Analisis Komparatif Algoritma Machine Learning untuk Prediksi Kekambuhan Kanker Payudara Berdasarkan Karakteristik Tumor

Comparative Analysis of Machine Learning Algorithms for Breast Cancer Recurrence Prediction Based on Tumor Characteristics

Chairunnisa Desti Arzety*[1], Vanya Dwi Nabila[2], Aprilia Herawati[3], Allsela Meiriza[4], Ken Ditha Tania[5]

Program Studi Sistem Informasi, Universitas Sriwijaya, Jalan Palembang - Prabumulih Km. 32 Indralaya, Ogan Ilir, Telp. (0711) 580069, 580225, 580169, 580275, Fax. (0711) 580644

E-mail : 09031282328032@student.unsri.ac.id[1], 09031382328119@student.unsri.ac.id[2], 09031382328133@student.unsri.ac.id[3], allsela_meiriza@yahoo.com[4], kenya.tania@gmail.com[5]*

**Corresponding author*

Received 10 April 2026; Revised 18 April 2026; Accepted 24 April 2026

Abstrak - Kanker payudara merupakan salah satu penyebab kematian tertinggi pada perempuan di dunia, di mana tantangan utamanya terletak pada risiko kekambuhan (*recurrence*). Penelitian ini bertujuan untuk membangun model prediksi kekambuhan kanker payudara dengan membandingkan tiga algoritma *machine learning*, yaitu Logistic Regression, Decision Tree, dan Random Forest, berdasarkan karakteristik tumor dari dataset METABRIC. Tahapan penelitian meliputi pra-pemrosesan data, seleksi fitur klinis, dan pembagian data dengan rasio 80:20. Hasil evaluasi menunjukkan bahwa Logistic Regressio*n* memiliki performa terbaik dalam hal akurasi (0,661) dan ROC-AUC (0,689), sementara Random Forest menunjukkan keunggulan pada metrik *recall* (0,544) yang krusial untuk deteksi pasien berisiko. Analisis *feature importance* mengidentifikasi bahwa jumlah mutasi genetik (*Mutation Count*), *Nottingham Prognostic Index* (NPI), dan ukuran tumor merupakan faktor paling dominan dalam memprediksi kekambuhan. Penelitian ini menyimpulkan bahwa karakteristik biologis tumor memiliki pengaruh signifikan terhadap risiko kekambuhan dan penggunaan *machine learning* berpotensi besar menjadi sistem pendukung keputusan klinis untuk stratifikasi risiko pasien secara objektif.

Kata kunci - Kanker Payudara, Kekambuhan, *Machine Learning*, METABRIC, Karakteristik Tumor

Abstract - *Breast cancer is one of the leading causes of death among women worldwide, with recurrence remaining a major clinical challenge. This study aims to develop a breast cancer recurrence prediction model by comparing three machine learning algorithms, namely Logistic Regression, Decision Tree, and Random Forest, based on tumor characteristics derived from the METABRIC dataset. The research stages include data preprocessing, clinical feature selection, and data splitting using an 80:20 ratio. Evaluation results indicate that Logistic Regression achieved the best performance in terms of accuracy (0.661) and ROC-AUC (0.689), while Random Forest demonstrated superior recall (0.544), which is crucial for identifying high-risk patients. Feature importance analysis revealed that Mutation Count, Nottingham Prognostic Index (NPI), and Tumor Size were the most dominant predictors of recurrence. The findings suggest that biological tumor characteristics have a significant influence on recurrence risk, and the application of machine learning shows strong potential as a clinical decision support system for objective patient risk stratification.*

Keywords - *Breast Cancer, Recurrence, Machine Learning, METABRIC, Tumor Characteristics*

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15976

1. PENDAHULUAN

Kanker payudara merupakan salah satu jenis kanker dengan angka kejadian tertinggi pada perempuan secara global. Meskipun perkembangan terapi telah meningkatkan angka kelangsungan hidup pasien, risiko kekambuhan (*recurrence*) tetap menjadi tantangan utama dalam manajemen klinis. Kekambuhan dapat terjadi dalam bentuk lokal maupun metastasis jauh dan berhubungan dengan karakteristik biologis tumor, faktor molekuler, serta respons terhadap terapi. Studi melalui *systematic review* dan *meta-analysis* menunjukkan bahwa prediksi kekambuhan menjadi fokus penting dalam penelitian onkologi karena berpengaruh terhadap strategi pemantauan jangka panjang pasien [1]. Hal serupa ditegaskan dalam kajian evolusi model prediksi risiko kekambuhan oleh penelitian lain yang menyoroti pergeseran dari pendekatan statistik tradisional menuju metode berbasis kecerdasan buatan [2].

Pemanfaatan *machine learning* dalam prediksi kekambuhan kanker payudara telah berkembang pesat dalam beberapa tahun terakhir. Sebuah studi komparatif [3] menguji berbagai algoritma *machine learning* dan menemukan bahwa model *ensemble* memiliki performa yang lebih stabil dibandingkan model tunggal. Studi lain [4] juga melakukan analisis komparatif beberapa algoritma dan menunjukkan bahwa Random Forest dan model *boosting* cenderung memberikan keseimbangan yang lebih baik antara akurasi dan kemampuan deteksi kasus positif. Selain itu, penelitian yang dilakukan oleh [5] mengembangkan model berbasis LASSO dan *ensemble learning* untuk mendukung keputusan terapi personalisasi dan menunjukkan peningkatan performa prediksi dibandingkan pendekatan konvensional.

Pendekatan *survival analysis* yang dikombinasikan dengan *machine learning* juga telah digunakan untuk memprediksi waktu terjadinya kekambuhan. Sebagai contoh, sebuah penelitian [6] menerapkan teknik *ensemble survival* pada *dataset* METABRIC dan melaporkan performa yang kompetitif dalam memprediksi kelangsungan hidup pasien. Penelitian lain yang dipublikasikan dalam [7] mengintegrasikan *survival modeling* dengan algoritma *machine learning* untuk memprediksi kekambuhan dan metastasis secara simultan, menunjukkan bahwa pendekatan *hibrida* mampu meningkatkan kemampuan diskriminasi model.

Beberapa penelitian menekankan pentingnya integrasi berbagai jenis data dalam meningkatkan akurasi prediksi. Studi oleh [8] menunjukkan bahwa kombinasi karakteristik klinis dan fitur radiomik dari citra medis dapat meningkatkan kemampuan model dalam mendeteksi risiko kekambuhan. Sementara itu, penelitian yang dipublikasikan di [9] membandingkan penggunaan data terstruktur dan tidak terstruktur dari rekam medis elektronik dan menemukan bahwa integrasi keduanya menghasilkan performa yang lebih baik dibandingkan penggunaan satu jenis data saja. Penelitian lain pada [10] juga mengembangkan model *ensemble* berbasis fitur *histopatologi* untuk memprediksi *relapse* dan metastasis dengan hasil yang signifikan.

Dalam konteks komparasi algoritma, beberapa studi telah membandingkan Logistic Regression, Decision Tree, Random Forest, serta metode *boosting*. Studi yang dipublikasikan dalam [11] membandingkan algoritma *machine learning* dan *deep learning* untuk prediksi kekambuhan satu tahun dan menunjukkan variasi performa yang dipengaruhi oleh karakteristik *dataset*. Studi lain [12] menunjukkan bahwa pemilihan variabel prediktor yang tepat, termasuk kombinasi klinis dan biomarker, secara signifikan mempengaruhi peningkatan akurasi model *machine learning* dalam memprediksi waktu kekambuhan kanker payudara. Selain itu, studi yang diterbitkan dalam [13] mengembangkan model *machine learning* untuk memprediksi *distant recurrence* menggunakan data *real-world* dan menegaskan bahwa model *ensemble* memiliki kemampuan generalisasi yang lebih baik.

Aspek interpretabilitas model juga menjadi perhatian penting dalam penerapan *machine learning* di bidang kesehatan. Penelitian terbaru mengenai *explainable AI* dalam prediksi risiko kanker [14] menekankan bahwa transparansi model diperlukan untuk meningkatkan kepercayaan klinisi terhadap sistem prediksi. Sejalan dengan hal tersebut, penelitian lain [15] juga menunjukkan bahwa metode interpretasi seperti SHAP dapat digunakan untuk mengidentifikasi fitur dominan seperti ukuran tumor, stadium klinis, jumlah metastasis kelenjar getah bening, dan

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15976

usia yang secara signifikan berkontribusi terhadap prediksi risiko kekambuhan kanker payudara, sehingga model menjadi lebih transparan dan klinis dapat dipahami.

Meskipun berbagai penelitian telah menunjukkan potensi *machine learning* dalam memprediksi kekambuhan kanker payudara, masih terdapat variasi hasil performa antar studi. Perbedaan tersebut dapat disebabkan oleh karakteristik *dataset*, jumlah sampel, metode seleksi fitur, serta teknik evaluasi yang digunakan. Beberapa penelitian melaporkan nilai AUC yang tinggi di atas 0,70, sementara penelitian lain menunjukkan performa moderat. Lebih jauh, sebagian besar studi yang ada menggunakan kombinasi fitur klinis, molekuler, dan genomik yang kompleks, sehingga belum diketahui sejauh mana karakteristik tumor saja, tanpa variabel tambahan lain, dapat dijadikan dasar prediksi yang andal. Hal ini menunjukkan bahwa terdapat kesenjangan penelitian yang perlu diatasi, yakni perlunya evaluasi komparatif yang terfokus pada karakteristik tumor sebagai prediktor utama untuk menentukan model yang paling sesuai.

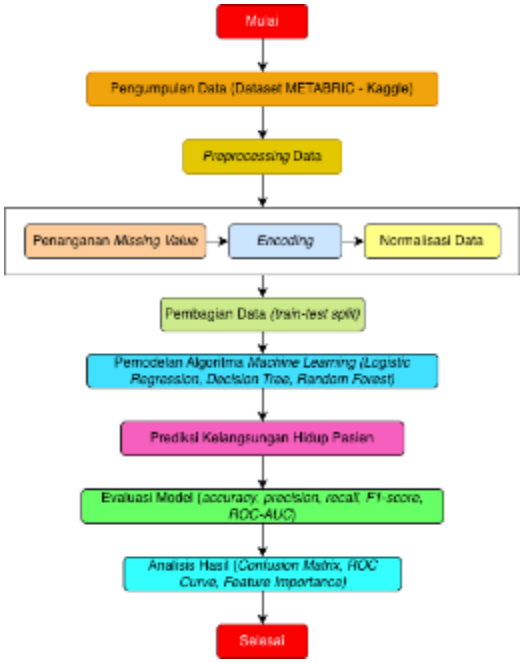
Studi komparasi algoritma *machine learning* juga banyak dilakukan pada berbagai domain untuk menentukan model yang paling optimal. Penelitian oleh [16] membandingkan beberapa algoritma klasifikasi seperti Naive Bayes, SVM, K-NN, Decision Tree, dan Random Forest pada analisis sentimen ulasan aplikasi digital dan menunjukkan bahwa performa model sangat dipengaruhi oleh karakteristik data yang digunakan. Penelitian lain [17] juga melakukan komparasi beberapa algoritma *machine learning* pada prediksi *customer churn* dan menemukan bahwa model *ensemble* mampu memberikan performa yang lebih tinggi dibandingkan model dasar. Temuan tersebut menunjukkan bahwa pemilihan algoritma yang tepat merupakan faktor penting dalam membangun model prediksi yang akurat, sehingga pendekatan komparatif juga diterapkan dalam penelitian ini untuk menentukan algoritma terbaik dalam memprediksi kekambuhan kanker payudara.

Berdasarkan tinjauan literatur tersebut, dapat diidentifikasi bahwa sebagian besar penelitian berfokus pada pengembangan model dengan berbagai kombinasi fitur dan algoritma, namun belum terdapat konsensus mengenai model klasifikasi yang paling optimal ketika hanya menggunakan karakteristik tumor sebagai variabel utama. Oleh karena itu, penelitian ini bertujuan untuk membangun dan membandingkan performa beberapa algoritma *machine learning*, yaitu Logistic Regression, Decision Tree, dan Random Forest, dalam memprediksi kekambuhan kanker payudara berdasarkan karakteristik tumor pasien menggunakan dataset METABRIC. Kebaruan penelitian ini terletak pada fokus eksklusif terhadap karakteristik biologis tumor sebagai fitur prediksi, tanpa menggabungkan data genomik atau klinis tambahan, sehingga menghasilkan model yang lebih sederhana namun tetap dapat dievaluasi secara komprehensif. Evaluasi dilakukan menggunakan beberapa metrik klasifikasi, meliputi *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*, disertai analisis *feature importance* untuk memperoleh gambaran performa yang menyeluruh. Hasil penelitian ini diharapkan dapat berkontribusi dalam pengembangan sistem pendukung keputusan klinis berbasis data untuk stratifikasi risiko kekambuhan pasien kanker payudara secara objektif.

2. METODE PENELITIAN

Bagian ini menguraikan tahapan penelitian yang dilakukan, mulai dari pengumpulan data, pra-pemrosesan, pemodelan *machine learning*, hingga evaluasi dan analisis hasil untuk memprediksi kekambuhan kanker payudara.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15976



Gambar 1. Alur Penelitian

2.1. Sumber dan Deskripsi Data

Data penelitian ini menggunakan *dataset Breast Cancer Recurrence* (METABRIC) yang diakses melalui Kaggle. *Dataset* ini merupakan hasil kolaborasi internasional yang mencakup rekam medis klinis, profil molekuler, hingga durasi kelangsungan hidup pasien secara mendalam. Dari total 1.986 pasien yang tercatat, setiap entri mewakili satu sampel tumor primer. Meskipun *dataset* aslinya menyediakan lebih dari 50 variabel, penelitian ini memfokuskan analisis pada karakteristik biologis tumor. Langkah ini diambil karena variabel biologis memiliki nilai prognostik yang kuat dalam menentukan risiko kekambuhan. Selain itu, pembatasan variabel ini bertujuan agar model yang dihasilkan lebih sederhana, mudah dipahami (*interpretable*), dan relevan saat diaplikasikan dalam pengambilan keputusan klinis di lapangan.

2.2. Pra-pemrosesan Data dan Seleksi Fitur

Tahap pra-pemrosesan dilakukan untuk menjamin kualitas dan konsistensi data sebelum masuk ke fase pemodelan. Dalam menangani nilai kosong (*missing value*), penelitian ini menerapkan metode penghapusan baris (*listwise deletion*) untuk menjaga integritas data asli [18]. Meskipun pendekatan ini mengakibatkan pengurangan jumlah data yang cukup signifikan, metode ini dipilih karena variabel yang digunakan merupakan karakteristik biologis tumor yang bersifat sensitif secara klinis.

Penanganan *missing value* merupakan tahap krusial dalam analisis data klinis karena berpengaruh langsung terhadap validitas hasil penelitian [18]. Penggunaan metode imputasi, seperti K-Nearest Neighbors (KNN), berpotensi menghasilkan nilai sintetis yang tidak sepenuhnya merepresentasikan kondisi biologis sebenarnya serta dapat menimbulkan bias dalam proses pemodelan [19]. Oleh karena itu, penelitian ini lebih mengutamakan kualitas dan konsistensi data dibandingkan kuantitas. Meskipun terjadi pengurangan data sekitar 36%, jumlah data yang tersisa masih mencukupi untuk proses pelatihan model machine learning secara optimal.

Penentuan fitur dilakukan dengan mempertimbangkan aspek klinis serta didukung oleh literatur terdahulu yang mengonfirmasi bahwa karakteristik biologis tumor merupakan prediktor kuat terhadap risiko kekambuhan [1], [5], [15]. Variabel yang diinklusikan dalam analisis ini meliputi *tumor size*, *tumor stage*, *neoplasm histologic grade*, *lymph nodes examined positive*, *Nottingham Prognostic Index (NPI)*, *mutation count*, dan *cellularity*.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15976

Selain itu, variabel target *Relapse Free Status* dikonversi menjadi format biner untuk keperluan klasifikasi, dengan kategori "Not Recurred" sebagai kelas negatif dan "Recurred" sebagai kelas positif. Proses ini bertujuan untuk memastikan kompatibilitas data dengan algoritma machine learning yang digunakan.

Selanjutnya, dilakukan normalisasi data pada fitur numerik menggunakan teknik *scaling* dengan metode *StandardScaler* untuk menyamakan rentang nilai antarvariabel. Proses ini bertujuan untuk meningkatkan kinerja model, khususnya pada algoritma Logistic Regression yang sensitif terhadap skala data. Dengan demikian, dataset yang digunakan telah berada dalam kondisi bersih, terstruktur, dan siap untuk tahap pembagian data serta pelatihan model.

2.3. Pembagian Data dan Pemodelan *Machine Learning*

Dataset yang telah melalui tahap pra-pemrosesan kemudian dibagi menjadi data pelatihan (*training set*) dan data pengujian (*testing set*) dengan rasio 80:20 menggunakan metode *train-test split*. Proporsi ini dipilih untuk memberikan keseimbangan optimal antara kedalaman proses pembelajaran model dan validitas evaluasi pada data yang belum pernah terbaca sebelumnya. Guna menjamin reproduktibilitas eksperimen, parameter *random_state* ditetapkan pada nilai 42. Selain itu, teknik *stratified sampling* diterapkan pada variabel target *Relapse Free Status* untuk menjaga konsistensi distribusi kelas di kedua subset data.

Penelitian ini mengimplementasikan tiga algoritma klasifikasi, yaitu Logistic Regression, Decision Tree, dan Random Forest. Logistic Regression digunakan sebagai model dasar (*baseline*) karena tingkat interpretabilitasnya yang tinggi dalam analisis medis [2], [12]. Decision Tree digunakan untuk menangkap hubungan non-linear antarvariabel, sedangkan Random Forest diterapkan sebagai metode *ensemble* untuk meningkatkan stabilitas dan akurasi prediksi [3], [4], [13]. Model yang telah dilatih kemudian digunakan untuk melakukan prediksi terhadap data pengujian guna menentukan status kekambuhan pasien.

2.4. Evaluasi dan Analisis Model

Performa model diukur melalui metrik *accuracy, precision, recall, F1-score*, serta *Receiver Operating Characteristic – Area Under Curve (ROC-AUC)*. *ROC-AUC* digunakan sebagai indikator utama karena kemampuannya dalam merepresentasikan daya diskriminasi model secara objektif [1], [11]. Dalam konteks medis, *recall* menjadi perhatian utama karena berkaitan dengan kemampuan model dalam mendeteksi pasien yang benar-benar berisiko mengalami kekambuhan.

Selain evaluasi kuantitatif, dilakukan analisis lebih lanjut menggunakan *confusion matrix* pada model Random Forest untuk melihat distribusi prediksi benar dan salah pada masing-masing kelas, mengingat model ini menunjukkan keseimbangan performa terbaik di antara ketiga algoritma yang dibandingkan. Kurva *ROC* juga dianalisis untuk mengevaluasi kemampuan model dalam membedakan kelas pada berbagai nilai ambang.

Selanjutnya, analisis *feature importance* pada model Random Forest digunakan untuk mengidentifikasi variabel yang paling berkontribusi terhadap prediksi risiko kekambuhan. Pendekatan ini mendukung prinsip *explainable artificial intelligence* (XAI) dalam sektor kesehatan [14], [15].

3. HASIL DAN PEMBAHASAN

3.1. Pengolahan dan *Preprocessing* Data

Dataset *Breast Cancer Recurrence* (METABRIC) yang digunakan dalam penelitian ini terdiri dari 1.986 sampel pasien dengan berbagai atribut klinis dan molekuler. Pada tahap awal, dilakukan seleksi fitur dengan memfokuskan analisis pada karakteristik biologis tumor yang secara klinis relevan terhadap risiko kekambuhan, yaitu *tumor size*, *tumor stage*, *neoplasm histologic grade*, *lymph nodes examined positive*, *Nottingham Prognostic Index (NPI)*, *mutation count*, dan *cellularity*. Tahap *preprocessing* diawali dengan identifikasi dan penanganan nilai

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15976

kosong (*missing values*). Data yang memiliki nilai tidak lengkap pada atribut yang digunakan dihapus menggunakan metode *listwise deletion* untuk menjaga konsistensi dan kualitas data pelatihan. Setelah proses pembersihan data, diperoleh 1.269 data yang siap digunakan untuk tahap pemodelan. Selain itu, variabel target *Relapse Free Status* dikonversi menjadi format biner untuk keperluan klasifikasi, dengan kategori "Not Recurred" sebagai kelas negatif dan "Recurred" sebagai kelas positif. Proses ini bertujuan untuk memastikan kompatibilitas data dengan algoritma machine learning yang digunakan. Berdasarkan hasil pra-pemrosesan, distribusi kelas pada variabel target menunjukkan bahwa terdapat 734 data (57,84%) dengan kategori "Not Recurred" dan 535 data (42,16%) dengan kategori "Recurred". Hal ini menunjukkan bahwa dataset relatif seimbang, meskipun terdapat sedikit dominasi pada kelas negatif. Selanjutnya, dilakukan normalisasi data pada seluruh fitur numerik menggunakan teknik *scaling* dengan bantuan *StandardScaler*. Proses ini bertujuan untuk menyamakan skala antar fitur sehingga model, khususnya Logistic Regression, dapat bekerja secara optimal. Normalisasi dilakukan setelah pembagian data, di mana proses *fit* diterapkan pada data pelatihan dan *transform* pada data pengujian untuk menghindari *data leakage*. Penyederhanaan fitur pada karakteristik tumor dilakukan untuk mengurangi kompleksitas model serta meminimalkan pengaruh variabel non-klinis yang tidak berkaitan langsung dengan risiko kekambuhan. Dengan demikian, dataset akhir yang digunakan telah berada dalam kondisi bersih, terstruktur, dan siap untuk tahap pembagian data serta pelatihan model.

3.2. Pembagian Data dan Pelatihan Model

Dataset selanjutnya dibagi menjadi data pelatihan dan data pengujian menggunakan metode *train–test split* dengan rasio 80:20. Parameter *random_state* ditetapkan sebesar 42 untuk menjaga reproduktibilitas eksperimen, sementara teknik *stratified sampling* diterapkan agar distribusi kelas pada variabel target tetap seimbang pada kedua subset data.

Sebanyak 1.015 data digunakan sebagai data pelatihan, sedangkan 254 data digunakan sebagai data pengujian. Tiga algoritma klasifikasi diterapkan dalam penelitian ini, yaitu Logistic Regression, Decision Tree, dan Random Forest. Ketiga model dilatih menggunakan data pelatihan dan kemudian diuji pada data pengujian untuk menilai kemampuan prediksi terhadap status kekambuhan pasien. Penggunaan beberapa algoritma memungkinkan perbandingan pendekatan linear, berbasis aturan, dan *ensemble learning* dalam mempelajari hubungan antara karakteristik tumor dan risiko kekambuhan.

3.3. Evaluasi Performa Model

Evaluasi performa dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*. Hasil evaluasi ketiga model ditunjukkan pada Tabel 1.

Tabel 1. Perbandingan Performa Model Machine Learning

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*ROC-AUC*
Logistic Regression	0.661	0.661	0.402	0.500	0.689
Decision Tree	0.547	0.462	0.458	0.460	0.527
Random Forest	0.614	0.544	0.544	0.529	0.619

Logistic Regression menghasilkan nilai *accuracy* sebesar 0,661, diikuti Random Forest sebesar 0,614 dan Decision Tree sebesar 0,547. Nilai *ROC-AUC* masing-masing model adalah 0,689, 0,619, dan 0,527.

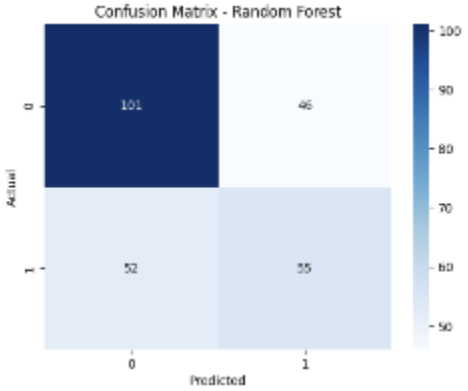
e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15976

Hasil tersebut menunjukkan bahwa pendekatan linear mampu memberikan performa klasifikasi yang stabil pada dataset penelitian. Namun demikian, Random Forest memperlihatkan nilai *recall* yang lebih tinggi dibandingkan model lainnya, yang mengindikasikan kemampuan lebih baik dalam mendeteksi pasien yang mengalami kekambuhan. Dalam konteks prediksi medis, kemampuan mendeteksi kasus positif menjadi aspek penting karena kesalahan klasifikasi dapat berdampak pada keterlambatan penanganan pasien.

Sebaliknya, performa Decision Tree yang relatif lebih rendah menunjukkan keterbatasan model pohon keputusan tunggal dalam melakukan generalisasi terhadap data baru. Kondisi ini sering dikaitkan dengan kecenderungan overfitting pada data pelatihan, terutama ketika pola data bersifat kompleks.

3.4. Analisis *Confusion Matrix* dan *ROC Curve*

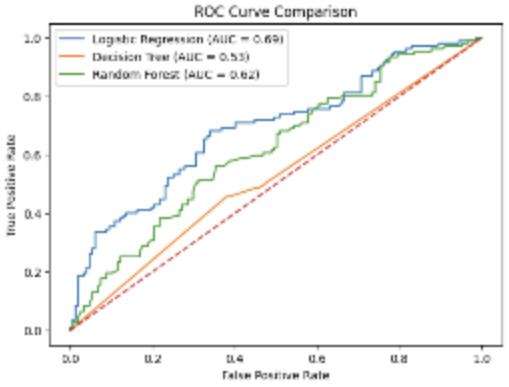
Untuk memperoleh gambaran lebih rinci mengenai kinerja klasifikasi, dilakukan analisis *confusion matrix* menggunakan model Random Forest. Hasil pengujian ditampilkan pada Gambar 2.



Gambar 2. *Confusion Matrix* Model Random Forest

Confusion matrix menunjukkan distribusi prediksi benar dan salah pada kedua kelas. Model mampu mengidentifikasi sebagian besar pasien yang tidak mengalami kekambuhan, namun masih ditemukan sejumlah kasus *false negative* yang menunjukkan bahwa beberapa pasien berisiko belum berhasil dideteksi oleh model.

Evaluasi tambahan dilakukan menggunakan kurva *ROC* yang ditampilkan pada Gambar 3.



Gambar 3. *ROC Curve* Perbandingan Model

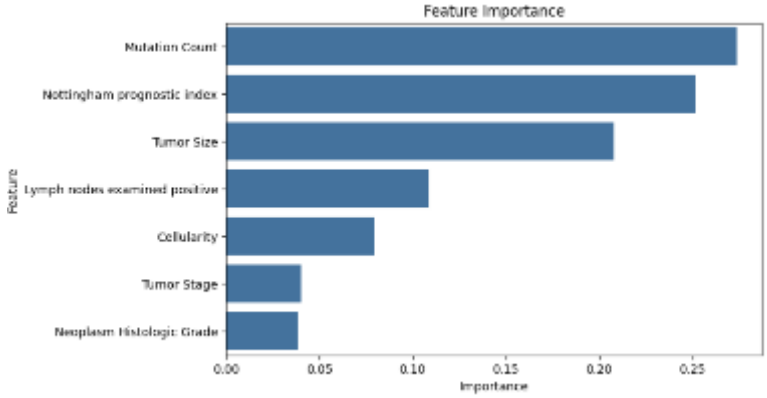
Nilai AUC menunjukkan bahwa Logistic Regression memiliki kemampuan diskriminasi kelas tertinggi, diikuti oleh Random Forest dan Decision Tree. Kurva *ROC* Decision Tree yang mendekati garis diagonal menunjukkan kemampuan klasifikasi yang relatif terbatas dibandingkan model lainnya.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15976

Temuan ini memperlihatkan bahwa pendekatan *ensemble* pada Random Forest mampu meningkatkan stabilitas prediksi dibandingkan pohon keputusan tunggal, terutama dalam mempelajari pola data dengan karakteristik non-linear.

3.5. Analisis *Feature Importance*

Interpretasi model dilakukan menggunakan analisis *feature importance* pada algoritma Random Forest untuk mengidentifikasi variabel yang paling berkontribusi terhadap prediksi kekambuhan. Hasil analisis ditunjukkan pada Gambar 4.



Gambar 4. Grafik *Feature Importance*

Mutation Count muncul sebagai fitur dengan kontribusi terbesar, diikuti Nottingham Prognostic Index dan Tumor Size. Dominasi ketiga variabel tersebut menunjukkan bahwa faktor biologis tumor dan indikator prognostik memiliki pengaruh kuat terhadap kemungkinan terjadinya kekambuhan.

Kontribusi lymph nodes examined positive dan cellularity berada pada tingkat moderat, sedangkan tumor stage dan histologic grade memberikan pengaruh yang relatif lebih kecil. Meskipun demikian, seluruh variabel tetap berperan dalam meningkatkan kemampuan model dalam mengenali pola klinis pasien.

Hasil ini sejalan dengan konsep klinis bahwa perubahan genetik dan karakteristik biologis tumor berperan penting dalam menentukan agresivitas kanker serta peluang terjadinya kekambuhan.

3.6. Implikasi Hasil Penelitian

Secara keseluruhan, hasil penelitian menunjukkan bahwa pendekatan machine learning mampu digunakan untuk memprediksi risiko kekambuhan kanker payudara berdasarkan karakteristik tumor pasien. Logistic Regression menunjukkan kemampuan diskriminasi yang baik, sementara Random Forest memberikan keseimbangan performa terutama dalam mendeteksi kasus berisiko.

Temuan ini mengindikasikan bahwa model prediksi berbasis data berpotensi dikembangkan sebagai sistem pendukung keputusan klinis dalam membantu proses stratifikasi risiko pasien. Dengan identifikasi risiko yang lebih dini, pemantauan dan perencanaan terapi lanjutan dapat dilakukan secara lebih terarah.

Namun demikian, performa model yang masih berada pada kategori moderat menunjukkan bahwa kekambuhan kanker payudara kemungkinan dipengaruhi oleh faktor lain di luar karakteristik tumor, seperti jenis terapi, faktor genetik individu, maupun kondisi gaya hidup pasien yang belum dimasukkan dalam penelitian ini.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15976

4. KESIMPULAN DAN SARAN

Hasil evaluasi menunjukkan adanya perbedaan performa di antara model Logistic Regression, Decision Tree, dan Random Forest dalam memprediksi kekambuhan kanker payudara berdasarkan karakteristik tumor. Hasil penelitian menunjukkan bahwa Logistic Regression memberikan performa terbaik dengan *accuracy* sebesar 0,661 dan *ROC-AUC* sebesar 0,689, sedangkan Random Forest memberikan keseimbangan performa yang lebih baik, terutama ditunjukkan melalui nilai recall tertinggi dalam mendeteksi pasien yang berisiko mengalami kekambuhan. Decision Tree memiliki performa paling rendah, yang mengindikasikan keterbatasan model pohon tunggal dalam melakukan generalisasi terhadap data yang kompleks.

Analisis *feature importance* menunjukkan bahwa Mutation Count, Nottingham Prognostic Index, dan Tumor Size merupakan variabel yang paling berpengaruh dalam proses prediksi, sementara variabel lain memberikan kontribusi yang relatif lebih kecil. Temuan ini menegaskan bahwa karakteristik biologis tumor memiliki peran penting dalam proses stratifikasi risiko pasien.

Secara keseluruhan, penelitian ini merupakan bukti konsep awal (*proof of concept*) yang menunjukkan bahwa pendekatan *machine learning* berpotensi dikembangkan lebih lanjut untuk mendukung prediksi kekambuhan kanker payudara. Namun demikian, dengan nilai akurasi tertinggi sebesar 0,661, model yang dihasilkan belum memenuhi standar yang dipersyaratkan untuk penerapan klinis secara langsung dan tidak boleh digunakan sebagai satu-satunya dasar pengambilan keputusan medis. Peningkatan performa model masih sangat diperlukan, antara lain melalui integrasi variabel terapi, faktor genetik individu, dan data gaya hidup pasien. Selain itu, validasi pada dataset yang lebih besar dan beragam merupakan prasyarat penting sebelum model ini dapat dipertimbangkan untuk implementasi klinis yang lebih luas.

DAFTAR PUSTAKA

[1] Lu D, Long X, Fu W, Liu B, Zhou X, Sun S. Predictive value of machine learning for breast cancer recurrence: a systematic review and meta-analysis. Journal of Cancer Research and Clinical Oncology. 2023.

[2] El Haji H, et al. Evolution of breast cancer recurrence risk prediction: a systematic review of statistical and machine learning–based models. JCO Clinical Cancer Informatics. 2023;7.

[3] Zuo D, Yang L, Jin Y, Qi H, Liu Y, Ren L. Machine learning-based models for the prediction of breast cancer recurrence risk. BMC Medical Informatics and Decision Making. 2023;23(1).

[4] Abbas NR, Alkattan H, Aldallal IB. Machine learning approaches for predicting breast cancer recurrence: a comparative analysis. Mesopotamian Journal of Artificial Intelligence in Healthcare. 2025;2025:208–218.

[5] Lee TF, et al. A machine learning model for predicting breast cancer recurrence and supporting personalized treatment decisions through comprehensive feature selection and explainable ensemble learning. Cancer Management and Research. 2025;17:917–932.

[6] Buyrukoglu G. Survival analysis in breast cancer: evaluating ensemble learning techniques for prediction. PeerJ Computer Science. 2024;10.

[7] Noman SM, et al. Leveraging survival analysis and machine learning for accurate prediction of breast cancer recurrence and metastasis. Scientific Reports. 2025;15(1).

[8] Azeroual S, Ben-Bouazza FE, Naqi A, Sebihi R. Predicting disease recurrence in breast cancer patients using machine learning models with clinical and radiomic characteristics: a retrospective study. Journal of the Egyptian National Cancer Institute. 2024;36(1).

[9] González-Castro L, et al. Machine learning algorithms to predict breast cancer recurrence using structured and unstructured sources from electronic health records. Cancers. 2023;15(10).

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15976

[10] Sahoo G, et al. Predicting breast cancer relapse from histopathological images with ensemble machine learning models. Current Oncology. 2024;31(11):6577–6597.
[11] Nopour R. Prediction of one-year recurrence among breast cancer patients undergone surgery using artificial intelligence-based algorithms: a retrospective study on prognostic factors. BMC Cancer. 2025;25(1).
[12] Gupta SR. Prediction time of breast cancer tumor recurrence using machine learning. Cancer Treatment Research Communications. 2022;32.
[13] Howard FM, et al. Machine learning-based prediction of distant recurrence risk and ribociclib treatment effect in HR+/HER2- early breast cancer using real-world and NATALEE data. Clinical Cancer Research. 2026;32(2):428–437.
[14] Ghasemi A, Hashtarkhani S, Schwartz DL, Shaban-Nejad A. Explainable artificial intelligence in breast cancer detection and risk prediction: a systematic scoping review. Cancer Informatics. 2024.
[15] Liu Y, Fu Y, Peng Y, Ming J. Clinical decision support tool for breast cancer recurrence prediction using SHAP value in cooperative game theory. Heliyon. 2024;10(2).
[16] Al Fachrozi M, Tania KD. Comparison of Naïve Bayes, SVM, K-NN, Decision Tree, and Random Forest in sentiment analysis based on Seabank application aspects. JURTEKSI (Jurnal Teknologi dan Sistem Informasi). 2025;12(1):69–80.
[17] Ardhani DA, Tania KD. Knowledge discovery on e-commerce customer churn using interpretable machine learning: a comparative study of SHAP-based classifiers. Journal of Applied Artificial Intelligence and Computing. 2025.
[18] A. Afkanpour *et al*., "Handling missing data in clinical research: a comprehensive review," *BMC Medical Research Methodology*, vol. 24, no. 1, 2024.
[19] E. Curnow *et al*., "The impact of imputation methods on bias in epidemiological studies," *Frontiers in Epidemiology*, vol. 3, 2023.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15976