itu, penelitian ini bertujuan untuk melakukan perbandingan kinerja antara algoritma Decision Tree dan Random Forest menggunakan dataset yang sama serta metrik evaluasi yang seragam. Dengan pendekatan tersebut, penelitian ini diharapkan dapat menunjukkan posisi dan keunggulan masing-masing algoritma serta memberikan rekomendasi metode yang lebih andal untuk mendukung sistem diagnosis kanker payudara berbasis data.

STUDI LITERATUR

Penelitian terkait klasifikasi kanker payudara menggunakan pendekatan machine learning telah banyak dilakukan dalam beberapa tahun terakhir. Pemanfaatan data medis sebagai dasar pengambilan keputusan dinilai mampu meningkatkan akurasi diagnosis dibandingkan metode konvensional. Salah satu dataset yang paling sering digunakan dalam penelitian ini adalah Wisconsin Breast Cancer Dataset karena memiliki atribut numerik yang merepresentasikan karakteristik sel kanker secara detail.

(Hiba, Mousannif, Moatassime, & Noel, 2016) melakukan penelitian dengan menerapkan beberapa algoritma machine learning pada dataset Wisconsin Breast Cancer dan menunjukkan bahwa metode berbasis ensemble memiliki tingkat akurasi yang lebih tinggi dibandingkan algoritma tunggal. Hasil penelitian tersebut mengindikasikan bahwa kombinasi beberapa model dapat meningkatkan kemampuan generalisasi dalam proses klasifikasi kanker payudara.

Decision Tree merupakan salah satu algoritma yang banyak digunakan karena memiliki struktur yang sederhana dan mudah diinterpretasikan. Namun, algoritma ini memiliki kelemahan utama berupa kecenderungan terhadap overfitting, terutama ketika digunakan pada data dengan kompleksitas tinggi tanpa pengaturan parameter yang optimal (Quinlan, 2014). Untuk mengatasi permasalahan tersebut, Breiman (2001) mengembangkan Random Forest sebagai metode ensemble yang menggabungkan banyak pohon keputusan sehingga mampu menghasilkan model yang lebih stabil dan akurat.

Berdasarkan studi literatur tersebut, dapat disimpulkan bahwa perbandingan antara algoritma Decision Tree dan Random Forest masih relevan untuk dilakukan, khususnya dalam konteks klasifikasi kanker payudara. Oleh karena itu, penelitian ini difokuskan pada analisis komparatif kedua algoritma tersebut guna mengetahui perbedaan kinerja dan keandalannya dalam menangani dataset kanker payudara.

METODE

Metode penelitian yang digunakan dalam penelitian ini terdiri dari beberapa tahapan sebagai berikut:

1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah *Wisconsin Breast Cancer Dataset* yang diperoleh dari

UCI Machine Learning Repository. Dataset ini terdiri dari 569 data dengan 30 atribut numerik yang merepresentasikan karakteristik inti sel kanker, serta satu label kelas yang menunjukkan kategori kanker jinak (benign) atau ganas (malignant).

2. Prapemrosesan Data

Tahap prapemrosesan meliputi pemeriksaan data kosong (missing value), normalisasi data jika diperlukan, serta pengkodean label kelas menjadi bentuk numerik agar dapat diproses oleh algoritma *machine learning*.

3. Pembagian Data

Dataset dibagi menjadi dua bagian, yaitu data latih dan data uji, dengan perbandingan 80% untuk data latih dan 20% untuk data uji. Data latih digunakan untuk membangun model klasifikasi, sedangkan data uji digunakan untuk mengevaluasi kinerja model terhadap data yang belum pernah dilihat sebelumnya.

4. Penerapan Algoritma

Pada tahap ini, algoritma *Decision Tree* dan *Random Forest* diterapkan pada data latih. Parameter model disesuaikan untuk memperoleh kinerja yang optimal.

5. Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan data uji dengan metrik akurasi, presisi, *recall*, dan *F1-score*. Metrik tersebut digunakan untuk memberikan gambaran menyeluruh mengenai kemampuan model dalam melakukan klasifikasi kanker payudara. Hasil evaluasi dari kedua algoritma kemudian dibandingkan untuk menentukan metode yang memiliki kinerja terbaik.

HASIL DAN PEMBAHASAN

Berdasarkan hasil pengujian yang dilakukan menggunakan Google Colab, diperoleh nilai akurasi untuk masing-masing algoritma pada data latih (training) dan data uji (testing). Hasil pengujian menunjukkan bahwa algoritma *Random Forest* memiliki nilai akurasi sebesar 98,68% pada data latih dan 95,61% pada data uji. Sementara itu, algoritma *Decision Tree* menghasilkan akurasi sebesar 96,92% pada data latih dan 91,23% pada data uji. Hasil Uji ditampilkan pada Gambar 1 dan Gambar 2.

```

Accuracy score train: 96.92%
Accuracy score test: 91.23%
Classification Report

```

	precision	recall	f1-score	support
0	0.92	0.94	0.93	72
1	0.90	0.86	0.88	42
accuracy			0.91	114
macro avg	0.91	0.90	0.90	114
weighted avg	0.91	0.91	0.91	114

Gambar 1. Akurasi Random Forest

```

Accuracy score train: 98.68%
Accuracy score test: 95.61%
Classification Report

```

	precision	recall	f1-score	support
0	0.94	1.00	0.97	72
1	1.00	0.88	0.94	42
accuracy			0.96	114
macro avg	0.97	0.94	0.95	114
weighted avg	0.96	0.96	0.96	114

Gambar 2. Akurasi Decision Tree

Nilai akurasi yang tinggi pada data latih menunjukkan bahwa kedua algoritma mampu mempelajari pola data dengan baik. Namun, terdapat perbedaan yang cukup signifikan pada nilai akurasi data uji. *Random Forest* menunjukkan penurunan akurasi yang relatif kecil dari data latih ke data uji, yang mengindikasikan bahwa model memiliki kemampuan generalisasi yang baik. Hal ini sesuai dengan teori yang dikemukakan oleh (Breinman, 2001) bahwa *Random Forest* mampu mengurangi risiko overfitting melalui mekanisme ensemble dari banyak pohon keputusan.

Sebaliknya, *Decision Tree* mengalami penurunan akurasi yang lebih besar pada data uji. Perbedaan antara akurasi data latih dan data uji ini mengindikasikan bahwa model *Decision Tree* cenderung mengalami overfitting, yaitu model terlalu menyesuaikan diri dengan data latih sehingga performanya menurun ketika diuji pada data baru. Kondisi ini sejalan dengan penelitian (Quinlan, 2014) yang menyatakan bahwa *Decision Tree* rentan terhadap overfitting apabila tidak dilakukan pengaturan parameter atau pruning yang optimal.

Pada Gambar 3 ditampilkan tabel hasil perbandingan antara *Random Forest* dan *Decision Tree*.

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	91.23%	0.91	0.91	0.91
Random Forest	95.61%	0.96	0.96	0.96

Gambar 3. Tabel Hasil Perbandingan

Berdasarkan hasil perbandingan kinerja model yang ditampilkan pada Gambar 3, *Random Forest* menunjukkan performa yang lebih baik dibandingkan *Decision Tree* pada seluruh metrik evaluasi. Perbedaan ini tidak hanya mencerminkan keunggulan numerik, tetapi juga berkaitan erat dengan karakteristik *dataset* serta mekanisme pembentukan model pada masing-masing algoritma. *Dataset Breast Cancer* memiliki sejumlah fitur numerik dengan hubungan yang kompleks, sehingga menuntut model yang mampu menangkap pola data secara lebih stabil.

Decision Tree membangun model klasifikasi menggunakan satu struktur pohon keputusan, sehingga sangat sensitif terhadap variasi data latih. Pada *dataset* dengan jumlah fitur yang relatif banyak, model ini cenderung menyesuaikan diri secara berlebihan terhadap data latih, yang menyebabkan penurunan kinerja pada data uji. Kondisi tersebut tercermin dari selisih akurasi antara data latih dan data uji yang lebih besar, yang mengindikasikan terjadinya *overfitting*.

Sebaliknya, *Random Forest* mengombinasikan banyak pohon keputusan yang dibangun dari subset data dan fitur yang berbeda. Pendekatan ini memungkinkan model untuk mengurangi variansi dan pengaruh noise pada *dataset*. Proses agregasi prediksi dari banyak pohon keputusan menghasilkan model yang lebih robust dan memiliki kemampuan generalisasi yang lebih baik. Hal ini menjelaskan mengapa *Random Forest* mampu mencapai nilai presisi, *recall*, dan *F1-score* yang lebih tinggi dibandingkan *Decision Tree*.

Selain itu, kemampuan *Random Forest* dalam melakukan pemilihan fitur secara acak pada setiap pohon turut berkontribusi terhadap peningkatan kinerja model. Mekanisme ini memungkinkan *Random Forest* untuk menangani variasi skala dan korelasi antar fitur dengan lebih efektif. Dengan demikian, hasil penelitian ini menunjukkan bahwa *Random Forest* lebih sesuai digunakan untuk klasifikasi kanker payudara pada dataset dengan kompleksitas tinggi, sedangkan *Decision Tree* lebih rentan terhadap *overfitting* apabila tidak disertai dengan pengaturan parameter yang optimal.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma *Random Forest* menunjukkan kinerja yang lebih baik dibandingkan *Decision Tree* dalam klasifikasi dataset *Breast Cancer*. Keunggulan *Random Forest* terlihat dari nilai akurasi, presisi, *recall*, dan *F1-score* yang lebih tinggi serta selisih kinerja yang lebih kecil antara data latih dan data uji, yang mengindikasikan kemampuan generalisasi model yang lebih baik.

Meskipun demikian, penelitian ini memiliki beberapa keterbatasan. Dataset yang digunakan berasal dari satu sumber, yaitu *Wisconsin Breast Cancer Dataset*, sehingga variasi data yang merepresentasikan kondisi klinis yang lebih beragam belum sepenuhnya tercakup. Selain itu, penelitian ini belum melakukan optimasi parameter secara mendalam maupun penerapan teknik validasi silang (*cross-validation*) yang lebih ekstensif, sehingga potensi peningkatan kinerja model masih terbuka untuk diteliti lebih lanjut.

Dalam konteks penerapan nyata, hasil penelitian ini menunjukkan bahwa *Random Forest* memiliki potensi untuk digunakan sebagai komponen dalam sistem pendukung keputusan medis, khususnya dalam membantu proses klasifikasi awal kanker payudara. Model ini dapat dimanfaatkan sebagai alat bantu bagi tenaga medis untuk memberikan rekomendasi awal berdasarkan data pasien, namun tetap harus digunakan sebagai sistem pendukung, bukan sebagai pengganti keputusan klinis. Oleh karena itu, integrasi model *machine learning* ke dalam sistem medis perlu disertai dengan evaluasi lanjutan, validasi klinis, serta pertimbangan aspek etika dan keamanan data pasien.

Sebagai arah penelitian selanjutnya, disarankan untuk menggunakan dataset yang lebih beragam, menerapkan optimasi parameter dan teknik validasi yang lebih komprehensif, serta menguji integrasi model ke dalam prototipe sistem pendukung keputusan medis guna menilai efektivitasnya dalam lingkungan klinis nyata.

REFERENSI

- Breiman, L. (2001). *Random Forests*. *Statistics Department, University of California, Berkeley, CA 94720*.
- Breiman, L. (2001). *Random Forests*.
- Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Dermatologist-level classification of skin cancer with deep neural networks*.
- Gayathri, B., Sumathi, C., & Santhanam, T. (2013). BREAST CANCER DIAGNOSIS USING MACHINE. *International Journal of Distributed and Parallel Systems*.
- Hiba, A., Mousannif, H., Moatassime, H., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*.
- Liu, C., Wang, J., Xiao, D., & Liang, Q. (2016). Forecasting S&P 500 Stock Index Using Statistical Learning Models.



Open Journal of Statistics.

Quinlan, J. (2014). *programs for machine learning*.

