

# Analisis Perbandingan Algoritma K-Nearest Neighbor (K-NN) dan Random Forest dalam Klasifikasi Kanker Payudara

Mahattir Rahman Muhammad<sup>1\*</sup>, Ahmad Homaidi<sup>2</sup>, Syarif Aminul Khoiri<sup>3</sup>

<sup>1,2,3</sup>Teknologi Informasi, Universitas Ibrahimy, Situbondo, Jawa Timur

\*Korespondensi penulis. E-mail: [mahattirahmad@gmail.com](mailto:mahattirahmad@gmail.com), Telp: 081335443430

## Abstrak

Kanker payudara merupakan salah satu penyakit yang memiliki tingkat kejadian dan kematian yang tinggi pada perempuan di seluruh dunia. Deteksi dini yang akurat sangat penting untuk meningkatkan peluang kesembuhan pasien dan mendukung pengambilan keputusan medis. Penelitian ini bertujuan untuk menganalisis dan membandingkan kinerja algoritma K-Nearest Neighbor (K-NN) dan Random Forest dalam klasifikasi kanker payudara menggunakan Breast Cancer Wisconsin (Diagnostic) Dataset. Penelitian menggunakan metode Cross Industry Standard Process for Data Mining (CRISP-DM) yang meliputi tahapan Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Dataset yang digunakan terdiri dari 569 data dengan 30 atribut numerik dan satu atribut target diagnosis. Data dibagi menjadi data latih sebesar 80% dan data uji sebesar 20%, serta dilakukan normalisasi menggunakan Min-Max Scaling. Evaluasi model dilakukan menggunakan metrik accuracy, precision, recall, F1-score, dan Area Under Curve (AUC). Hasil penelitian menunjukkan bahwa nilai k terbaik pada algoritma K-NN diperoleh pada k=5 dengan accuracy sebesar 96,49%, precision sebesar 100%, recall sebesar 90,48%, F1-score sebesar 95,00%, dan AUC sebesar 99,62%. Sementara itu, algoritma Random Forest memperoleh accuracy sebesar 97,37%, precision sebesar 100%, recall sebesar 92,86%, F1-score sebesar 96,30%, dan AUC sebesar 99,29%. Berdasarkan hasil evaluasi, Random Forest menunjukkan performa yang lebih baik dibandingkan K-NN karena memiliki nilai accuracy, recall, dan F1-score yang lebih tinggi. Oleh karena itu, Random Forest dapat menjadi algoritma yang lebih optimal untuk klasifikasi kanker payudara pada Breast Cancer Wisconsin (Diagnostic) Dataset.

Kata kunci: Kanker Payudara, Klasifikasi, Machine Learning, K-Nearest Neighbor, Random Forest, CRISP-DM.

## Abstract

*Breast cancer is a disease with a high incidence and mortality rate in women worldwide. Accurate early detection is crucial to improve the chances of patient recovery and support medical decision-making. This study aims to analyze and compare the performance of the K-Nearest Neighbor (K-NN) and Random Forest algorithms in breast cancer classification using the Breast Cancer Wisconsin (Diagnostic) Dataset. The study uses the Cross Industry Standard Process for Data Mining (CRISP-DM) method, which includes the stages of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The dataset used consists of 569 data with 30 numeric attributes and one diagnostic target attribute. The data is divided into 80% training data and 20% testing data, and normalization is performed using Min-Max Scaling. Model evaluation is carried out using accuracy, precision, recall, F1-score, and Area Under Curve (AUC) metrics. The results showed that the best k value in the K-NN algorithm was obtained at k = 5 with an accuracy of 96.49%, precision of 100%, recall of 90.48%, F1-score of 95.00%, and AUC of 99.62%. Meanwhile, the Random Forest algorithm obtained an accuracy of 97.37%, precision of 100%, recall of 92.86%, F1-score of 96.30%, and AUC of 99.29%. Based on the evaluation results, Random Forest showed better performance than K-NN because it had higher accuracy, recall, and F1-score values. Therefore, Random Forest can be a more optimal algorithm for breast cancer classification in the Breast Cancer Wisconsin (Diagnostic) Dataset.*

Keyword: Breast Cancer, Classification, Machine Learning, K-Nearest Neighbor, Random Forest, CRISP-DM.

## PENDAHULUAN

Kanker payudara merupakan salah satu jenis kanker yang paling banyak dialami oleh perempuan di seluruh dunia dan menjadi penyebab utama kematian akibat kanker (Riadinni Alita, Uum Safari, Nova Riani, Fariha Khoerunisa 2020). Menurut berbagai laporan kesehatan global, angka kejadian kanker payudara terus mengalami peningkatan setiap tahunnya sehingga menjadi

---

permasalahan serius dalam bidang kesehatan masyarakat. Tingginya angka mortalitas akibat kanker payudara sering kali disebabkan oleh keterlambatan diagnosis dan penanganan yang tidak dilakukan pada stadium awal. Oleh karena itu, deteksi dini dan diagnosis yang akurat menjadi faktor penting dalam meningkatkan peluang kesembuhan pasien serta menurunkan tingkat kematian akibat penyakit tersebut(Eka mei dianata, Lie Liana 2025).

Perkembangan teknologi informasi dan kecerdasan buatan (Artificial Intelligence) telah memberikan kontribusi yang signifikan dalam bidang kesehatan, khususnya dalam proses diagnosis penyakit(Idris et al. 2025). Salah satu cabang kecerdasan buatan yang banyak digunakan adalah machine learning. Machine learning memungkinkan sistem komputer untuk mempelajari pola dari data historis dan menghasilkan prediksi atau klasifikasi secara otomatis tanpa diprogram secara eksplisit(Tia Siti Nurazizah 2025). Dalam bidang medis, penerapan machine learning telah digunakan untuk membantu tenaga kesehatan dalam mengidentifikasi berbagai penyakit, termasuk kanker payudara, berdasarkan karakteristik data klinis pasien(Hulaifah Al Abrori dan Subhiyakto 2025).

Data mining sebagai bagian dari machine learning memiliki kemampuan untuk mengekstraksi pola dan pengetahuan tersembunyi dari kumpulan data dalam jumlah besar. Melalui teknik klasifikasi, data mining dapat digunakan untuk membedakan kondisi pasien yang terdiagnosis kanker payudara ganas (malignant) dan jinak (benign). Salah satu dataset yang paling sering digunakan dalam penelitian klasifikasi kanker payudara adalah Breast Cancer Wisconsin (Diagnostic) Dataset. Dataset ini berisi berbagai atribut hasil pemeriksaan sitologi yang dapat digunakan untuk memprediksi kondisi kanker payudara dengan tingkat akurasi yang tinggi.

Berbagai algoritma klasifikasi telah diterapkan dalam penelitian kanker payudara, seperti Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Artificial Neural Network, Random Forest, dan K-Nearest Neighbor (K-NN). Setiap algoritma memiliki karakteristik, kelebihan, dan kelemahan yang berbeda dalam menangani data klasifikasi. K-Nearest Neighbor (K-NN) merupakan algoritma berbasis kemiripan data yang bekerja dengan mengelompokkan objek berdasarkan kedekatan jarak terhadap sejumlah tetangga terdekat. Algoritma ini dikenal sederhana, mudah diimplementasikan, serta memiliki performa yang baik pada berbagai kasus klasifikasi(Malik Namus Akbar 2024). Namun, performa K-NN sangat dipengaruhi oleh pemilihan nilai parameter k dan skala data yang digunakan.

Di sisi lain, Random Forest merupakan algoritma berbasis ensemble learning yang menggabungkan sejumlah pohon keputusan (decision tree) untuk menghasilkan prediksi yang lebih stabil dan akurat. Random Forest memiliki kemampuan yang baik dalam mengatasi masalah overfitting, menangani data berdimensi tinggi, serta mampu memberikan tingkat akurasi yang tinggi pada berbagai permasalahan klasifikasi(Salman, Kalakech, dan Steiti 2024). Oleh karena itu, algoritma ini banyak digunakan dalam penelitian medis yang melibatkan data kompleks dan jumlah atribut yang besar.

Beberapa penelitian terdahulu menunjukkan bahwa algoritma machine learning mampu memberikan performa yang baik dalam klasifikasi kanker payudara. Penelitian (Vincent Angkasa1 2022)menunjukkan bahwa algoritma K-NN memperoleh nilai ROC sebesar 0,9959, sedangkan Random Forest memperoleh nilai ROC sebesar 0,9951 pada Breast Cancer Wisconsin Dataset. Hasil tersebut menunjukkan bahwa kedua algoritma memiliki kemampuan klasifikasi yang sangat baik, meskipun terdapat perbedaan performa yang relatif kecil. Selain itu, berbagai penelitian lainnya juga melaporkan tingkat akurasi yang tinggi pada penerapan algoritma machine learning untuk diagnosis kanker payudara, sehingga menunjukkan potensi besar teknologi ini dalam mendukung pengambilan keputusan di bidang kesehatan.

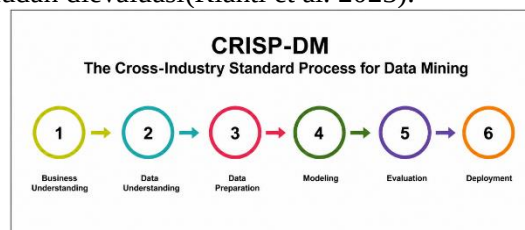
Meskipun telah banyak dilakukan penelitian mengenai klasifikasi kanker payudara, hasil performa algoritma sering kali berbeda tergantung pada teknik prapemrosesan data, pembagian data latih dan data uji, serta parameter yang digunakan. Oleh karena itu, diperlukan penelitian lebih lanjut untuk melakukan analisis perbandingan performa algoritma K-Nearest Neighbor dan Random Forest

pada Breast Cancer Wisconsin (Diagnostic) Dataset guna memperoleh informasi mengenai algoritma yang memiliki kemampuan klasifikasi terbaik berdasarkan metrik evaluasi yang digunakan.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis dan membandingkan kinerja algoritma K-Nearest Neighbor (K-NN) dan Random Forest dalam melakukan klasifikasi kanker payudara menggunakan Breast Cancer Wisconsin (Diagnostic) Dataset. Evaluasi performa dilakukan menggunakan metrik accuracy, precision, recall, F1-score, dan Area Under Curve (AUC) sehingga dapat diketahui algoritma yang lebih efektif dalam mendukung proses diagnosis kanker payudara secara cepat dan akurat.

## METODE

Penelitian ini menggunakan metode Cross Industry Standard Process for Data Mining (CRISP-DM) sebagai kerangka kerja dalam proses pengolahan dan analisis data. CRISP-DM dipilih karena merupakan metodologi yang sistematis dan banyak digunakan dalam penelitian data mining maupun machine learning. Metode ini terdiri atas enam tahapan utama, yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Melalui tahapan tersebut, proses pembangunan model klasifikasi dapat dilakukan secara terstruktur sehingga menghasilkan model yang optimal dan mudah dievaluasi (Rianti et al. 2023).



**Gambar 1. Metode CRISP-DM**

### 1. Business Understanding

Tahap Business Understanding bertujuan untuk mengidentifikasi permasalahan penelitian dan menentukan tujuan yang ingin dicapai. Permasalahan yang diangkat adalah perbandingan performa algoritma K-Nearest Neighbor (K-NN) dan Random Forest dalam klasifikasi kanker payudara. Tujuan penelitian adalah menentukan algoritma terbaik berdasarkan nilai accuracy, precision, recall, F1-score, dan AUC.

### 2. Data Understanding

Dataset yang digunakan adalah Breast Cancer Wisconsin (Diagnostic) Dataset yang diperoleh dari Kaggle dan berasal dari UCI Machine Learning Repository. Dataset terdiri dari 569 data dengan 32 atribut, yaitu 1 atribut identitas (id), 1 atribut target (diagnosis), dan 30 atribut numerik hasil pemeriksaan Fine Needle Aspiration (FNA). Kelas diagnosis terdiri dari Malignant (M) dan Benign (B).

### 3. Data Preparation

Tahap Data Preparation bertujuan untuk mempersiapkan data sebelum digunakan dalam proses pemodelan. Proses yang dilakukan meliputi:

#### a) Data Cleaning

Pada tahap ini dilakukan penghapusan atribut id karena tidak memiliki kontribusi terhadap proses klasifikasi. Hasil pemeriksaan menunjukkan bahwa dataset tidak memiliki missing value sehingga tidak diperlukan proses penanganan data kosong.

#### b) Transformasi Label

Atribut diagnosis yang memiliki nilai kategorikal diubah menjadi data numerik, yaitu:

- Malignant (M) = 1
- Benign (B) = 0

#### c) Normalisasi Data

Karena algoritma K-NN sensitif terhadap skala data, maka dilakukan proses normalisasi menggunakan metode Min-Max Scaling agar seluruh atribut memiliki rentang nilai yang sama.

#### d) Pembagian Dataset

Dataset dibagi menjadi data latih (training set) dan data uji (testing set) dengan proporsi Data latih : 80% dan Data uji : 20%. Pembagian data dilakukan menggunakan metode train-test split dengan nilai random state tertentu agar hasil eksperimen dapat direproduksi. Berdasarkan jumlah data sebanyak 569 record, maka diperoleh yaitu training Data = 455 data serta Testing Data = 114 data. Proses pembagian data dilakukan menggunakan fungsi `train_test_split()` dengan nilai `random_state = 42` untuk menjaga konsistensi hasil eksperimen.

### 4. Modeling

Pada tahap Modeling, dilakukan pembangunan model klasifikasi menggunakan dua algoritma machine learning, yaitu K-Nearest Neighbor (K-NN) dan Random Forest.

#### a) K-Nearest Neighbor (K-NN)

Algoritma K-NN bekerja dengan menentukan kelas suatu data berdasarkan mayoritas kelas dari sejumlah tetangga terdekat (nearest neighbors). Kedekatan antar data dihitung menggunakan metode Euclidean Distance. Persamaan Euclidean Distance (Yudhana, Sunardi, dan Hartanta 2020):

$$d(P, Q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

Pada penelitian ini dilakukan pengujian beberapa nilai parameter k, yaitu: k=3,5,7,9 untuk memperoleh nilai k yang menghasilkan performa terbaik.

#### b) Random Forest

Random Forest merupakan algoritma berbasis ensemble learning yang menggabungkan sejumlah pohon keputusan (decision trees) untuk menghasilkan prediksi yang lebih akurat. Setiap pohon dibangun menggunakan teknik bootstrap sampling dan random feature selection, kemudian hasil klasifikasi ditentukan berdasarkan mekanisme majority voting (Salman et al. 2024). Model Random Forest dibangun menggunakan 100 decision tree dengan parameter `n_estimators=100` dan `criterion='gini'`. Parameter utama yang digunakan meliputi:

**Tabel 1. Parameter yang Digunakan**

| Parameter                      | Nilai   |
|--------------------------------|---------|
| <code>n_estimators</code>      | 100     |
| <code>criterion</code>         | gini    |
| <code>random_state</code>      | 42      |
| <code>max_depth</code>         | Default |
| <code>min_samples_split</code> | 2       |

### 5. Evaluation

Tahap Evaluation dilakukan untuk mengukur performa kedua model klasifikasi. Evaluasi dilakukan menggunakan Confusion Matrix yang menghasilkan nilai:

---

Keterangan:

TP (True Positive) : jumlah data positif yang diprediksi positif dengan benar.

TN (True Negative) : jumlah data negatif yang diprediksi negatif dengan benar.

FP (False Positive) : jumlah data negatif yang diprediksi positif.

FN (False Negative) : jumlah data positif yang diprediksi negatif.

Berdasarkan Confusion Matrix dihitung beberapa metrik evaluasi sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Selain itu, penelitian ini juga menggunakan Receiver Operating Characteristic (ROC) dan Area Under Curve (AUC) untuk mengetahui kemampuan model dalam membedakan kelas tumor ganas dan tumor jinak. Nilai AUC yang mendekati 1 menunjukkan performa klasifikasi yang semakin baik.

## 6. Deployment

Tahap Deployment merupakan tahap penyajian hasil penelitian. Tahap Deployment dilakukan dengan mengimplementasikan model terbaik hasil penelitian, yaitu Random Forest, ke dalam aplikasi berbasis web menggunakan framework Streamlit. Sistem memungkinkan pengguna memasukkan nilai atribut pemeriksaan dan memperoleh hasil klasifikasi kanker payudara secara otomatis.

## HASIL DAN PEMBAHASAN

### Hasil Pengujian Algoritma K-Nearest Neighbor (K-NN)

Pengujian algoritma K-Nearest Neighbor dilakukan menggunakan beberapa nilai parameter k, yaitu 3, 5, 7, dan 9. Tujuan pengujian ini adalah untuk menentukan nilai k yang menghasilkan performa klasifikasi terbaik pada Breast Cancer Wisconsin (Diagnostic) Dataset.

**Tabel 2. Hasil Pengujian Nilai k pada Algoritma K-NN**

| Nilai k | Accuracy |
|---------|----------|
| 3       | 95,61%   |
| 5       | 96,49%   |
| 7       | 95,61%   |
| 9       | 95,61%   |

Berdasarkan Tabel 1, nilai k terbaik diperoleh pada k=5 dengan tingkat akurasi sebesar 96,49%. Oleh karena itu, nilai k=5 digunakan sebagai model K-NN terbaik untuk dibandingkan dengan algoritma Random Forest.

Selanjutnya dilakukan evaluasi menggunakan confusion matrix untuk mengetahui jumlah data yang berhasil dan gagal diklasifikasikan oleh model.

**Tabel 3. Confusion Matrix K-NN**

| Aktual    | Prediksi Benign | Prediksi Malignant |
|-----------|-----------------|--------------------|
| Benign    | 72              | 0                  |
| Malignant | 4               | 38                 |

Berdasarkan confusion matrix tersebut, model K-NN berhasil mengklasifikasikan seluruh data benign dengan benar tanpa menghasilkan false positive. Namun masih terdapat 4 data malignant yang salah diklasifikasikan sebagai benign.

**Tabel 4. Hasil Evaluasi Algoritma K-NN**

| Metrik    | Nilai   |
|-----------|---------|
| Accuracy  | 96,49%  |
| Precision | 100,00% |
| Recall    | 90,48%  |
| F1-Score  | 95,00%  |
| AUC       | 99,62%  |

Hasil evaluasi menunjukkan bahwa algoritma K-NN memiliki kemampuan klasifikasi yang sangat baik dengan nilai accuracy sebesar 96,49%. Nilai precision sebesar 100% menunjukkan bahwa seluruh data yang diprediksi sebagai kanker ganas merupakan data yang benar-benar termasuk kelas kanker ganas. Selain itu, nilai AUC sebesar 99,62% menunjukkan kemampuan model yang sangat tinggi dalam membedakan antara kelas tumor ganas dan tumor jinak.

### Hasil Pengujian Algoritma Random Forest

Algoritma Random Forest dibangun menggunakan parameter  $n\_estimators = 100$ ,  $criterion = gini$ , dan  $random\_state = 42$  sesuai dengan rancangan penelitian.

**Tabel 5. Confusion Matrix Random Forest**

| Aktual    | Prediksi Benign | Prediksi Malignant |
|-----------|-----------------|--------------------|
| Benign    | 72              | 0                  |
| Malignant | 3               | 39                 |

Berdasarkan confusion matrix tersebut, Random Forest berhasil mengklasifikasikan seluruh data benign dengan benar dan hanya menghasilkan 3 kesalahan klasifikasi pada kelas malignant.

**Tabel 6. Hasil Evaluasi Algoritma Random Forest**

| Metrik    | Nilai   |
|-----------|---------|
| Accuracy  | 97,37%  |
| Precision | 100,00% |
| Recall    | 92,86%  |
| F1-Score  | 96,30%  |
| AUC       | 99,29%  |

Hasil evaluasi menunjukkan bahwa Random Forest memperoleh accuracy sebesar 97,37%, precision sebesar 100%, recall sebesar 92,86%, F1-score sebesar 96,30%, dan AUC sebesar 99,29%. Nilai tersebut menunjukkan bahwa Random Forest mampu melakukan klasifikasi kanker payudara dengan tingkat ketepatan yang sangat tinggi.

### Perbandingan Kinerja K-NN dan Random Forest

Perbandingan performa kedua algoritma dilakukan menggunakan metrik accuracy, precision, recall, F1-score, dan AUC.

**Tabel 7. Perbandingan Hasil Evaluasi K-NN dan Random Forest**

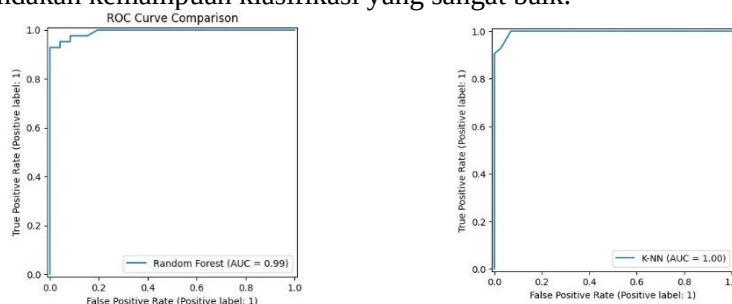
| Metrik    | K-NN    | Random Forest |
|-----------|---------|---------------|
| Accuracy  | 96,49%  | 97,37%        |
| Precision | 100,00% | 100,00%       |
| Recall    | 90,48%  | 92,86%        |
| F1-Score  | 95,00%  | 96,30%        |
| AUC       | 99,62%  | 99,29%        |

Berdasarkan Tabel 6, algoritma Random Forest memperoleh nilai accuracy, recall, dan F1-score yang lebih tinggi dibandingkan algoritma K-NN. Accuracy Random Forest mencapai 97,37%, sedangkan K-NN memperoleh 96,49%. Perbedaan ini menunjukkan bahwa Random Forest memiliki kemampuan yang lebih baik dalam melakukan klasifikasi terhadap data yang belum pernah dilihat sebelumnya. Pada metrik recall, Random Forest memperoleh nilai 92,86%, lebih tinggi dibandingkan K-NN sebesar 90,48%. Hal ini menunjukkan bahwa Random Forest lebih mampu mendeteksi kasus kanker ganas (malignant) dibandingkan K-NN. Dalam konteks diagnosis medis, recall merupakan metrik yang sangat penting karena berkaitan dengan kemampuan model dalam mengidentifikasi pasien yang benar-benar menderita kanker.

Nilai precision kedua algoritma mencapai 100%, yang menunjukkan bahwa seluruh prediksi positif yang dihasilkan model merupakan prediksi yang benar. Dengan kata lain, tidak terdapat kasus benign yang salah diklasifikasikan sebagai malignant pada kedua model. Pada metrik F1-score, Random Forest memperoleh nilai 96,30%, sedangkan K-NN memperoleh 95,00%. Nilai F1-score yang lebih tinggi menunjukkan bahwa Random Forest memiliki keseimbangan yang lebih baik antara precision dan recall. Meskipun demikian, algoritma K-NN memperoleh nilai AUC sebesar 99,62%, sedikit lebih tinggi dibandingkan Random Forest yang memperoleh nilai AUC sebesar 99,29%. Nilai AUC yang mendekati 1 menunjukkan bahwa kedua algoritma memiliki kemampuan yang sangat baik dalam membedakan kelas tumor ganas dan tumor jinak. Perbedaan nilai AUC yang relatif kecil menunjukkan bahwa kedua model sama-sama memiliki performa klasifikasi yang sangat tinggi.

### Analisis Kurva ROC

Kurva Receiver Operating Characteristic (ROC) digunakan untuk mengevaluasi kemampuan model dalam membedakan kelas positif dan negatif pada berbagai nilai threshold. Hasil pengujian menunjukkan bahwa kurva ROC kedua algoritma berada sangat dekat dengan sudut kiri atas grafik, yang menandakan kemampuan klasifikasi yang sangat baik.

**Gambar 2. ROC Curve Kedua Algoritma**

Nilai AUC sebesar 99,62% pada K-NN dan 99,29% pada Random Forest menunjukkan bahwa kedua algoritma mampu membedakan kelas malignant dan benign dengan tingkat akurasi yang sangat tinggi. Berdasarkan kategori interpretasi AUC, nilai di atas 90% termasuk dalam kategori excellent classification. Oleh karena itu, baik K-NN maupun Random Forest dapat digunakan sebagai model klasifikasi kanker payudara yang efektif.



## Pembahasan

Hasil penelitian menunjukkan bahwa kedua algoritma machine learning yang digunakan mampu menghasilkan performa klasifikasi yang sangat baik pada Breast Cancer Wisconsin (Diagnostic) Dataset. Tingginya nilai accuracy dan AUC menunjukkan bahwa karakteristik sel kanker yang terdapat pada dataset dapat dipelajari dengan baik oleh kedua model.

Algoritma K-NN bekerja berdasarkan prinsip kedekatan jarak antar data. Setelah dilakukan normalisasi menggunakan Min-Max Scaling, K-NN mampu menghasilkan accuracy sebesar 96,49%. Hal ini menunjukkan bahwa normalisasi berhasil meningkatkan efektivitas pengukuran jarak antar data sehingga proses klasifikasi menjadi lebih akurat.

Sementara itu, Random Forest menghasilkan performa yang sedikit lebih unggul dibandingkan K-NN. Keunggulan tersebut disebabkan oleh kemampuan Random Forest dalam membangun banyak pohon keputusan yang kemudian digabungkan melalui mekanisme majority voting. Pendekatan ensemble learning ini membuat model lebih tahan terhadap overfitting dan mampu menangkap pola hubungan antar atribut yang lebih kompleks. Berdasarkan keseluruhan hasil evaluasi, Random Forest dipilih sebagai model terbaik karena menghasilkan nilai accuracy, recall, dan F1-score yang lebih tinggi dibandingkan K-NN. Meskipun K-NN memperoleh nilai AUC yang sedikit lebih tinggi, perbedaan tersebut relatif kecil sehingga tidak mengubah kesimpulan bahwa Random Forest memiliki performa klasifikasi yang lebih unggul pada dataset yang digunakan dalam penelitian ini.

## Implementasi Sistem

Setelah proses pelatihan dan evaluasi model selesai dilakukan, algoritma dengan performa terbaik yaitu Random Forest diimplementasikan ke dalam sebuah aplikasi berbasis web menggunakan framework Streamlit. Implementasi ini bertujuan untuk menunjukkan bahwa model yang telah dibangun tidak hanya mampu menghasilkan performa klasifikasi yang tinggi, tetapi juga dapat digunakan sebagai alat bantu prediksi kanker payudara secara interaktif.

Tampilan utama sistem terdiri dari beberapa bagian, yaitu informasi model, formulir input data pasien, dan hasil prediksi. Pada bagian informasi model ditampilkan ringkasan performa model Random Forest yang diperoleh dari hasil penelitian, yaitu accuracy sebesar 97,37%, jumlah dataset sebanyak 569 data, serta 30 atribut yang digunakan dalam proses klasifikasi.

**Tentang Sistem**  
Implementasi Random Forest untuk klasifikasi kanker payudara berdasarkan Breast Cancer Wisconsin Dataset.

**Sistem Klasifikasi Kanker Payudara**  
Random Forest (Model Terbaik Penelitian)

Accuracy: 97.37%    Dataset: 569 Data    Atribut: 30 Atribut

**Input Data Pasien**

|                        |                         |                        |
|------------------------|-------------------------|------------------------|
| mean radius            | mean texture            | mean perimeter         |
| 14.23700               | 15.29940                | 11.06030               |
| mean area              | mean circ-imp           | mean compactness       |
| 854.08910              | 0.09630                 | 0.04341                |
| mean concavity         | mean concave points     | mean symmetry          |
| 0.08670                | 0.04030                 | 0.03142                |
| mean fractal dimension | radius error            | texture error          |
| 0.06270                | 0.05172                 | 1.21053                |
| perimeter error        | compactness error       | symmetry error         |
| 2.00009                | 40.37070                | 0.07011                |
| compactness error      | compactness error       | compactness error      |
| 0.02910                | 0.03104                 | 0.01790                |
| symmetry error         | fractal dimension error | mean texture           |
| 0.00942                | 0.00370                 | 16.30310               |
| compactness error      | compactness error       | mean area              |
| 28.87720               | 107.15120               | 805.88120              |
| compactness error      | compactness error       | mean symmetry          |
| 0.03200                | 0.04400                 | 0.02200                |
| mean compactness error | mean symmetry           | mean fractal dimension |
| 0.01400                | 0.00070                 | 0.00040                |

**Hasil Prediksi**  
Benign (Tumor jinak)  
Probability: 65.00%

Gambar 3. Tampilan Implementasi Sistem



---

Pengguna dapat memasukkan nilai dari 30 atribut pemeriksaan yang terdapat pada Breast Cancer Wisconsin (Diagnostic) Dataset, seperti mean radius, mean texture, mean perimeter, mean area, mean smoothness, dan atribut lainnya. Data yang dimasukkan kemudian diproses oleh model Random Forest yang telah dilatih sebelumnya untuk menghasilkan prediksi diagnosis.

Setelah tombol prediksi dijalankan, sistem akan menampilkan hasil klasifikasi berupa kategori Benign (tumor jinak) atau Malignant (kanker ganas) beserta nilai probabilitas prediksi. Pada contoh implementasi yang ditunjukkan pada Gambar 3, sistem menghasilkan prediksi Benign (Tumor Jinak) dengan tingkat probabilitas sebesar 65,00%. Implementasi sistem ini menunjukkan bahwa model Random Forest yang dihasilkan pada penelitian mampu diterapkan dalam bentuk aplikasi berbasis web yang mudah digunakan. Dengan tingkat akurasi sebesar 97,37%, sistem berpotensi digunakan sebagai alat bantu pendukung keputusan dalam proses identifikasi awal kanker payudara. Namun demikian, hasil prediksi yang diberikan sistem tidak dimaksudkan untuk menggantikan diagnosis medis, melainkan sebagai sarana pendukung yang dapat membantu proses analisis data kesehatan.

## SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan menggunakan Breast Cancer Wisconsin (Diagnostic) Dataset, dapat disimpulkan bahwa algoritma K-Nearest Neighbor (K-NN) dan Random Forest mampu melakukan klasifikasi kanker payudara dengan performa yang sangat baik. Hasil pengujian menunjukkan bahwa nilai k terbaik pada algoritma K-NN diperoleh pada k=5 dengan accuracy sebesar 96,49%. Berdasarkan perbandingan performa, algoritma Random Forest memperoleh hasil yang lebih baik dibandingkan K-NN dengan nilai accuracy sebesar 97,37%, recall sebesar 92,86%, dan F1-score sebesar 96,30%. Sementara itu, K-NN memperoleh accuracy sebesar 96,49%, recall sebesar 90,48%, dan F1-score sebesar 95,00%. Meskipun K-NN memiliki nilai AUC sedikit lebih tinggi yaitu 99,62%, perbedaannya relatif kecil sehingga tidak mempengaruhi hasil akhir penelitian. Dengan demikian, Random Forest dapat direkomendasikan sebagai algoritma yang lebih optimal untuk klasifikasi kanker payudara pada Breast Cancer Wisconsin (Diagnostic) Dataset karena mampu memberikan performa klasifikasi yang lebih baik secara keseluruhan. Selain itu, model Random Forest berhasil diimplementasikan ke dalam aplikasi berbasis web menggunakan Streamlit sehingga dapat digunakan sebagai alat bantu klasifikasi kanker payudara secara interaktif. Penelitian selanjutnya disarankan dapat menggunakan teknik validasi yang lebih komprehensif seperti K-Fold Cross Validation, melakukan optimasi parameter pada masing-masing algoritma, serta membandingkan Random Forest dengan algoritma machine learning lainnya untuk memperoleh performa klasifikasi yang lebih optimal.

## DAFTAR PUSTAKA

- Eka mei dianata, Lie Liana, Azmi Avi. 2025. *Peningkatan PENGEYAHUAN TERADAP DETEKSI DINI Kanker Payudar PADA REMAJA*. Vol. 2.
- Hulaifah Al Abrori, Zahra Zul, dan Egia Rosi Subhiyakto. 2025. "Analisis Komparatif Akurasi Prediksi Kanker Payudara Menggunakan Algoritma Random Forest dan Logistic Regression." *Jurnal Algoritma* 22(1):300–311. doi: 10.33364/algoritma/v.22-1.2164.
- Idris, Mohamad, Angga Wijaya, Linda Septiani, Terza Aflika Happy, dan Risti Graharti. 2025. "Pemanfaatan Kecerdasan Buatan sebagai Alat Bantu Diagnosis di Bidang Kesehatan : Literatur Review Utilization of Artificial Intelligence as a Diagnostic Aid in the Health Sector : A Literature Review." *JK Unila* 9(1):117–21.
- Malik Namus Akbar, Fahmi. 2024. "Metode KNN (K-Nearest Neighbor) untuk Menentukan Kualitas Air." *Jurnal Tekno Kompak* 18(1):28–40.
- Riadinni Alita1, Uum Safari1, Nova Riani1, Fariha Khoerunisa1, Dhea Amanda Sulistiani1. 2020. "PENINGKATAN PENGETAHUAN TENTANG Pencegahan Kanker Payudara MELALUI 'SADARI' PADA REMAJA DI SMK PELITA ALAM." *JMH* 02:434–40.

- 
- Rianti, Afika, Nuur Wachid, Abdul Majid, dan Ahmad Fauzi. 2023. "CRISP-DM : Metodologi Proyek Data Science." *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB) 2023* 107–14.
- Salman, Hasan Ahmed, Ali Kalakech, dan Amani Steiti. 2024. "Random Forest Algorithm Overview." *Babylonian Journal of Machine Learning* 2024:69–79. doi: 10.58496/BJML/2024/007.
- Tia Siti Nurazizah. 2025. "Artificial Intelligence Dan Machine Learning Dalam Kehidupan Manusia." *Jurnal Teknologi Informasi Indonesia* 9(2):45–60.
- Vincent Angkasa<sup>1</sup>, Jefri Junifer Pangaribuan<sup>2</sup>. 2022. "Komparasi Tingkat Akurasi Random Forest dan KNN untuk Mendiagnosis Penyakit Kanker Payudara." *Informasi System Development X*(January):34–41.
- Yudhana, Anton, Sunardi Sunardi, dan Agus Jaka Sri Hartanta. 2020. "Algoritma K-Nn Dengan Euclidean Distance Untuk Prediksi Hasil Penggergajian Kayu Sengon." *Transmisi* 22(4):123–29. doi: 10.14710/transmisi.22.4.123-129.