

ANALISIS PREDIKTIF KELANGSUNGAN HIDUP PASIEN KANKER PAYUDARA MENGGUNAKAN ALGORITMA RANDOM FOREST

Vanya Dwi Nabila, Chairunnisa Desti Arzety, Aprilia Herawati, Fathoni, Ali Ibrahim

Sistem Informasi, Universitas Sriwijaya
Jalan Palembang-Prabumulih KM 32 Indralaya, Indonesia
09031382328119@student.unsri.ac.id

ABSTRAK

Kanker payudara merupakan salah satu penyebab kematian tertinggi pada perempuan di dunia. Permasalahan utama dalam penanganan penyakit ini adalah rendahnya akurasi prediksi kelangsungan hidup akibat kompleksitas hubungan *non-linear* antar variabel klinis dan tingginya dimensi data. Penelitian ini bertujuan untuk membangun model prediksi kelangsungan hidup pasien kanker payudara menggunakan algoritma *Random Forest*. Metode yang digunakan meliputi *preprocessing data*, pelatihan model, serta evaluasi menggunakan *10-Fold Cross-Validation* dengan metrik *accuracy*, *precision*, *recall*, dan *F1-Score* pada dataset METABRIC yang terdiri dari 1.986 data pasien. Hasil penelitian menunjukkan bahwa model memperoleh *accuracy* sebesar 71,04%, *precision* 72,31%, *recall* 81,50%, dan *F1-Score* 76,63%, dengan rata-rata akurasi *cross-validation* sebesar 66,75%. Analisis *feature importance* menunjukkan bahwa *Age at Diagnosis*, *Nottingham Prognostic Index*, dan *Tumor Size* merupakan faktor yang paling berpengaruh terhadap prediksi kelangsungan hidup pasien.

Kata kunci : Kanker Payudara, Machine Learning, Random Forest, Survival Prediction, Klasifikasi, Feature Importance.

1. PENDAHULUAN

Kanker payudara merupakan salah satu jenis kanker dengan angka kejadian tertinggi pada perempuan di dunia dan masih menjadi penyebab utama kematian akibat kanker. Data global terkini mengonfirmasi adanya tren kenaikan kasus tahunan secara konsisten di berbagai negara, tanpa memandang tingkat kemajuan ekonominya. Fenomena ini menegaskan bahwa penanganan kanker payudara merupakan isu kesehatan publik mendesak yang memerlukan keterlibatan serius dari seluruh pemangku kepentingan [1]. Secara biologis, penyakit ini dipicu oleh mutasi genetik pada sel epitelial di saluran atau lobulus payudara, yang mengakibatkan proliferasi sel abnormal yang tidak terkendali dan berisiko menyebar ke organ vital lain melalui jalur metastasis [2]. Karakteristiknya yang sangat heterogen menyebabkan variasi signifikan dalam kondisi klinis, efektivitas pengobatan, dan proyeksi prognosis pada setiap individu pasien [3]. Konsekuensinya, diperlukan instrumen yang lebih menyeluruh untuk membantu proses diagnosis, penentuan prognosis, serta pemilihan strategi terapi yang presisi.

Hambatan fundamental dalam penanggulangan penyakit ini adalah seringnya terjadi keterlambatan diagnosis dan rendahnya akurasi dalam memprediksi harapan hidup pasien. Minimnya skrining dini dan rendahnya kesadaran publik untuk melakukan pemeriksaan mandiri mengakibatkan banyak kasus baru teridentifikasi saat sudah mencapai stadium lanjut. Kondisi tersebut secara langsung memperburuk probabilitas kesembuhan dan memicu lonjakan angka kematian. [6][7]. Di sisi lain, prognosis pasien dipengaruhi oleh banyak faktor yang saling berinteraksi, seperti usia saat diagnosis, ukuran tumor, status kelenjar getah bening, stadium kanker, riwayat

keluarga, ekspresi biomarker, komorbiditas, serta kondisi fisiologis pasien [4][5][8]. Interaksi antarvariabel yang rumit ini membuat model statistik tradisional sering kali tidak memadai, terutama saat mengolah data yang memiliki dimensi tinggi, pola *non-linear*, maupun masalah multikolinearitas antar fitur.

Kemajuan dalam teknologi kecerdasan buatan, khususnya bidang *machine learning*, menyajikan solusi inovatif untuk mengatasi hambatan tersebut. Sejumlah algoritma populer seperti *Decision Tree*, *Naive Bayes*, *Support Vector Machine* (SVM), serta *Random Forest* telah diimplementasikan dalam klasifikasi medis dengan hasil yang menjanjikan [9][10]. Studi literatur sistematis bahkan mengonfirmasi bahwa pemanfaatan AI pada kasus kanker payudara mampu mencapai performa prediksi yang superior berkat kemampuannya dalam mengekstraksi pola tersembunyi dari data klinis yang kompleks [11]. Terkait prediksi kelangsungan hidup, beberapa temuan menunjukkan bahwa *machine learning* lebih unggul dibanding teknik konvensional karena sifatnya yang adaptif dan mampu memproses beragam variabel secara simultan [12][13].

Namun demikian, mayoritas studi terdahulu masih terpaku pada pemisahan tumor jinak dan ganas, sementara kajian khusus mengenai prediksi kelangsungan hidup pasien masih cukup langka. Selain itu, model-model mutakhir seperti *deep learning* sering kali dikritik karena sifatnya yang sulit dipahami (*black-box*), sehingga menyulitkan dokter dalam melakukan interpretasi medis. Padahal, transparansi model sangat krusial dalam dunia medis agar hasil prediksi dapat dipertanggungjawabkan sebagai basis pengambilan tindakan klinis. Oleh karena itu, masih ada urgensi untuk mengembangkan

model prediktif yang tidak hanya memiliki akurasi tinggi dan stabil, tetapi juga mampu memberikan penjelasan mengenai faktor-faktor kunci yang memengaruhi prognosis.

Berdasarkan permasalahan tersebut, penelitian ini menawarkan rencana penyelesaian melalui penerapan algoritma Random Forest untuk menganalisis kelangsungan hidup pasien kanker payudara menggunakan dataset METABRIC. Pemilihan *Random Forest* didasarkan pada ketangguhannya dalam mengolah data berdimensi tinggi, ketahanan terhadap data yang berderau (*noise*), kemampuan memetakan hubungan non-linear, serta penyediaan fitur *feature importance* untuk mengukur signifikansi setiap variabel. [14][16]. Alur penelitian ini meliputi akuisisi data, prapemrosesan, pembagian data, pelatihan serta pengujian model, hingga validasi menggunakan teknik *10-Fold Cross Validation*. Kualitas model akan diukur melalui parameter *accuracy*, *precision*, *recall*, dan *F1-Score*.

Penelitian ini bertujuan untuk merancang dan menganalisis performa model prediksi kelangsungan hidup pasien kanker payudara berbasis *Random Forest*, sekaligus memetakan variabel klinis yang paling menentukan prognosis. Luaran dari studi ini diharapkan dapat berkontribusi dalam penciptaan sistem pendukung keputusan klinis berbasis AI, yang membantu praktisi medis memahami pola penyakit secara lebih mendalam dan merumuskan strategi perawatan yang lebih efisien. Lebih jauh lagi, penelitian ini diproyeksikan dapat memperkaya literatur kesehatan digital, terutama terkait penerapan analisis prediktif kecerdasan buatan di lingkungan domestik Indonesia.

2. TINJAUAN PUSTAKA

Bagian ini menguraikan konsep teoritis yang relevan dengan penelitian, meliputi kanker payudara, *machine learning*, serta algoritma *Random Forest* sebagai metode yang digunakan.

2.1. Kanker Payudara dan Faktor Prognosis

Kanker payudara adalah penyakit dengan karakteristik heterogen yang dipengaruhi oleh berbagai faktor klinis dan biologis. Faktor-faktor seperti usia saat diagnosis, ukuran tumor, status kelenjar getah bening, serta ekspresi *biomarker* seperti *estrogen receptor (ER)*, *progesterone receptor (PR)*, dan *HER2* berperan penting dalam menentukan prognosis pasien [8]. Salah satu indikator *prognostik* yang banyak digunakan secara klinis adalah *Nottingham Prognostic Index (NPI)*, yaitu nilai komposit yang mencerminkan tingkat keparahan tumor berdasarkan karakteristik biologis dan penyebaran kanker dalam tubuh pasien [8]. Selain itu, kondisi *komorbid* dan tingkat penyebaran kanker juga menjadi indikator utama dalam menentukan tingkat kelangsungan hidup pasien [7].

Untuk menghasilkan prediksi *prognostik* yang lebih akurat, diperlukan pendekatan berbasis

data, karena hubungan antara variabel-variabel ini kompleks dan seringkali non-linear, sehingga pendekatan analisis konvensional tidak dapat menangkap pola interaksi antar faktor [6].

2.2. Machine Learning dalam Prediksi Kanker Payudara

Dalam bidang kesehatan, penggunaan *machine learning* telah berkembang pesat, khususnya dalam analisis kanker payudara. Untuk meningkatkan akurasi klasifikasi dan prediksi penyakit, berbagai algoritma telah digunakan, termasuk *Random Forest*, *Decision Tree*, *Naive Bayes*, dan *Support Vector Machine (SVM)*. Masing-masing algoritma memiliki karakteristik yang berbeda; *Naive Bayes* relatif sederhana namun mengasumsikan independensi antar fitur, *Decision Tree* mudah diinterpretasikan tetapi rentan terhadap *overfitting*, sedangkan *SVM* efektif pada data berdimensi tinggi namun membutuhkan komputasi yang lebih besar. *Random Forest* hadir sebagai pendekatan *ensemble* yang mengatasi kelemahan tersebut dengan menggabungkan secara acak hasil dari berbagai pohon keputusan. [9] [10].

Berdasarkan studi tinjauan sistematis, penerapan kecerdasan buatan dalam prediksi kanker payudara menunjukkan performa yang tinggi dengan kemampuan mengolah sejumlah data besar serta mengidentifikasi pola kompleks yang tidak dapat dideteksi dengan metode konvensional [11]. Selain itu, pendekatan berbasis *machine learning* juga terbukti lebih adaptif dalam menangani data berdimensi tinggi.

Namun demikian, sebagian besar penelitian masih berfokus pada klasifikasi tumor menjadi jinak dan ganas. Pendekatan ini belum sepenuhnya menjawab kebutuhan klinis yang memerlukan analisis lebih mendalam terkait prognosis dan kelangsungan hidup pasien [12].

2.3. Prediksi Kelangsungan Hidup (Survival Prediction)

Prediksi kelangsungan hidup pasien kanker payudara merupakan aspek penting dalam mendukung pengambilan keputusan klinis. Pendekatan ini mempertimbangkan berbagai faktor *prognostik* secara simultan untuk memperkirakan kemungkinan bertahan hidup pasien dalam periode tertentu.

Beberapa penelitian menunjukkan bahwa pendekatan *machine learning* dapat lebih akurat dalam memprediksi kelangsungan hidup dibandingkan metode tradisional. Studi oleh [12] menunjukkan bahwa penggunaan pendekatan *multimethod* dapat meningkatkan performa model dengan memanfaatkan kombinasi algoritma dan pemilihan fitur yang optimal. Penelitian lain oleh [13] menunjukkan bahwa penggunaan algoritma berbasis *ensemble* seperti *Random Forest*, *XGBoost*, dan *LightGBM* mampu mencapai performa tinggi dalam memprediksi metastasis dan kekambuhan kanker.

Selain itu, penelitian oleh [17] dan [18] menunjukkan bahwa model *machine learning* mampu

mengidentifikasi hubungan kompleks antar variabel klinis yang berpengaruh terhadap kelangsungan hidup pasien. Penelitian oleh [16] juga menunjukkan bahwa *Random Forest* memiliki kinerja yang lebih baik dibandingkan regresi logistik dalam prediksi prognosis kanker payudara.

Hasil-hasil tersebut menunjukkan bahwa metode berbasis *machine learning* memiliki keunggulan dalam menangani data yang sangat besar dan juga mampu memberikan prediksi yang lebih akurat dan komprehensif.

2.4. Algoritma Random Forest dalam Prediksi Kanker Payudara

Random Forest adalah salah satu algoritma *ensemble learning* yang banyak digunakan dalam analisis data medis karena kemampuannya untuk meningkatkan akurasi dan mengurangi risiko *overfitting*. Algoritma ini bekerja dengan menghasilkan sejumlah pohon keputusan secara acak dan kemudian menggabungkan hasilnya melalui mekanisme *majority voting*. Dua parameter utama yang memengaruhi performa model adalah *n_estimators* yang menentukan jumlah pohon keputusan yang dibangun, dan *max_depth* yang membatasi kedalaman setiap pohon guna mencegah *overfitting* pada data latih [14].

Random Forest memiliki keunggulan dalam menangani data kompleks, termasuk data berdimensi tinggi dan data dengan *noise*. Selain itu, algoritma ini memiliki kemampuan untuk mengumpulkan informasi tentang pentingnya fitur (*feature importance*), sehingga memberikan nilai tambah dalam interpretasi hasil model. Studi oleh [9] dan [10] menunjukkan bahwa *Random Forest* menghasilkan tingkat akurasi yang lebih tinggi dibandingkan *Naive Bayes* dan *Decision Tree*. Penelitian lain dalam [14] juga menunjukkan bahwa kombinasi *Random Forest* dengan teknik optimasi seperti *feature selection* dapat secara signifikan meningkatkan performa model. Ini berarti bahwa *Random Forest* adalah algoritma yang efisien dan fleksibel untuk memprediksi kelangsungan hidup pada pasien kanker payudara.

2.5. Penelitian Terdahulu

Berbagai penelitian terdahulu menunjukkan bahwa metode *machine learning* memiliki performa yang baik dalam klasifikasi maupun prediksi prognosis kanker payudara. Algoritma *Random Forest* dilaporkan memiliki tingkat akurasi yang lebih tinggi dibandingkan *Naive Bayes* dalam klasifikasi kanker payudara, terutama dalam menangani kompleksitas data [9]. Selain itu, dibandingkan dengan *Decision Tree*, *Random Forest* juga menunjukkan performa yang lebih stabil serta kemampuan generalisasi yang lebih baik pada dataset dengan fitur yang beragam [10].

Dalam konteks prediksi kelangsungan hidup pasien, pendekatan berbasis *multimethod* yang menggabungkan beberapa algoritma *machine*

learning terbukti mampu meningkatkan performa model secara signifikan [12]. Penggunaan algoritma *ensemble* seperti *Random Forest*, *XGBoost*, dan *LightGBM* juga menghasilkan nilai akurasi yang tinggi dalam memprediksi kekambuhan dan metastasis kanker payudara [13]. Selain itu, model berbasis *Random Forest* terbukti lebih unggul dibandingkan regresi logistik dalam memprediksi *outcome* klinis pasien [16].

Lebih lanjut, penelitian lain menunjukkan bahwa *machine learning* mampu mengidentifikasi hubungan kompleks antar variabel klinis yang berpengaruh terhadap prognosis pasien, termasuk pada kasus metastasis tulang [17][18]. Temuan-temuan tersebut menegaskan bahwa pendekatan berbasis *machine learning*, khususnya *Random Forest*, memiliki potensi besar dalam meningkatkan akurasi prediksi serta memberikan wawasan yang lebih mendalam dalam analisis kanker payudara.

2.6. Research Gap dan Posisi Penelitian

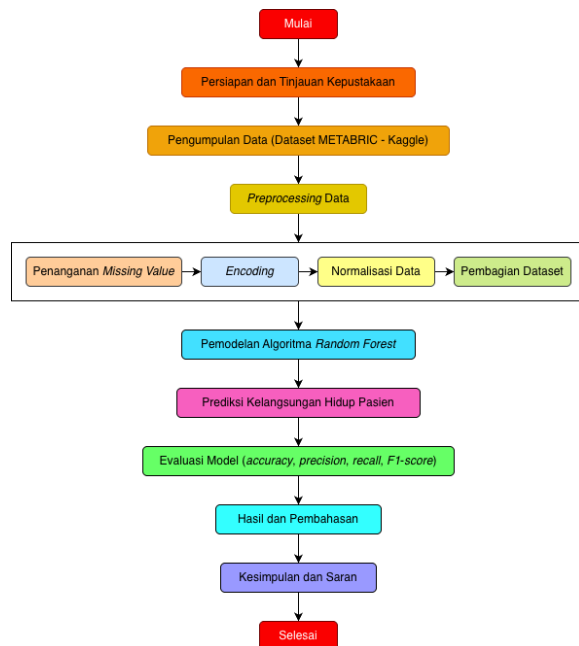
Meskipun berbagai penelitian telah mengkaji penerapan *machine learning* dalam kanker payudara, namun ada beberapa celah yang perlu diperhatikan. Sebagian besar penelitian masih berfokus pada klasifikasi tumor, sementara penelitian yang secara khusus membahas prediksi kelangsungan hidup pasien masih relatif terbatas. Selain itu, beberapa penelitian masih menggunakan metode statistik konvensional, yang memiliki keterbatasan ketika menangani hubungan antara data berdimensi tinggi dan non-linear. Di sisi lain, pendekatan berbasis *deep learning* meskipun memiliki akurasi tinggi, sering kali sulit diinterpretasikan sehingga kurang sesuai untuk kebutuhan klinis yang memerlukan transparansi.

Penelitian oleh [12] dan [13] memang telah menggunakan *Random Forest* dalam konteks *survival prediction*, namun keduanya berfokus pada prediksi kekambuhan dan metastasis dengan kombinasi beberapa algoritma sekaligus. Berbeda dengan pendekatan tersebut, penelitian ini secara spesifik memusatkan perhatian pada prediksi status kelangsungan hidup pasien menggunakan *Random Forest* sebagai algoritma tunggal berbasis *dataset* METABRIC, dengan penekanan pada analisis *feature importance* sebagai keluaran interpretatif yang bernilai klinis. Pendekatan ini dipilih karena mampu menggabungkan keunggulan dari sisi akurasi dan *interpretabilitas*, serta mengatasi keterbatasan metode sebelumnya, sehingga diharapkan dapat berkontribusi pada pengembangan model prediktif di bidang kesehatan yang lebih efektif dan aplikatif.

3. METODE PENELITIAN

Bagian ini menguraikan rangkaian prosedur sistematis yang ditempuh untuk mengembangkan model estimasi probabilitas hidup penderita kanker payudara melalui implementasi metode *Random Forest*. Secara komprehensif, tahapan kerja yang

dilakukan dalam studi ini dipresentasikan melalui Gambar 1.



Gambar 1. Kerangka Penelitian

Berdasarkan Gambar 1, proses penelitian ini mencakup beberapa tahapan utama, yaitu persiapan dan tinjauan kepustakaan, pengumpulan data, *preprocessing* data, pemodelan, prediksi, evaluasi, dan validasi model.

3.1. Persiapan dan Tinjauan Kepustakaan

Tahap awal penelitian ini mencakup perumusan masalah, penentuan target riset, serta pengkajian literatur dari berbagai referensi yang relevan. Tinjauan kepustakaan dilakukan untuk mendalami tren pemanfaatan *machine learning* pada prediksi kelangsungan hidup penderita kanker payudara, sekaligus untuk menemukan celah penelitian (*research gap*) sebagai landasan pengembangan model ini.

3.2. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang bersumber dari dataset METABRIC di platform Kaggle. Sebagai repositori terbuka yang terstandarisasi, Kaggle dipilih karena menyediakan data yang terstruktur untuk kebutuhan analisis *machine learning*.

Dataset yang digunakan berisi data klinis dari 1.986 pasien kanker payudara yang mencakup berbagai atribut penting seperti usia pasien, jenis kelamin, stadium tumor (*tumor stage*), jenis histologi, status *biomarker* (*ER*, *PR*, *HER2*), jenis operasi, serta status pasien (hidup atau meninggal). Data tersebut memungkinkan analisis hubungan antara karakteristik klinis pasien dengan tingkat kelangsungan hidup.

Pemanfaatan data sekunder ini dianggap efisien karena menyediakan basis data historis yang nyata dan sistematis. Hal ini sejalan dengan penelitian oleh [12] yang menyatakan bahwa pemanfaatan data klinis memungkinkan pengembangan model prediksi kelangsungan hidup secara sistematis dan berbasis data nyata. Selain itu, kualitas dan kelengkapan data menjadi faktor penting dalam meningkatkan performa model prediksi [11].

Secara teknis, akuisisi data dimulai dengan mengunduh berkas dari Kaggle, yang dilanjutkan dengan penyaringan fitur-fitur yang memiliki relevansi terhadap tujuan riset, khususnya variabel medis yang memengaruhi prognosis. Sebelum memasuki tahap *preprocessing*, dilakukan proses validasi guna memastikan seluruh data telah lengkap dan konsisten.

3.3. Preprocessing Data

Sebelum memasuki tahap pemodelan, serangkaian prosedur pengolahan awal diterapkan untuk menjamin kualitas data yang optimal. Proses ini dilaksanakan untuk memperbaiki kekurangan pada data serta menyesuaikan formatnya agar kompatibel dengan kebutuhan algoritma *machine learning*.

- Penanganan Nilai yang Hilang (*Missing Value*):** Sebelum data digunakan untuk pemodelan, dilakukan serangkaian tahapan *preprocessing* untuk memastikan tidak ada data kosong yang masuk ke tahap pemodelan. Penanganan yang tepat sangat penting agar informasi yang tidak lengkap tidak memengaruhi kualitas hasil prediksi.
- Pengodean Variabel Kategorikal (*Encoding*):** Fitur kategorikal seperti jenis histologi dan status *biomarker* (*ER*, *PR*, *HER2*) dikonversi menjadi bentuk numerik melalui teknik *Label Encoding* dan *One-Hot Encoding*. Proses ini dilakukan karena algoritma *Random Forest* memerlukan input dalam format numerik untuk proses komputasinya.
- Normalisasi Data:** Seluruh variabel numerik dikalibrasi menggunakan metode *Min-Max Scaler* ke dalam rentang $[0,1]$ guna menyelaraskan skala antar fitur [14]. Meskipun *Random Forest* pada dasarnya tidak terlalu dipengaruhi oleh perbedaan skala, normalisasi ini tetap diterapkan untuk menjaga konsistensi prosedur dan memudahkan analisis data.
- Pembagian *Dataset*:** Data dipisahkan menjadi dua bagian, yaitu data latih (*training*) dan data uji (*testing*) dengan perbandingan 80:20. Strategi ini diterapkan untuk mengoptimalkan proses pembelajaran model melalui mayoritas data, sekaligus sebagai sarana pengujian performa model pada data baru yang belum pernah diproses sebelumnya. Prosedur ini dilakukan dengan mengikuti metodologi yang juga digunakan dalam studi [9].

3.4. Pemodelan Algoritma Random Forest

Pada penelitian ini, pengembangan model prediksi kelangsungan hidup pasien dilakukan dengan menerapkan algoritma *Random Forest*. Metode ini termasuk dalam kategori *ensemble learning* yang bekerja dengan membentuk sekumpulan pohon keputusan (*decision trees*). Hasil dari setiap pohon kemudian diintegrasikan guna mengoptimalkan akurasi prediksi serta meminimalisir potensi terjadinya *overfitting*.

Setiap pohon keputusan dalam struktur *Random Forest* dikonstruksi melalui pemilihan subset data dan fitur secara acak. Mekanisme ini bertujuan menciptakan model yang lebih tangguh (*robust*) terhadap fluktuasi data dan efektif dalam mengolah dataset dengan dimensi yang kompleks.

Dalam proses pembentukan pohon, digunakan kriteria Gini Impurity untuk mengukur tingkat kemurnian suatu simpul. Nilai Gini yang lebih kecil menunjukkan bahwa simpul tersebut semakin homogen.

Di setiap percabangan, algoritma menentukan pemisahan terbaik dengan memilih sejumlah fitur acak dari keseluruhan fitur yang ada. Strategi tersebut diterapkan untuk memperluas keberagaman antar pohon di dalam model. Prediksi akhir diperoleh melalui mekanisme *majority voting*, yaitu kelas dengan jumlah prediksi terbanyak dari seluruh pohon keputusan akan menjadi hasil akhir model.

Parameter model ditentukan berdasarkan referensi penelitian sebelumnya dengan memperhatikan efisiensi antara performa dan tingkat kerumitan model. Parameter utama yang diimplementasikan dalam riset ini adalah $n_estimators$ sebesar 100 dan max_depth senilai 10. Seluruh proses pemodelan dijalankan menggunakan bahasa pemrograman Python dengan bantuan pustaka Scikit-learn di platform Google Colab [10].

3.5. Prediksi Kelangsungan Hidup Pasien

Model *Random Forest* yang sudah melewati tahap pelatihan kemudian diterapkan untuk mengestimasi status kelangsungan hidup subjek pada data uji. Setiap pasien diklasifikasikan ke dalam dua kategori, yaitu hidup (*Alive*) atau meninggal (*Dead*), berdasarkan fitur klinis yang tersedia dalam dataset METABRIC.

3.6. Evaluasi Model

Tahap evaluasi bertujuan untuk menilai efektivitas algoritma *Random Forest* dalam mengelompokkan status kelangsungan hidup pasien. Pengujian ini dilakukan dengan menggunakan data uji (*testing data*), yaitu kumpulan data yang sebelumnya tidak dilibatkan dalam proses pembelajaran model.

Performa model diukur melalui empat indikator utama: *accuracy*, *precision*, *recall*, dan *F1-Score*. Pemilihan keempat metrik ini dilakukan karena dapat memberikan representasi kinerja klasifikasi yang

menyeluruh, terutama dalam menghadapi potensi ketidakseimbangan kelas pada dataset.

- Accuracy*: Menilai rasio ketepatan prediksi terhadap keseluruhan jumlah data.
- Precision*: Menggambarkan tingkat akurasi model dalam menentukan kelas positif.
- Recall*: Mengukur sejauh mana model mampu menjaring seluruh data yang termasuk dalam kategori positif.
- F1-Score*: Merupakan nilai harmonik yang menyeimbangkan antara *precision* dan *recall*.

Selain metrik di atas, *confusion matrix* juga diterapkan untuk memetakan distribusi prediksi terhadap label yang sebenarnya. *Confusion matrix* membantu dalam mengidentifikasi jenis kekeliruan klasifikasi, baik pada kelas positif maupun negatif, dapat diidentifikasi secara lebih mendalam.

3.7. Validasi Model

Validasi dilakukan untuk menjamin model memiliki tingkat generalisasi yang optimal serta terhindar dari kendala *overfitting* pada data latih. Berbeda dengan evaluasi standar, validasi lebih menitikberatkan pada penilaian stabilitas kinerja model terhadap variasi pembagian data.

Metode yang digunakan dalam penelitian ini adalah *10-Fold Cross-Validation*. Dalam teknik ini, dataset dipisahkan menjadi sepuluh bagian (*fold*) yang seimbang. Di setiap iterasi, model dilatih dengan sembilan bagian dan diuji menggunakan satu bagian sisa secara bergantian hingga seluruh bagian pernah berfungsi sebagai data uji.

Rangkaian proses tersebut menghasilkan sepuluh nilai performa yang kemudian dihitung rata-ratanya untuk mendapatkan hasil yang lebih konsisten dan representatif. Penggunaan pendekatan ini efektif dalam meminimalisir bias yang sering muncul pada pembagian data tunggal (*train-test split*). Hasil validasi tersebut juga diperiksa kembali menggunakan *confusion matrix* untuk meninjau konsistensi prediksi model melalui komponen *True Positive*, *True Negative*, *False Positive*, dan *False Negative*.

4. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil dari penerapan model serta analisis mendalam mengenai kinerja dan efektivitas model dalam mengestimasi probabilitas kelangsungan hidup pasien.

4.1. Evaluasi dan Analisis Performa Model

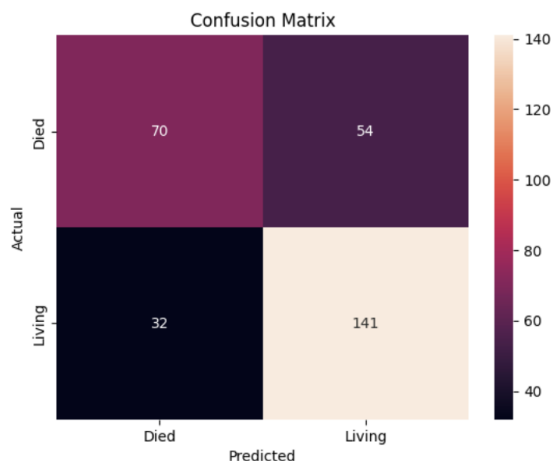
Penelitian ini mengimplementasikan algoritma *Random Forest* untuk menyusun model prediksi tingkat kelangsungan hidup penderita kanker payudara. Guna menguji kualitas dan keandalan model yang telah dibangun, dilakukan pengukuran menggunakan empat metrik evaluasi standar, yaitu *accuracy*, *precision*, *recall*, dan *F1-Score*. Ringkasan hasil pencapaian dari masing-masing metrik tersebut dirinci pada Tabel 1.

Tabel 1. Perbandingan Performa Model Machine Learning

Metrik	Nilai(%)
Accuracy	71.04
Precision	72.31
Recall	81.50
F1-Score	76.63

Berdasarkan hasil evaluasi pada Tabel 1, model mencatatkan tingkat akurasi senilai 71,04%, Angka tersebut merepresentasikan bahwa mayoritas prediksi model selaras dengan kondisi aktual. Sementara itu, perolehan *precision* sebesar 72,31% menggambarkan bahwa estimasi positif model memiliki derajat ketepatan yang memadai. Di sisi lain, skor *recall* yang menyentuh 81,50% membuktikan efektivitas model dalam mengenali pasien yang mampu bertahan hidup.

Nilai *recall* yang lebih tinggi dibandingkan *precision* menandakan bahwa model memiliki sensitivitas yang lebih menonjol dalam mendeteksi kategori positif. Pada ranah medis, sifat ini sangat krusial karena berfungsi menekan risiko kegagalan identifikasi pasien yang berpotensi hidup, sekalipun terdapat kemungkinan peningkatan kesalahan klasifikasi (*false positive*) pada kelompok negatif.



Gambar 2. Confusion matrix model random forest

Analisis mendalam dilakukan melalui instrumen *confusion matrix* seperti yang dipresentasikan pada Gambar 2. Data tersebut mengungkap bahwa model memiliki kecenderungan lebih optimal dalam mengidentifikasi subjek yang bertahan hidup, yang mengarah pada kemungkinan adanya ketidakseimbangan distribusi data atau dominasi model terhadap kelas mayoritas. Hasil ini selaras dengan perolehan skor *recall* yang melampaui *precision*, yang membuktikan bahwa model memiliki sensitivitas tinggi terhadap kategori positif.

Meski demikian, kekeliruan prediksi tetap muncul, ditunjukkan oleh adanya nilai *False Positive* dan *False Negative*. Kondisi *False Positive* menggambarkan adanya pasien yang diprediksi akan bertahan hidup namun pada kenyataannya tidak, di mana secara klinis hal ini berisiko memicu

pengambilan keputusan medis yang kurang akurat. Di sisi lain, *False Negative* menunjukkan keterbatasan model dalam mengenali sebagian pasien yang semestinya terdeteksi.

Hal ini menunjukkan bahwa walaupun sensitivitas model sudah memadai, kemampuan dalam mendiferensiasi kedua kelompok pasien masih memerlukan penyempurnaan, terutama pada aspek spesifisitas. Langkah perbaikan dapat diupayakan melalui penyesuaian parameter model (optimasi) atau penggunaan teknik penyeimbangan data guna mereduksi bias pada kelas mayoritas.

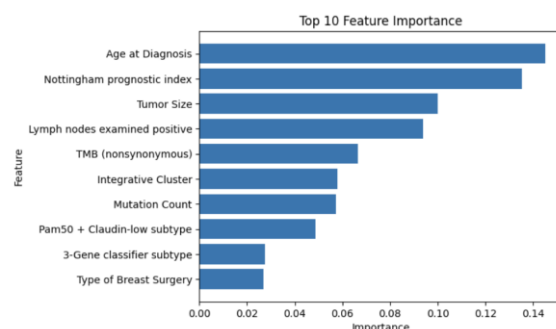
4.2. Validasi dan Generalisasi Model

Untuk mengevaluasi kestabilan dan kemampuan generalisasi model, dilakukan pengujian ulang melalui metode *10-Fold Cross-Validation*. Dalam teknik ini, dataset dipartisi menjadi sepuluh bagian untuk proses pelatihan dan pengujian secara simultan, sehingga setiap data pernah berfungsi sebagai data uji.

Berdasarkan pengujian tersebut, diperoleh rata-rata akurasi senilai 66,75%. Apabila dikomparasikan dengan hasil uji sebelumnya, memang ditemukan sedikit deviasi penurunan akurasi. Namun, selisih tersebut masih tergolong wajar dan membuktikan bahwa model tidak terindikasi mengalami *overfitting* yang ekstrem. Stabilitas kinerja ini menunjukkan bahwa struktur Random Forest yang dirancang memiliki kapabilitas generalisasi yang mumpuni terhadap data baru. Hal ini menandakan model tidak sekadar unggul pada data latih, namun tetap konsisten saat diimplementasikan pada dataset yang belum pernah diproses sebelumnya.

4.3. Interpretasi Model (Feature Importance)

Selain menyediakan *output* prediksi, algoritma *Random Forest* juga memberikan kapabilitas untuk menelaah derajat kontribusi setiap variabel dalam mekanisme klasifikasi. Detail mengenai tingkat kepentingan fitur (*feature importance*) tersebut ditampilkan pada Gambar 3.



Gambar 3. Grafik Feature Importance

Berdasarkan Gambar 3, variabel usia saat diagnosis (*Age at Diagnosis*) tercatat sebagai fitur dengan tingkat kepentingan paling tinggi dalam memengaruhi prediksi model. Hal ini mengindikasikan bahwa umur pasien ketika pertama

kali terdiagnosis menjadi faktor paling dominan bagi kelangsungan hidup penderita kanker payudara. Di samping itu, variabel lain seperti *Nottingham Prognostic Index* serta ukuran tumor (*Tumor Size*) turut memberikan kontribusi yang berarti.

Hasil temuan ini selaras dengan prinsip medis yang menempatkan kondisi fisiologis serta derajat keparahan penyakit sebagai determinan utama dalam prognosis kanker payudara. Indikator seperti indeks prognostik dan dimensi tumor merepresentasikan sejauh mana perkembangan kanker telah terjadi, sedangkan usia cenderung berkorelasi dengan kemampuan imunitas tubuh serta efektivitas respons terhadap pengobatan.

Selain itu, variabel jumlah kelenjar getah bening yang terinfeksi juga memengaruhi *output* prediksi, yang mencerminkan intensitas penyebaran sel kanker pada tubuh subjek. Peningkatan nilai pada variabel tersebut menunjukkan risiko metastasis yang lebih besar, sehingga berpotensi menurunkan probabilitas bertahan hidup pasien.

Secara komprehensif, analisis ini membuktikan bahwa algoritma *Random Forest* tidak sekadar menghasilkan prediksi dengan akurasi yang mumpuni, namun juga mampu menyajikan wawasan klinis yang relevan mengenai faktor-faktor penentu prognosis pasien.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, algoritma *Random Forest* terbukti cukup efektif dalam memprediksi kelangsungan hidup pasien kanker payudara, ditunjukkan oleh nilai *accuracy* sebesar 71,04%, *precision* 72,31%, *recall* 81,50%, dan *F1-Score* 76,63%; tingginya nilai *recall* mengindikasikan kemampuan model yang baik dalam mengidentifikasi pasien yang bertahan hidup. Hasil validasi menggunakan *10-Fold Cross-Validation* menunjukkan rata-rata akurasi 66,75%, yang mengindikasikan kemampuan *generalisasi* model yang cukup stabil; sementara analisis *feature importance* mengungkap bahwa variabel *Age at Diagnosis*, *Nottingham Prognostic Index*, dan *Tumor Size* menjadi faktor dominan dalam menentukan prognosis pasien, sehingga model tidak hanya bersifat prediktif tetapi juga memberikan wawasan klinis yang bermakna. Namun demikian, masih terdapat kesalahan klasifikasi yang menunjukkan perlunya peningkatan *performa* model; oleh karena itu, untuk penelitian selanjutnya disarankan melakukan optimasi parameter, menerapkan teknik penyeimbangan data seperti SMOTE untuk mengatasi ketidakseimbangan kelas, membandingkan dengan algoritma lain seperti *XGBoost* atau *LightGBM*, serta menggunakan *dataset* yang lebih besar dan beragam untuk membuat hasil prediksi lebih akurat dan *robust*.

DAFTAR PUSTAKA

- [1] A. N. Giaquinto *et al.*, "Breast cancer statistics 2024," *CA. Cancer J. Clin.*, vol. 74, no. 6, pp. 477–495, 2024, doi: 10.3322/caac.21863.
- [2] M. Dasegowda, M. Brunda, M. Sajana, and K. M. Basalingappa, "Breast Cancer Basics: Understanding the Disease and Its Challenges," in *Nano Theragnostics in Breast Cancer: Advances, Challenges, and Future Prospects*, Singapore: Springer Nature Singapore, 2026, pp. 51–88.
- [3] B. Smolarz, A. Zadrozna-Nowak, and H. Romanowicz, "Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature)," *Cancers*, vol. 14, no. 10, p. 2569, May 2022, doi: 10.3390/cancers14102569.
- [4] S. R. Alimun, S. Rijal, I. M. Musa, R. Purnamasari, and F. Irsandy, "Analisis Faktor Risiko Kanker Payudara," *Fakumi Medical Journal*, vol. 4, no. 6, pp. 473–484, 2024.
- [5] E. I. Obeagu and G. U. Obeagu, "Breast cancer: A review of risk factors and diagnosis," *Medicine*, vol. 103, no. 3, p. e36905, 2024.
- [6] K. Suparna and L. M. K. K. S. Sari, "Kanker payudara: Diagnostik, faktor risiko, dan stadium," *Ganesha Medicina*, vol. 2, no. 1, pp. 42–48, 2022.
- [7] H. Xu and B. Xu, "Breast cancer: Epidemiology, risk factors and screening," *Chinese Journal of Cancer Research*, vol. 35, no. 6, pp. 565–583, 2023.
- [8] H. Abdul Rahman, S. N. N. Zaim, U. S. Suhaime, and A. A. Jamain, "Prognostic Factors Associated with Breast Cancer-Specific Survival from 1995 to 2022: A Systematic Review and Meta-Analysis of 1,386,663 Cases from 30 Countries," *Diseases*, vol. 12, no. 6, 2024, doi: 10.3390/diseases12060111.
- [9] R. N. Gurning, A. A. Sulaeman, and D. Afandi, "Breast Cancer Classification Using Naïve Bayes and Random Forest Algorithms," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 7, no. 3, pp. 920–933, Aug. 2025, doi: 10.47709/cnahpc.v7i3.6609.
- [10] U. Firdaus, A. Alfiah, and L. Mohdo, "Analisis Kinerja Decision Tree dan Random Forest Menggunakan Dataset Breast Cancer," *J. Pendidik. Sains Komput.*, vol. 6, no. 01, pp. 37–42, 2026, doi: 10.47709/jpsk.v6i01.7892.
- [11] Z. Javanmard *et al.*, "Artificial intelligence in breast cancer survival prediction: a comprehensive systematic review and meta-analysis," *Front. Oncol.*, vol. 14, no. January, pp. 1–16, 2024, doi: 10.3389/fonc.2024.1420328.
- [12] S. Z. Hamedi *et al.*, "Application of machine learning in breast cancer survival prediction using a multimethod approach," *Sci. Rep.*, vol. 14, no. 1, p. 30147, Dec. 2024, doi: 10.1038/s41598-024-81734-y.

- [13] S. M. Noman *et al.*, "Leveraging survival analysis and machine learning for accurate prediction of breast cancer recurrence and metastasis," *Sci. Rep.*, vol. 15, no. 1, pp. 1–16, 2025, doi: 10.1038/s41598-025-87622-3.
- [14] Y. J. Triadin, K. Kusriani, and K. Kusnawi, "KLASIFIKASI RANDOM FOREST TERHADAP DIAGNOSA PENYAKIT KANKER PAYUDARA BERDASARKAN STATUS KEGANASAN," *Tek. Teknol. Inf. dan Multimed.*, vol. 6, no. 1, pp. 129–134, Jun. 2025, doi: 10.46764/teknimedia.v6i1.259.
- [15] F. J. Kaunang, B. Hakim, F. Fraderic, S. Hartono, and A. K. Mulyanto, "Breast Cancer Detection using Decision Tree and Random Forest," *J. Appl. Informatics Comput.*, vol. 9, no. 2, pp. 302–309, Mar. 2025, doi: 10.30871/jaic.v9i2.9073.
- [16] Y. Jin, A. Lan, Y. Dai, L. Jiang, and S. Liu, "Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy," *Eur. J. Med. Res.*, vol. 28, no. 1, pp. 1–12, 2023, doi: 10.1186/s40001-023-01361-7.
- [17] X. Zhong, Y. Lin, W. Zhang, and Q. Bi, "Predicting diagnosis and survival of bone metastasis in breast cancer using machine learning," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-45438-z.
- [18] C. Li, M. Liu, J. Li, W. Wang, C. Feng, Y. Cai, F. Wu, X. Zhao, C. Du, Y. Zhang, Y. Wang, S. Zhang, and J. Qu, "Machine learning predicts the prognosis of breast cancer patients with initial bone metastases," *Frontiers in Public Health*, vol. 10, p. 1003976, 2022, doi: 10.3389/fpubh.2022.1003976