Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 5, No. 5, Mei 2025, Hal. 1407-1416
DOI: https://doi.org/10.52436/1.jpti. 802
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Optimasi Akurasi Prediksi Penyakit Kanker Payudara Menggunakan Metode Random Forest

Diky Misdiantoro[*1], Erliyan Redy Susanto[2]

[1]Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia, Indonesia
[2]Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia, Indonesia
Email: [1]diky_misdiantoro@teknokrat.ac.id, [2]erliyan.redy@teknokrat.ac.id

Abstrak

Kanker payudara merupakan ancaman serius bagi kesehatan wanita, dengan deteksi dini dan diagnosis akurat yang sangat penting untuk meningkatkan hasil pasien dan mengurangi kematian; namun, akurasi dan keandalan metode diagnosis saat ini masih perlu ditingkatkan. Penelitian ini bertujuan untuk mengoptimalkan metode *Random Forest* dalam prediksi kanker payudara, meningkatkan akurasi dan efisiensi model diagnosa. Menggunakan dataset Kaggle yang mencakup 569 pasien dengan 30 fitur karakteristik sel payudara, penelitian ini melalui proses pemuatan data, pra-pemrosesan (pembersihan dan normalisasi), pembagian data latih dan uji, pelatihan model Random Forest, prediksi, evaluasi model dengan metrik akurasi, presisi, recall, F1-score, dan AUC-ROC, serta visualisasi hasil. Hasilnya, penelitian mencapai akurasi 96%, peningkatan dari 95% yang dicapai sebelumnya, dengan model Random Forest yang dioptimalkan menunjukkan performa sangat baik dalam mengklasifikasikan kanker payudara, memiliki presisi 98%, recall 93%, dan F1-Score 95%, serta kemampuan klasifikasi sempurna dengan AUC 1.00. Model ini dapat menjadi alat yang efektif dalam diagnosis medis kanker payudara, meningkatkan hasil pasien, dan mengurangi kematian, sehingga penelitian ini merekomendasikan penggunaan metode Random Forest yang dioptimalkan sebagai alat prediksi kanker payudara yang efektif dan efisien.

Kata kunci: *Akurasi model, Diagnosis medis, Kanker payudara, Optimasi metode, Random forest.*

Optimization of Breast Cancer Disease Prediction Accuracy Using Random Forest Method

Abstract

Breast cancer poses a significant threat to women's health, with early detection and accurate diagnosis being crucial for improving patient outcomes and reducing mortality; however, the accuracy and reliability of current diagnostic methods still need to be enhanced. This study aims to optimize the Random Forest method for breast cancer prediction, thereby improving the accuracy and efficiency of the diagnostic model. Utilizing the Kaggle dataset, which includes 569 patients with 30 breast cell characteristic features, the study follows a process of data loading, pre-processing (cleaning and normalization), splitting into training and testing data, training the Random Forest model, making predictions, evaluating the model using accuracy, precision, recall, F1-score, and AUC-ROC metrics, and visualizing the results. The study achieved an accuracy of 96%, an improvement from the previous 95%, with the optimized Random Forest model demonstrating excellent performance in classifying breast cancer, achieving a precision of 98%, recall of 93%, and F1-Score of 95%, as well as perfect classification ability with an AUC of 1.00. This model can serve as an effective tool in breast cancer diagnostics, improving patient outcomes and reducing mortality, leading to the recommendation of the optimized Random Forest method as an effective and efficient breast cancer prediction tool.

Keywords: *Breast cancer, Model accuracy, Medical diagnosis, Method optimization,Random forest*

1. PENDAHULUAN

Kanker payudara merupakan salah satu bentuk kanker yang paling umum dialami oleh wanita di seluruh dunia, dan menjadi salah satu ancaman serius bagi kesehatan masyarakat .Penyakit ini tidak hanya mempengaruhi kesehatan fisik, tetapi juga memiliki dampak psikologis dan sosial yang signifikan bagi para penderitanya[1]. Deteksi dini dan diagnosis yang akurat sangat penting untuk meningkatkan hasil pada pasien dan mengurangi tingkat kematian[2]. Semakin dini kanker payudara dideteksi, semakin besar peluang untuk melakukan pengobatan yang efektif dan meningkatkan harapan hidup pasien[3]. Selama beberapa dekade terakhir, telah terjadi kemajuan yang signifikan dalam teknik pencitraan medis dan metode diagnosis, yang telah secara signifikan meningkatkan

kemampuan untuk mendeteksi kanker payudara pada tahap awal[4]. Namun, akurasi dan keandalan metode diagnosis ini tetap menjadi faktor kritis dalam pengelolaan pasien yang efektif[5]. Meskipun telah ada banyak kemajuan, masih ada celah dalam penelitian yang perlu diisi untuk meningkatkan akurasi dan keandalan diagnosis kanker payudara[6].

Dalam beberapa tahun terakhir, algoritma pembelajaran mesin telah muncul sebagai alat kuat untuk membantu diagnosis medis, terutama dalam analisis gambar medis dan data pasien[7]. Algoritma ini memiliki kemampuan untuk memproses dataset besar dan mengidentifikasi pola kompleks yang mungkin tidak segera terlihat bagi pengamat manusia[8]. Salah satu algoritma tersebut adalah klasifikasi *Random Forest*, yang telah menunjukkan janji dalam berbagai aplikasi medis karena ketahanannya dan kemampuannya dalam menangani data berdimensi tinggi[9]. Algoritma *Random Forest* telah menjadi fokus penelitian karena kemampuannya untuk meningkatkan akurasi diagnosis dan efisiensi dalam pengolahan data medis[10]. Penelitian sebelumnya telah menunjukkan bahwa algoritma ini dapat memberikan hasil yang lebih akurat dibandingkan dengan metode diagnosis tradisional, sehingga memiliki potensi besar untuk meningkatkan hasil pengobatan kanker payudara[11].

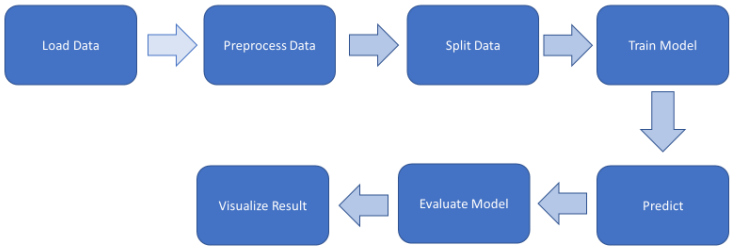
Algoritma *Random Forest* adalah metode pembelajaran berbasis ensemble yang membangun beberapa pohon keputusan selama pelatihan dan mengeluarkan mode kelas (klasifikasi) atau prediksi rata-rata (regresi) dari pohon individu[12]. Pendekatan ini tidak hanya meningkatkan akurasi prediksi tetapi juga meningkatkan ketahanan model terhadap *overfitting*[13]. Dalam konteks diagnosis kanker payudara, klasifikasi *Random Forest* telah digunakan untuk menganalisis berbagai fitur yang berasal dari gambar medis, seperti *mammogram* dan *scan ultrasonografi*, serta data klinis dari rekam medis pasien[14]. Penggunaan algoritma ini telah membuka kemungkinan baru dalam diagnosis kanker payudara, dengan kemampuan untuk mengidentifikasi pola yang sulit terdeteksi secara manual. Hal ini sangat penting dalam konteks kanker payudara, di mana keakuratan diagnosis dapat berdampak besar pada hasil pengobatan dan kualitas hidup pasien[15].

Penelitian sebelumnya telah menguji performa akurasi dari tujuh algoritma pembelajaran mesin yang berbeda untuk prediksi penyakit jantung. Dari tujuh algoritma yang diuji, empat algoritma, yaitu *Decision Tree*, *Naive Bayes*, *K-nearest neighbor*, dan *Logistic Regression*, mencapai tingkat akurasi yang sama yang cukup tinggi, yaitu 95,00%. Algoritma *Neural Network* menghasilkan sedikit kurang dengan akurasi 94,25%, sementara *Random Forest* dan *Support Vector Machines* memiliki akurasi yang sama yaitu 94,75%[11].

Penelitian ini bertujuan untuk mengoptimalkan metode *Random Forest* dalam prediksi kanker payudara dengan memperbaiki akurasi prediksi dan efisiensi model[16]. Penelitian sebelumnya telah mencapai akurasi 95%, dan penelitian ini berusaha untuk meningkatkannya menjadi 96%[11]. Dengan merampingkan *algoritma Random Forest*, studi ini berusaha untuk memenuhi kebutuhan kritis akan alat diagnostik yang lebih akurat dan dapat diandalkan dalam manajemen kanker payudara[17]. Penelitian ini sangat relevan karena membangun pengetahuan yang ada dan berkontribusi pada upaya berkelanjutan untuk meningkatkan deteksi dini dan diagnosis kanker payudara, yang pada akhirnya akan bermanfaat bagi perawatan dan hasil pasien[18]. Dengan meningkatkan akurasi dan efisiensi algoritma *Random Forest*, penelitian ini berharap dapat memberikan kontribusi signifikan dalam pengelolaan kanker payudara dan meningkatkan kualitas hidup para penderitanya[19].

2. METODE PENELITIAN

Dalam penelitian ini, tahapan studi dilakukan secara sistematis untuk memastikan bahwa model *Random Forest* dapat menghasilkan prediksi yang akurat terhadap penyakit kanker payudara. Setiap langkah dalam proses ini memiliki peran penting, mulai dari pemuatan data hingga visualisasi hasil, guna meningkatkan pemahaman terhadap performa model yang digunakan. Gambar 1 adalah diagram yang menggambarkan alur tahapan studi dalam penelitian ini :



Gambar 1. Tahapan Penelitian

a) Load Data

Tahap pertama dalam penelitian adalah memuat dataset dari file CSV yang berisi informasi tentang 569 pasien dengan 30 fitur yang mencerminkan karakteristik sel-sel payudara.

b) Preprocess Data

Setelah data dimuat, langkah selanjutnya adalah pra-pemrosesan data. Tahap ini melibatkan pembersihan data, normalisasi.

c) Split Data

Setelah data diproses, dataset dibagi menjadi dua bagian: data latih dan data uji. Ini dilakukan untuk memungkinkan evaluasi objektif dari performa model. Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk mengevaluasi performa model.

d) Train Model

Pada tahap ini, model Random Forest dilatih menggunakan data latih. Model belajar pola dan hubungan antara fitur-fitur dan label target (diagnosis kanker payudara) dengan membuat sejumlah pohon keputusan dan menggabungkan hasilnya,.

e) Predict

Setelah model dilatih, langkah berikutnya adalah untuk menggunakan model untuk membuat prediksi pada data uji. Hasil prediksi kemudian dievaluasi pada tahap selanjutnya.

f) Evaluate Model

Hasil prediksi dievaluasi menggunakan metrik-metrik seperti akurasi, presisi, recall, F1-score, dan AUC-ROC untuk menilai seberapa baik model bekerja.

g) Visualize Results

Terakhir, hasil evaluasi dan prediksi divisualisasikan untuk mempermudah interpretasi dan penyampaian temuan secara jelas dan informatif.

2.1 Sumber Data

Dalam penelitian ini, data yang digunakan untuk mengembangkan dan mengoptimalkan model prediksi penyakit kanker payudara menggunakan metode *Random Forest* diperoleh dari platform Kaggle. Dataset yang digunakan berasal dari proyek dengan judul *"Breast Cancer Wisconsin (Diagnostic)"*. Dataset ini merupakan salah satu dataset yang paling sering digunakan dalam penelitian terkait prediksi kanker payudara karena kualitas dan jumlah datanya yang cukup besar[20].

Dataset ini berisi informasi tentang 569 pasien dengan 30 fitur yang mencerminkan karakteristik sel-sel payudara yang telah diukur secara mikroskopis. Fitur-fitur tersebut termasuk radius, tekstur, perimeter, area, kehalusan, kompak, *concavity*, *concave points*, simetri, dan fraktal dimension dari inti sel payudara[21]. Setiap fitur memiliki nilai rata-rata, standar deviasi, dan "*worst*" atau nilai maksimum untuk setiap pasien. Selain itu, dataset juga mencakup label target yang menunjukkan apakah pasien memiliki kanker payudara (*malignant*) atau tidak (*benign*).

Penggunaan dataset ini dalam penelitian ini didasarkan pada beberapa alasan. Pertama, dataset ini telah banyak digunakan dalam penelitian sebelumnya, sehingga hasil penelitian ini dapat dibandingkan dengan studi-studi lain yang menggunakan dataset yang sama[22]. Kedua, kualitas dan variasi data dalam dataset ini memungkinkan pengujian dan validasi model dengan lebih baik. Ketiga, dataset ini tersedia secara publik dan dapat diakses oleh peneliti lain, sehingga meningkatkan transparansi dan reproduktibilitas penelitian.

Sebagai langkah awal dalam proses pengembangan model, dataset ini dimuat dan diproses menggunakan teknik pra-pemrosesan data seperti membersihkan data, normalisasi. Kemudian, dataset dibagi menjadi dua bagian: data latih dan data uji, untuk memungkinkan evaluasi objektif dari performa model. Dengan demikian, dataset ini menjadi dasar yang kuat untuk mengoptimalkan model prediksi kanker payudara menggunakan metode *Random Forest* dalam penelitian ini[23][24].

2.2 Variable

Dalam dataset yang digunakan untuk penelitian ini, terdapat total 32 variabel. Variabel pertama adalah "id", yang berfungsi sebagai pengidentifikasi unik setiap sampel data. Variabel kedua adalah "diagnosis" (atau "diagnosis"), yang merupakan label kelas yang menunjukkan apakah sampel tersebut termasuk ke dalam kategori ganas (M) atau jinak (B). Sisanya, 30 variabel lainnya adalah fitur-fitur yang diukur dari citra sel-sel payudara pasien, yang meliputi radius rata-rata ("*radius_mean*"), tekstur rata-rata ("*texture_mean*"), keliling rata-rata ("*perimeter_mean*"), luas rata-rata ("*area_mean*"), kehalusan rata-rata ("*smoothness_mean*"), ketidaksimetrisan

rata-rata ("*compactness_mean*"), simetri rata-rata ("*symmetry_mean*"), dan dimensi fraktal rata-rata ("*fractal_dimension_mean*"). Setiap fitur ini juga memiliki versi standar deviasi ("_se") dan versi terburuk ("*_worst*"), sehingga menghasilkan total 30 fitur. Dengan demikian, dataset ini mencakup 32 variabel yang akan digunakan dalam analisis dan pembangunan model prediksi menggunakan metode Random Forest.

Tabel 1. Variable dalam dataset

Nama variable	keterangan
id	ID unik untuk setiap pasien.
diagnosis	Label klasifikasi kanker (M: Malignant, B: Benign).
radius_mean	Rata-rata radius dari area tumor.
texture_mean	Rata-rata tekstur dari area tumor.
perimeter_mean	Rata-rata perimeter dari area tumor.
area_mean	Rata-rata luas area tumor.
smoothness_mean	Rata-rata smoothness dari area tumor
compactness_mean	Rata-rata compactness dari area tumor.
concavity_mean	Rata-rata concavity dari area tumor.
concave points_mean	Rata-rata jumlah titik cekung pada kontur tumor
symmetry_mean	Rata-rata simetri dari area tumor.
fractal_dimension_mean	Rata-rata dimensi fraktal dari area tumor.
radius_se	Standar error dari radius tumor.
texture_se	Standar error dari tekstur tumor.
perimeter_se	Standar error dari perimeter tumor.
area_se	Standar error dari luas area tumor
smoothness_se	Standar error dari smoothness tumor.
compactness_se	Standar error dari compactness tumor.
concavity_se	Standar error dari concavity tumor.
concave points_se	Standar error dari jumlah titik cekung pada kontur tumor
symmetry_se	Standar error dari simetri tumor.
fractal_dimension_se	Standar error dari dimensi fraktal tumor.
radius_worst	Radius terburuk (nilai maksimum) dari area tumor.
texture_worst	Tekstur terburuk (nilai maksimum) dari area tumor
perimeter_worst	Perimeter terburuk (nilai maksimum) dari area tumor.
area_worst	Luas area terburuk (nilai maksimum) dari area tumor.
smoothness_worst	Smoothness terburuk (nilai maksimum) dari area tumor.
compactness_worst	Compactness terburuk (nilai maksimum) dari area tumor.
concavity_worst	Concavity terburuk (nilai maksimum) dari area tumor.
concave points_worst	Jumlah titik cekung terburuk (nilai maksimum) dari kontur tumor
symmetry_worst	Simetri terburuk (nilai maksimum) dari area tumor.
fractal_dimension_worst	Dimensi fraktal terburuk (nilai maksimum) dari area tumor

2.3 Random Forest

Random Forest adalah algoritma pembelajaran mesin berbasis *ensemble learning* yang digunakan untuk tugas klasifikasi dan regresi[25]. Algoritma ini merupakan pengembangan dari *Decision Tree* dengan membangun

banyak pohon untuk meningkatkan akurasi dan ketahanan terhadap *overfitting*[26]. Pada beberapa penelitian dahulu, random forest diketahui memiliki akurasi lebih baik dibandingkan yang lain[27][28][29]. Dalam penelitian ini, *Random Forest* digunakan untuk mengklasifikasikan tumor sebagai *malignant* (ganas) atau *benign* (jinak) berdasarkan dataset kanker payudara[30]. Beberapa parameter penting dalam *Random Forest* adalah *n_estimators*, yaitu jumlah pohon dalam hutan, serta *max_depth*, yang menentukan kedalaman maksimum setiap pohon. Parameter *random_state* digunakan untuk memastikan hasil eksperimen yang konsisten. Dalam penelitian ini, *random_state* diatur ke 42 untuk menjaga reprodusibilitas. Prediksi akhir dari *Random Forest* ditentukan berdasarkan *majority voting* dari semua pohon keputusan yang ada[31].

$$\hat{y} = mode(T1(X), T2(X), \ldots, TN(X)) \quad (1)$$

Di mana Ti(X) adalah prediksi dari pohon keputusan ke-i, dan mode adalah nilai yang paling sering muncul. Evaluasi model dilakukan menggunakan metrik akurasi, precision, recall, F1-score, dan ROC-AUC untuk mengukur kinerja klasifikasi.

2.3. Confusion Matrix

Confusion Matrix yang diberikan menunjukkan hasil prediksi model *Random Forest* dalam mengklasifikasikan kasus kanker payudara menjadi dua kelas: *Benign* (jinak) dan *Malignant* (ganas).

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad (2)$$

Keterangan :
TP : True Positives.
FP : False Positives.
FN : False Negatives.
TN : True Negatives.

Metrik Evaluasi dari *Confusion Matrix* :

a) ***Accuracy*** *: Mengukur seberapa sering model membuat prediksi yang benar.*

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

b) ***Precision :*** *Mengukur proporsi prediksi positif yang benar-benar positif.*

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

c) ***Recall :*** *Mengukur kemampuan model untuk mendeteksi semua kasus positif yang sebenarnya.*

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

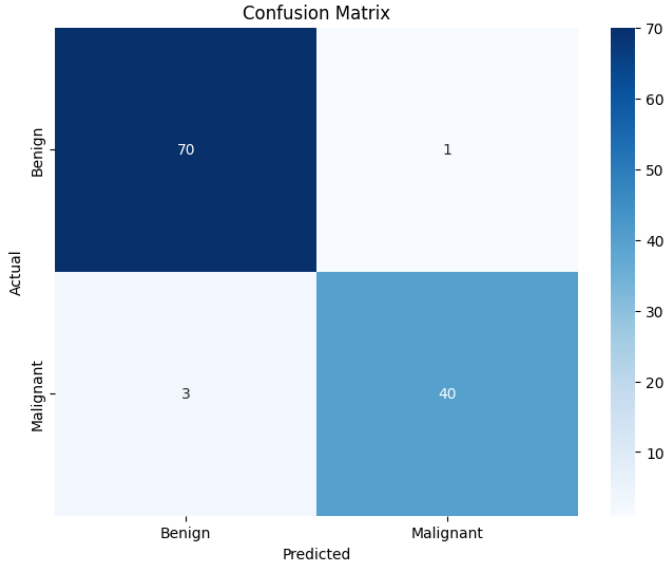
d) ***F1-Score :*** *Menggabungkan Precision dan Recall dalam satu metrik dengan mengambil rata-rata harmonisnya.*

$$f1 - Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision\ x\ Recall} \quad (6)$$

3. HASIL DAN PEMBAHASAN

3.1 Metode Random Forest

Dalam penelitian ini, kami menggunakan *confusion matrix* untuk menilai kinerja model prediksi penyakit kanker payudara yang dibangun menggunakan metode *random forest*. Gambar 2 menunjukkan hasil dari *confusion matrix* tersebut, yang menggambarkan jumlah kasus yang diprediksi dengan benar dan salah.

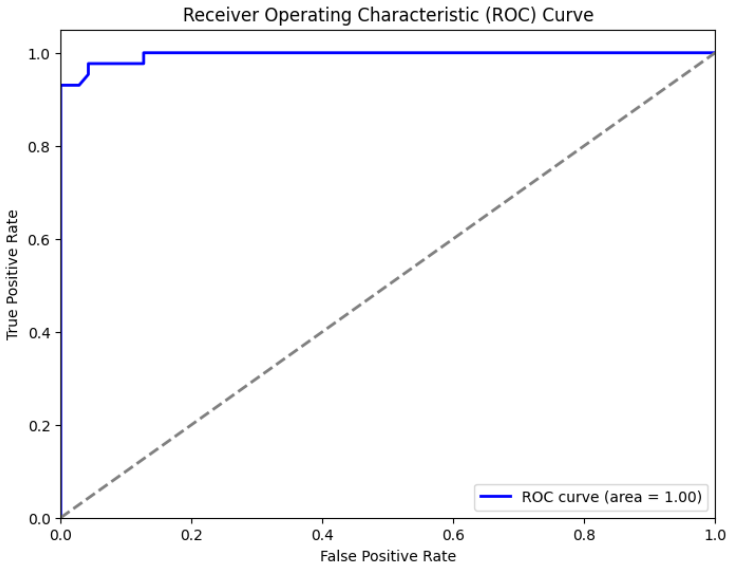


Gambar 1. Confusion matrix menggunakan *random forest*

Gambar 2 menunjukkan *confusion matrix* dari model *random forest* yang digunakan untuk prediksi penyakit kanker payudara. Dari matriks ini, terlihat bahwa model memiliki performa yang cukup baik dalam mengidentifikasi kanker payudara yang tidak berbahaya (*benign*) dengan akurasi 70 dari total 71 kasus yang sebenarnya tidak berbahaya. Hal ini menunjukkan tingkat kebenaran yang tinggi (*True Negative Rate*) dan menandakan bahwa model cukup efisien dalam mengelkan kasus yang tidak berbahaya. Namun, ada 1 kasus yang sebenarnya tidak berbahaya yang diprediksi salah sebagai berbahaya (*malignant*), yang menunjukkan tingkat kecenderungan model untuk memberikan hasil yang salah (*False Positive Rate*) yang cukup rendah.

Selain itu, dalam hal kasus kanker payudara yang sebenarnya berbahaya, model berhasil mengidentifikasi 40 dari total 43 kasus, yang menunjukkan tingkat sensitivitas yang baik (*True Positive Rate*). Namun, ada 3 kasus yang sebenarnya berbahaya yang diprediksi salah sebagai tidak berbahaya, yang merupakan area yang bisa dioptimalkan lebih lanjut dalam model. Dalam analisis ini, *confusion matrix* menjadi alat yang berharga untuk menilai kinerja model dan mengidentifikasi area yang perlu perbaikan, terutama dalam mengurangi jumlah *false negatives* untuk meningkatkan keselamatan dan efektivitas prediksi kanker payudara. Dengan demikian, penelitian ini menunjukkan potensi metode *random forest* dalam prediksi penyakit kanker payudara, sambil juga menunjukkan area yang perlu diperbaiki untuk meningkatkan akurasi prediksi.

Pada penelitian ini, kinerja model prediksi sangat penting. ROC Curve dapat memberikan wawasan dalam performa model klasifikasi. Berikut adalah ROC Curve dari model Random Forest untuk prediksi penyakit kanker payudara :



Gambar 2. ROC Curve menggunakan *random forest*

Gambar 3 menunjukkan ROC Curve yang dihasilkan oleh model *Random Forest* untuk prediksi penyakit kanker payudara. ROC Curve adalah alat analisis yang digunakan untuk menilai kinerja model klasifikasi dalam menentukan seberapa baik model tersebut mampu membedakan antara kelas positif dan negatif.

Pada gambar 3, sumbu x mewakili FPR atau rasio dari contoh negatif yang diprediksi sebagai positif, sedangkan sumbu y mewakili TPR atau rasio dari contoh positif yang diprediksi dengan benar. Garis putus-putus (*dashed line*) pada gambar ini mewakili performa acak dari model, di mana FPR dan TPR sama.

ROC Curve dari model *Random Forest* ini sangat dekat dengan titik (0,1), yang menunjukkan bahwa model mampu mengidentifikasi semua contoh positif dengan benar tanpa adanya contoh negatif yang dianggap positif. AUC dari ROC Curve ini adalah 1.00, yang merupakan nilai terbaik dan menunjukkan bahwa kinerja model sangat baik. Nilai AUC ini menunjukkan bahwa model mampu memisahkan antara kelas positif dan negatif dengan sangat baik dan memiliki performa yang optimal dalam prediksi penyakit kanker payudara.

Dalam penelitian sebelumnya menggunakan 7 metode pembelajaran mesin dalam prediksi penyakit jantung yang teridiri dari *neural network, decision tree, naïve bayes, k-nearest neighbor, logistic regresion, random forest, dan support vector machines* didapatkan hasil akurasi seperti pada Tabel 2.

Tabel 2 Performa berbagai model ML pada riset sebelumnya[11]

Nama Algoritma	**Nilai Akurasi**
Neural Network	94,25%
Decision Tree	95,00%
Naïve Bayes	95,00%
K-nearest neighbor	95,00%
Logistic regression	95,00%
Random Forest	94,75 %
Support Vector Machine	94,75 %

Berdasarkan pada Tabel 2, *algoritma logistic regresion, decision tree, naïve bayes dan k-nearest neighbor* memiliki nilai akurasi yang sama tinggi yaitu sebesar 95,00%. sedangkan algoritma *random forest* dan *support vector machines* memiliki akurasi yang sama yaitu 94,75% dan *algoritma neural network* memperoleh akurasi 94,25%[11]. Pembagian data dengan uji 80 : 20 algoritma *Random Forest* didapatkan hasil seperti Tabel 3.

Tabel 3 Pembagian Data 80% : 20%

Confusion Matrix	**Prescision**	**Recall**	**F1Score**	**Accuracy**
Random Forest	0.98	0.93	0.95	0.96

Tabel 3 menunjukkan hasil evaluasi model *Random Forest* yang dilakukan dengan pembagian data sebesar 80% untuk pelatihan dan 20% untuk pengujian. Model memiliki nilai presisi sebesar 0,98, yang menunjukkan tingkat keakuratan prediksi positif dari keseluruhan prediksi positif yang dilakukan oleh model, nilai recall mencapai 0,93, menunjukkan kemampuan model dalam mendeteksi semua data positif yang benar, model memiliki nilai F1 Score sebesar 0,95, yang merupakan rata-rata harmonis dari presisi dan recall, menandakan keseimbangan yang baik antara keduanya, tingkat akurasi model sebesar 0,96, yang menunjukkan persentase prediksi model yang benar terhadap keseluruhan data uji.

3.2 Diskusi

Penelitian yang dilakukan telah mencapai tujuan utamanya, yaitu mengoptimalkan metode *Random Forest* dalam prediksi kanker payudara dengan memperbaiki akurasi prediksi dan efisiensi model. Hasil penelitian menunjukkan bahwa model *Random Forest* yang dioptimalkan mampu mencapai akurasi sebesar 96%, yang merupakan peningkatan dari penelitian sebelumnya yang mencapai 95%[11]. Hal ini menunjukkan bahwa penggunaan teknik pra-pemrosesan data yang tepat, seperti pembersihan data, normalisasi, serta pengaturan parameter yang optimal pada model *Random Forest*, telah berkontribusi secara signifikan dalam meningkatkan akurasi prediksi.

Dari hasil *Confusion Matrix*, terlihat bahwa model *Random Forest* yang dioptimalkan mampu mengklasifikasikan kasus kanker payudara dengan sangat baik. Dengan jumlah *True Positive* sebanyak 70 dan *True Negative* sebanyak 40, serta hanya 1 *False Positive* dan 3 *False Negative*, model ini menunjukkan kemampuan yang sangat baik dalam membedakan antara kelas *Benign* dan *Malignant*. Presisi sebesar 98% dan recall sebesar 93% menunjukkan bahwa model tidak hanya akurat dalam memprediksi kasus positif, tetapi juga efektif dalam mendeteksi semua kasus positif yang sebenarnya. Nilai F1-Score sebesar 95% mengindikasikan keseimbangan yang baik antara presisi dan recall, yang sangat penting dalam konteks medis untuk meminimalkan risiko kesalahan diagnosis.

ROC Curve yang dihasilkan dari penelitian ini menunjukkan bahwa model *Random Forest* memiliki kemampuan klasifikasi yang sempurna dengan AUC sebesar 1.00. Hal ini menunjukkan bahwa model mampu membedakan antara kelas *Benign* dan *Malignant* dengan sangat baik tanpa ada kesalahan klasifikasi. TPR atau *Sensitivity* mencapai 1.00, yang berarti model mampu mendeteksi semua kasus *Malignant*, sementara FPR adalah 0.00, yang berarti tidak ada kasus *Benign* yang salah diklasifikasikan sebagai *Malignant*.

Hasil penelitian ini menunjukkan bahwa metode *Random Forest* yang dioptimalkan dapat menjadi alat yang sangat efektif dalam membantu diagnosis medis. Dengan akurasi yang tinggi dan kemampuan klasifikasi yang sangat baik, model ini dapat memberikan kontribusi signifikan dalam meningkatkan hasil pada pasien dan mengurangi tingkat kematian akibat kanker payudara. Selain itu, penggunaan dataset yang berkualitas tinggi dan teknik pra-pemrosesan data yang tepat telah membantu dalam mencapai hasil yang optimal. Oleh karena itu, penelitian ini merekomendasikan penggunaan metode *Random Forest* yang dioptimalkan sebagai salah satu metode prediksi kanker payudara yang efektif dan efisien.

4. KESIMPULAN

Penelitian ini telah mengoptimalkan metode Random Forest untuk prediksi kanker payudara menggunakan dataset Kaggle yang mencakup 569 pasien dengan 30 fitur karakteristik sel payudara, mencapai akurasi 96%, presisi 98%, recall 93%, dan F1-Score 95%, serta AUC-ROC sempurna sebesar 1.00. Meskipun hasilnya sangat baik, penelitian ini memiliki keterbatasan, seperti dataset yang relatif kecil yang mungkin tidak mencakup variasi yang cukup besar untuk menggeneralisasi hasil ke populasi yang lebih luas, serta keterbatasan model dalam menangani data yang lebih kompleks. Untuk penelitian selanjutnya, disarankan untuk mengintegrasikan teknik *deep learning*, seperti *Convolutional Neural Networks* (CNN), yang sangat efektif dalam mengolah data gambar medis dan dapat meningkatkan akurasi prediksi. Selain itu, melakukan uji coba pada dataset yang lebih besar dan bervariasi serta mengembangkan model hibrida yang menggabungkan Random Forest dengan teknik lain dapat meningkatkan performa model secara keseluruhan.

DAFTAR PUSTAKA

[1] R. Mehrotra and K. Yadav, "Breast cancer in India: Present scenario and the challenges ahead," *World J Clin Oncol*, vol. 13, no. 3, pp. 209–218, Mar. 2022, doi: 10.5306/wjco.v13.i3.209.

[2] D. Crosby *et al.*, "Early detection of cancer," *Science (1979)*, vol. 375, no. 6586, Mar. 2022, doi: 10.1126/science.aay9040.

[3] R. A. Mas'ud and Junta Zeniarja, "Optimasi Convolutional Neural Networks untuk Deteksi Kanker Payudara menggunakan Arsitektur DenseNet," *Edumatic: Jurnal Pendidikan Informatika*, vol. 8, no. 1, pp. 310–318, Jun. 2024, doi: 10.29408/edumatic.v8i1.25883.

[4] A. Pulumati, A. Pulumati, B. S. Dwarakanath, A. Verma, and R. V. L. Papineni, "Technological advancements in cancer diagnostics: Improvements and limitations," *Cancer Rep*, vol. 6, no. 2, Feb. 2023, doi: 10.1002/cnr2.1764.

[5] B. Hunter, S. Hindocha, and R. W. Lee, "The Role of Artificial Intelligence in Early Cancer Diagnosis," *Cancers (Basel)*, vol. 14, no. 6, p. 1524, Mar. 2022, doi: 10.3390/cancers14061524.

[6] S. Hussain *et al.*, "Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review," *Biomed Res Int*, vol. 2022, no. 1, Jan. 2022, doi: 10.1155/2022/5164970.

[7] R. Oktafiani, A. Hermawan, and D. Avianto, "Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning," *Jurnal Sains dan Informatika*, pp. 19–28, Jun. 2023, doi: 10.34128/jsi.v9i1.622.

[8] M. Rana and M. Bhushan, "Machine learning and deep learning approach for medical image analysis: diagnosis to detection," *Multimed Tools Appl*, vol. 82, no. 17, pp. 26731–26769, Jul. 2023, doi: 10.1007/s11042-022-14305-w.

[9] A. R. Inture, B. S. Sai Nadh, A. Sha, S. Abhishek, A. T, and T. V. Mullapudi, "Leveraging Random Forests for Ovarian Cancer Detection and Precision Prediction," in *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, Nov. 2023, pp. 910–915. doi: 10.1109/ICECA58529.2023.10394819.

[10] M. Song, H. Jung, S. Lee, D. Kim, and M. Ahn, "Diagnostic Classification and Biomarker Identification of Alzheimer's Disease with Random Forest Algorithm," *Brain Sci*, vol. 11, no. 4, p. 453, Apr. 2021, doi: 10.3390/brainsci11040453.

[11] N. R. Muntiari and K. H. Hanif, "Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning," *Jurnal Ilmu Komputer dan Teknologi*, vol. 3, no. 1, pp. 1–6, May 2022, doi: 10.35960/ikomti.v3i1.766.

[12] J. Wang, X. Sun, Q. Cheng, and Q. Cui, "An innovative random forest-based nonlinear ensemble paradigm of improved feature extraction and deep learning for carbon price forecasting," *Science of The Total Environment*, vol. 762, p. 143099, Mar. 2021, doi: 10.1016/j.scitotenv.2020.143099.

[13] M. A. Hassan, H. Salem, N. Bailek, and O. Kisi, "Random Forest Ensemble-Based Predictions of On-Road Vehicular Emissions and Fuel Consumption in Developing Urban Areas," *Sustainability*, vol. 15, no. 2, p. 1503, Jan. 2023, doi: 10.3390/su15021503.

[14] S. S, S. A, and A. S, "Ultrasound Image Analysis in Breast Cancer: A Comparative Study of Decision Trees and Random Forests," in *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, Dec. 2024, pp. 1024–1029. doi: 10.1109/CICN63059.2024.10847499.

[15] S. M. Shah, R. A. Khan, S. Arif, and U. Sajid, "Artificial intelligence for breast cancer analysis: Trends & directions," *Comput Biol Med*, vol. 142, p. 105221, Mar. 2022, doi: 10.1016/j.compbiomed.2022.105221.

[16] S. N. S., "Prediction of Breast Cancer Through Random Forest," *Curr Med Imaging Rev*, vol. 19, no. 10, Sep. 2023, doi: 10.2174/1573405618666220930150625.

[17] S. Yalavarthi, S. S. Makkapati, H. Murari, K. S. Balamurugan, and P. Rajendran, "Advanced Breast Cancer Diagnostics through a Comparative Analysis of SVM, Random Forests, and Neural Networks in MRI Image Analysis," in *2024 Asian Conference on Communication and Networks (ASIANComNet)*, IEEE, Oct. 2024, pp. 1–7. doi: 10.1109/ASIANComNet63184.2024.10811015.

[18] N. M. ud din, R. A. Dar, M. Rasool, and A. Assad, "Breast cancer detection using deep learning: Datasets, methods, and challenges ahead," *Comput Biol Med*, vol. 149, p. 106073, Oct. 2022, doi: 10.1016/j.compbiomed.2022.106073.

[19] D. Kang *et al.*, "Prediction Model for Postoperative Quality of Life Among Breast Cancer Survivors Along the Survivorship Trajectory From Pretreatment to 5 Years: Machine Learning–Based Analysis," *JMIR Public Health Surveill*, vol. 9, p. e45212, Aug. 2023, doi: 10.2196/45212.

[20] M. H. Alshayeji, H. Ellethy, S. Abed, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomed Signal Process Control*, vol. 71, p. 103141, Jan. 2022, doi: 10.1016/j.bspc.2021.103141.

[21] M. R. Darbandi, M. Darbandi, S. Darbandi, I. Bado, M. Hadizadeh, and H. R. Khorram Khorshid, "Artificial intelligence breakthroughs in pioneering early diagnosis and precision treatment of breast cancer: A multimethod study," *Eur J Cancer*, vol. 209, p. 114227, Sep. 2024, doi: 10.1016/j.ejca.2024.114227.

[22] W. T. Mohammad, R. Teete, H. Al-Aaraj, Y. S. Y. Rubbai, and M. M. Arabyat, "Diagnosis of Breast Cancer Pathology on the Wisconsin Dataset with the Help of Data Mining Classification and Clustering Techniques," *Appl Bionics Biomech*, vol. 2022, pp. 1–9, Apr. 2022, doi: 10.1155/2022/6187275.

[23] D. Benaya, "Implementasi Random Forest dalam Klasifikasi Kanker Paru-Paru," *JOINTER : Journal of Informatics Engineering*, vol. 5, no. 01, pp. 27–31, Jun. 2024, doi: 10.53682/jointer.v5i01.331.

[24] J. Li, J. Shi, J. Chen, Z. Du, and L. Huang, "Self-attention random forest for breast cancer image classification," *Front Oncol*, vol. 13, Feb. 2023, doi: 10.3389/fonc.2023.1043463.

[25] Cecep Wahyu Cahyana and Akhsin Nurlayli, "Analisis Performa Logistic Regression, Naïve Bayes, dan Random Forest sebagai Algoritma Pendeteksi Kanker Payudara," *INSERT : Information System and Emerging Technology Journal*, vol. 4, no. 1, pp. 51–64, Jun. 2023, doi: 10.23887/insert.v4i1.62362.

[26] G. Shanmugasundar, M. Vanitha, R. Čep, V. Kumar, K. Kalita, and M. Ramachandran, "A Comparative Study of Linear, Random Forest and AdaBoost Regressions for Modeling Non-Traditional Machining," *Processes*, vol. 9, no. 11, p. 2015, Nov. 2021, doi: 10.3390/pr9112015.

[27] S. Asadi, S. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," *J Biomed Inform*, vol. 115, p. 103690, Mar. 2021, doi: 10.1016/j.jbi.2021.103690.

[28] R. Zhu, Y. Wang, J.-X. Liu, and L.-Y. Dai, "IPCARF: improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier," *BMC Bioinformatics*, vol. 22, no. 1, p. 175, Dec. 2021, doi: 10.1186/s12859-021-04104-9.

[29] M. Pal and S. Parija, "Prediction of Heart Diseases using Random Forest," *J Phys Conf Ser*, vol. 1817, no. 1, p. 012009, Mar. 2021, doi: 10.1088/1742-6596/1817/1/012009.

[30] Y. Amethiya, P. Pipariya, S. Patel, and M. Shah, "Comparative analysis of breast cancer detection using machine learning and biosensors," *Intelligent Medicine*, vol. 2, no. 2, pp. 69–81, May 2022, doi: 10.1016/j.imed.2021.08.004.

[31] M. Aria, C. Cuccurullo, and A. Gnasso, "A comparison among interpretative proposals for Random Forests," *Machine Learning with Applications*, vol. 6, p. 100094, Dec. 2021, doi: 10.1016/j.mlwa.2021.100094.