

IMPLEMENTASI *FEATURE SELECTION* RELIEFF DAN KLASIFIKASI STATUS KANKER PAYUDARA MENGGUNAKAN *RANDOM FOREST* DAN *LOGISTIC REGRESSION* DENGAN INTERPRETASI SHAP

Fanny Krisvyanti, Abdul Rahman

Informatika, Universitas Multi Data Palembang

Jl Rajawali No.14, 9 Ilir, Kec. Ilir Tim. II, Kota Palembang, Sumatera Selatan

fannykrisvyanti_2226250048@mhs.mdp.ac.id

ABSTRAK

Kanker payudara merupakan salah satu penyakit dengan tingkat prevalensi yang tinggi dan menjadi penyebab utama kematian pada perempuan di seluruh dunia, sehingga diperlukan metode deteksi dini yang akurat dan dapat diinterpretasikan dengan baik. Namun, banyak *model machine learning* yang masih bersifat kurang transparan sehingga sulit dipahami dalam pengambilan keputusan klinis. Penelitian ini bertujuan untuk membangun model klasifikasi status kanker payudara yang tidak hanya memiliki performa tinggi, tetapi juga mudah diinterpretasikan. Metode yang digunakan meliputi *preprocessing data* berupa *data cleaning*, *encoding*, dan *scaling*, serta pembagian data dengan rasio 80:20. Seleksi fitur dilakukan menggunakan ReliefF untuk meningkatkan relevansi fitur terhadap kelas target. Model klasifikasi yang digunakan adalah *Logistic Regression* dan *Random Forest*, dengan evaluasi menggunakan *5-Fold Cross Validation* berdasarkan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *AUC-ROC*. Selain itu, interpretasi model dilakukan menggunakan SHAP (*SHapley Additive exPlanations*) untuk menjelaskan kontribusi setiap fitur terhadap hasil prediksi. Hasil penelitian menunjukkan bahwa *Logistic Regression* memberikan performa terbaik dengan *accuracy* 83,10%, *F1-score* 82,86%, dan *AUC-ROC* 83,94%. Selain itu, SHAP mampu memberikan interpretasi yang jelas terhadap pengaruh fitur sehingga meningkatkan transparansi model. Pendekatan kombinasi ReliefF, *Logistic Regression*, dan SHAP terbukti efektif untuk klasifikasi kanker payudara.

Kata kunci : Kanker Payudara, Machine Learning, ReliefF, Logistic Regression, SHAP, Klasifikasi.

1. PENDAHULUAN

Kanker payudara masih menjadi salah satu masalah kesehatan utama pada perempuan di seluruh dunia. Organisasi Kesehatan Dunia (WHO) melaporkan bahwa kanker payudara merupakan jenis kanker yang paling sering didiagnosis pada perempuan secara global [1].

Tingginya angka kejadian tersebut menunjukkan pentingnya deteksi dini dan identifikasi status kanker secara akurat guna mendukung pengambilan keputusan medis yang tepat.

Dalam praktik klinis, identifikasi status kanker payudara tidak hanya bergantung pada pencitraan medis, tetapi juga melibatkan berbagai faktor klinis dan demografis pasien seperti usia, indeks massa tubuh (BMI), riwayat reproduksi, pola menyusui, dan pola konsumsi makanan berlemak [2], [3].

Faktor-faktor tersebut umumnya tersusun dalam bentuk data tabular. Namun, interpretasi terhadap kombinasi berbagai faktor risiko masih dapat dipengaruhi oleh subjektivitas klinis, terutama pada kasus dengan nilai yang berada di batas ambang. Hal ini berpotensi menyebabkan variasi dalam pengambilan keputusan sehingga diperlukan pendekatan berbasis data yang lebih objektif dan konsisten.

Metode statistik seperti regresi logistik telah lama digunakan dalam klasifikasi klinis karena mampu menjelaskan hubungan antar variabel [4].

Namun, metode ini memiliki keterbatasan dalam menangkap hubungan non-linear dan interaksi kompleks antar fitur. Oleh karena itu, algoritma *machine learning* semakin banyak diterapkan untuk klasifikasi kanker berbasis data tabular karena kemampuannya dalam mempelajari pola yang lebih kompleks [3], [5].

Meskipun demikian, meningkatnya kompleksitas model sering kali mengurangi transparansi hasil prediksi sehingga menyulitkan interpretasi dalam konteks medis [6].

Untuk mengatasi masalah tersebut, *Explainable Artificial Intelligence* (XAI) dikembangkan untuk memberikan penjelasan terhadap kontribusi fitur dalam proses prediksi [6], [7].

Namun, pada dataset dengan jumlah fitur yang besar, interpretasi dapat menjadi kurang terfokus. Oleh karena itu, *feature selection* seperti ReliefF diperlukan untuk memilih fitur yang paling relevan sebelum proses pemodelan, sehingga interpretasi model menjadi lebih efektif. Meski demikian, integrasi antara *feature selection* dan XAI masih relatif terbatas dalam penelitian kanker berbasis data tabular [8].

Berdasarkan hal tersebut, penelitian ini mengusulkan pendekatan terintegrasi dengan menggabungkan ReliefF sebagai metode *feature selection* dan SHAP sebagai metode *explainable AI* untuk klasifikasi kanker payudara (BC vs Non-BC). Selain itu, penelitian ini membandingkan model *Logistic Regression* dan *Random Forest*, baik sebelum

maupun sesudah seleksi fitur, untuk mengevaluasi pengaruhnya terhadap performa klasifikasi dan interpretabilitas model.

2. TINJAUAN PUSTAKA

Tinjauan pustaka membahas konsep kanker payudara, *machine learning* pada data tabular, metode ReliefF untuk seleksi fitur, algoritma *Logistic Regression* dan *Random Forest*, metode interpretasi SHAP, serta metrik evaluasi model sebagai landasan teoritis penelitian.

2.1. Penelitian Terdahulu

Penelitian klasifikasi kanker payudara berbasis *machine learning* telah banyak dilakukan dengan berbagai algoritma, teknik seleksi fitur, dan metode interpretabilitas. Penggunaan *feature selection* berbasis *grammatical evolution* pada dataset WDBC terbukti dapat meningkatkan performa klasifikasi hingga mencapai AUC sebesar 0,9995, meskipun aspek interpretabilitas model belum dibahas secara mendalam [9].

Penerapan metode *interpretable machine learning* berbasis LIME menunjukkan bahwa model *Decision Tree* dapat mencapai akurasi sebesar 97,95%, namun pendekatan tersebut belum mengintegrasikan proses seleksi fitur dalam pemodelan [10].

Selain itu, perbandingan berbagai algoritma klasifikasi menunjukkan bahwa *Support Vector Machine* (SVM) menghasilkan akurasi terbaik sebesar 97,2%, tetapi penelitian tersebut masih berfokus pada peningkatan performa prediksi tanpa mempertimbangkan aspek interpretabilitas model [11].

Integrasi *feature selection* dan *Explainable Artificial Intelligence* (XAI) mulai dikembangkan melalui kombinasi *feature selection* berbasis *Random Forest* dan SHAP, yang mampu menghasilkan akurasi hingga 99,7% [12].

Pendekatan serupa juga diterapkan pada data genomik dengan memanfaatkan SHAP untuk meningkatkan interpretabilitas hasil klasifikasi [13].

Meskipun demikian, sebagian besar penelitian terdahulu masih berfokus pada dataset luar negeri maupun data genomik, sehingga penelitian berbasis data tabular faktor risiko pada populasi wanita Indonesia masih relatif terbatas. Untuk mengatasi kesenjangan tersebut, penelitian ini mengusulkan integrasi metode ReliefF sebagai *feature selection*, *Logistic Regression* dan *Random Forest* sebagai model klasifikasi, serta SHAP sebagai metode interpretabilitas. Pendekatan ini diharapkan dapat menghasilkan model yang tidak hanya memiliki performa prediksi yang baik, tetapi juga mudah diinterpretasikan dalam konteks klinis.

2.2. Kanker Payudara

Kanker payudara merupakan kondisi patologi yang terjadi akibat tumbuhnya sel-sel abnormal pada jaringan payudara secara tidak terkendali, sehingga

berpotensi menginvasi dan menyebar ke organ tubuh lain melalui mekanisme metastasis. Menurut World Health Organization (WHO), kanker payudara merupakan jenis kanker yang paling sering didiagnosis pada perempuan secara global serta menjadi salah satu penyebab utama kematian akibat kanker [1].

Faktor risiko kanker payudara meliputi usia menarche, usia melahirkan pertama, obesitas, dan usia menopause. Kompleksitas faktor risiko tersebut mendorong pengembangan model prediksi berbasis data untuk membantu identifikasi dini dan mendukung pengambilan keputusan klinis.

2.3. Data Tabular dalam Analisis Medis

Data tabular adalah data heterogen yang tersusun atas fitur numerik dan fitur kategorikal [14].

Dalam bidang kesehatan, data ini digunakan untuk menyimpan informasi pasien, seperti data demografi, diagnosis, dan karakteristik klinis. Pemanfaatan *machine learning* pada data kesehatan memungkinkan identifikasi pola untuk memprediksi risiko penyakit dan mendukung pengambilan keputusan medis [15].

2.4. Machine Learning

Machine learning (ML) merupakan salah satu cabang kecerdasan buatan yang memungkinkan sistem mempelajari pola dari data untuk menghasilkan prediksi tanpa memerlukan pemrograman secara eksplisit. Pada penelitian ini digunakan metode *supervised learning*, yaitu pendekatan yang membangun model berdasarkan data masukan dan keluaran yang telah diketahui [16].

Hubungan antara fitur masukan dan label keluaran dinyatakan pada Persamaan (1).

$$y = f(x) \quad (1)$$

Pada Persamaan (1), variabel (x) merepresentasikan fitur masukan, (y) menunjukkan label keluaran, sedangkan ($f(x)$) merupakan fungsi yang dipelajari model untuk memetakan data input ke output [17].

2.5. Logistic Regression

Logistic Regression adalah metode *supervised learning* yang digunakan untuk menyelesaikan masalah klasifikasi biner dengan memprediksi probabilitas keanggotaan suatu kelas [18].

Metode ini menggunakan fungsi sigmoid untuk mentransformasikan kombinasi linear fitur menjadi probabilitas pada rentang 0–1 sebagaimana ditunjukkan pada Persamaan (2).

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (2)$$

Pada Persamaan (2), variabel (z) menunjukkan kombinasi linear dari fitur masukan yang digunakan untuk menghasilkan probabilitas klasifikasi [19].

2.6. Random Forest

Sebagai salah satu algoritma *ensemble learning*, *Random Forest* memanfaatkan kumpulan *decision tree* untuk menghasilkan prediksi yang lebih andal [20].

Dalam proses pembentukannya, setiap pohon menggunakan teknik *bootstrap aggregating (bagging)* serta subset fitur yang dipilih secara acak, kemudian prediksi akhir ditetapkan melalui suara mayoritas (*majority voting*) [21].

Pendekatan ini mampu mengurangi risiko *overfitting* dan meningkatkan performa klasifikasi.

2.7. Feature Selection dan ReliefF

Feature selection merupakan proses pemilihan fitur paling relevan untuk mengurangi dimensionalitas data, meningkatkan generalisasi model, serta mempermudah interpretasi hasil [22].

Salah satu metode *filter-based feature selection* adalah ReliefF, yang bekerja dengan memberikan bobot pada fitur berdasarkan kemampuannya membedakan antar kelas [23].

ReliefF menentukan relevansi fitur melalui perbandingan sampel dengan *nearest hit* (kelas sama) dan *nearest miss* (kelas berbeda) [24].

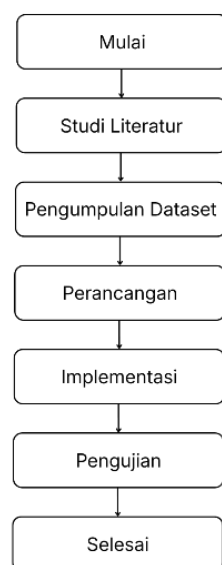
2.8. Explainable Artificial Intelligence (XAI) dan SHAP

Explainable Artificial Intelligence (XAI) merupakan pendekatan yang dikembangkan untuk meningkatkan transparansi model AI sehingga proses dan hasil prediksi dapat dipahami oleh manusia, khususnya pada bidang kritis seperti kesehatan [25].

Salah satu metode XAI yang banyak diterapkan adalah SHAP (*SHapley Additive exPlanations*), yaitu teknik interpretabilitas yang berlandaskan teori permainan kooperatif untuk menjelaskan pengaruh setiap fitur terhadap prediksi model [26].

SHAP sangat sesuai digunakan pada model black-box seperti Random Forest karena dapat menyeimbangkan kebutuhan akan performa prediksi yang tinggi dan interpretabilitas model [27].

3. METODE PENELITIAN



Gambar 1. Tahapan Metodologi Penelitian

Sebelum penelitian dilaksanakan, terlebih dahulu dilakukan perancangan metode yang akan digunakan. Metodologi yang dirancang bertujuan membangun model klasifikasi kanker payudara menggunakan algoritma *Logistic Regression* dan *Random Forest* dengan penerapan seleksi fitur ReliefF serta interpretasi model berbasis *Explainable Artificial Intelligence (XAI)* menggunakan SHAP. Rangkaian tahapan penelitian ditunjukkan pada Gambar 1.

Gambar 1 menggambarkan alur metodologi yang diterapkan dalam penelitian ini. Metodologi disusun secara sistematis untuk memastikan setiap tahapan dapat dilaksanakan secara terstruktur sesuai dengan tujuan penelitian. Proses penelitian diawali dengan studi literatur dan pengumpulan dataset, kemudian dilanjutkan dengan tahap perancangan, implementasi, serta pengujian.

3.1. Studi Literatur

Tahap ini diawali dengan studi literatur melalui pengumpulan jurnal, prosiding, dan penelitian terdahulu yang relevan dengan topik penelitian, yaitu klasifikasi kanker payudara menggunakan metode *machine learning*. Literatur yang dikaji mencakup konsep dasar *machine learning*, *Logistic Regression*, *Random Forest*, seleksi fitur menggunakan ReliefF, serta *Explainable Artificial Intelligence (XAI)* dengan metode SHAP.

3.2. Pengumpulan Dataset

Pada tahap ini dilakukan pengumpulan dataset yang digunakan sebagai sumber data penelitian. Dataset yang digunakan adalah *Dataset of Indonesian Women's Reproductive, High-Fat Diet and Body Mass Index Risk Factors for Breast Cancer* [28], yang diperoleh melalui platform *ScienceDirect*. Dataset tersebut dipilih karena memuat data faktor risiko kanker payudara pada wanita Indonesia berbasis data tabular, sehingga dinilai relevan dengan tujuan penelitian dalam melakukan klasifikasi status kanker payudara menggunakan pendekatan *machine learning*.

Dataset terdiri atas 400 data dengan 13 atribut, yang mencakup 12 atribut prediktor dan 1 atribut target sebagai label klasifikasi. Distribusi data terbagi secara seimbang menjadi 200 data kelas BC (*Breast Cancer*) dan 200 data kelas Non-BC, sehingga dapat membantu mengurangi potensi ketidakseimbangan kelas (*class imbalance*) pada tahap pemodelan. Deskripsi setiap atribut yang digunakan dalam penelitian ditampilkan pada Tabel 1.

Atribut pada dataset merepresentasikan berbagai faktor risiko yang berpotensi memengaruhi status kanker payudara, seperti usia (*Age*), riwayat reproduksi (*Parity*, *Breastfeeding*, *Menopause*, dan *First Pregnancy*), kebiasaan konsumsi makanan berlemak (*Highfat*), serta *Body Mass Index (BMI)*. Selain itu, terdapat atribut lain seperti status pernikahan, pendidikan, etnis, dan status pekerjaan yang turut digunakan sebagai variabel prediktor dalam proses klasifikasi. Dataset yang telah dikumpulkan

selanjutnya digunakan pada tahap *preprocessing*, seleksi fitur, pelatihan model, serta evaluasi performa klasifikasi status kanker payudara.

Tabel 1. Sampel Dataset

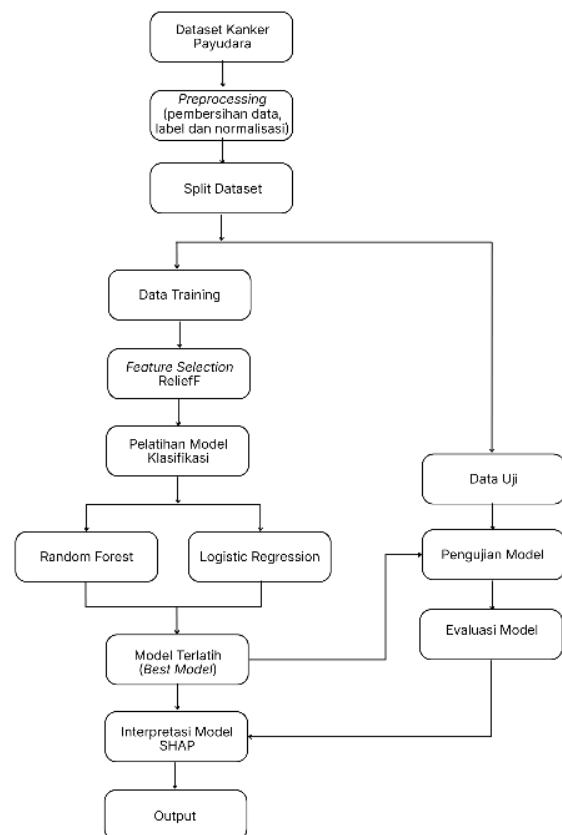
Grouping	Age (years)	Education	Working status	Marital status	Menarche (years)	Menopause	First pregnancy	Parity	Breastfeeding	Highfat	BMI	Ethnicity
BC	>= 50	Senior high school	Housewife	Marriage	12 to 13	< 50 years	20-29 years	>= Multiparous	>=12 months	Normal	Obesity	Minangnese
Non-BC	>= 50	Vocational degree	Master's student	Single/widow	>13	< 50 years	>30 years	>= Multiparous	<12 months	High	Normal	Minangnese
BC	>= 50	Senior high school	Housewife	Marriage	7 to 11	< 50 years	20-29 years	>= Multiparous	>=12 months	Normal	Normal	Javanese
Non-BC	>= 50	Senior high school	Master's student	Marriage	12 to 13	< 50 years	20-29 years	Primiparous	>=12 months	High	Normal	Javanese

3.3. Perancangan

Tahap perancangan menjelaskan alur pengembangan sistem klasifikasi status kanker payudara yang memanfaatkan algoritma *machine learning*, metode seleksi fitur ReliefF, dan pendekatan *Explainable Artificial Intelligence* (XAI) berbasis SHAP. Penelitian diawali dengan proses *preprocessing* yang meliputi pembersihan data (*data cleaning*), transformasi fitur kategorikal (*data encoding*), normalisasi data, serta pembagian dataset menjadi data latih dan data uji dengan rasio 80:20. Pemilihan rasio tersebut mengacu pada penelitian sebelumnya yang menunjukkan bahwa konfigurasi tersebut mampu memberikan performa model yang optimal. [29].

Tahap ini dilakukan untuk memastikan kualitas data sebelum digunakan pada proses pemodelan sehingga dapat mengurangi pengaruh data yang tidak relevan terhadap hasil klasifikasi.

Selanjutnya dilakukan *feature selection* menggunakan metode ReliefF untuk menentukan fitur paling relevan berdasarkan nilai bobot ($W[A]$) melalui *nearest hit* dan *nearest miss*. Dari total 12 fitur, dipilih 7 fitur dengan bobot tertinggi untuk mengurangi dimensionalitas, meminimalkan *overfitting*, serta meningkatkan efisiensi komputasi dan interpretabilitas model. Fitur hasil seleksi kemudian digunakan sebagai input pada dua model klasifikasi, yaitu *Logistic Regression* sebagai *baseline model* dan *Random Forest* untuk menangkap hubungan non-linear antar fitur. Proses pelatihan dilakukan menggunakan *5-Fold Cross Validation* guna meningkatkan stabilitas evaluasi dan mengurangi bias akibat pembagian data tertentu. Evaluasi model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *AUC-ROC*, dengan pemilihan model terbaik berdasarkan nilai *F1-score* dan *AUC-ROC* tertinggi. Tahap akhir dilakukan interpretasi model terbaik menggunakan SHAP melalui *global explanation* (*summary plot* dan *bar plot*), *local explanation* (*waterfall plot*), serta *dependence plot* untuk menganalisis pengaruh dan interaksi fitur terhadap hasil prediksi. Diagram perancangan penelitian dapat dilihat pada Gambar 2.



Gambar 2. Diagram Perancangan Penelitian

3.4. Implementasi

Pada tahap ini dilakukan implementasi sistem klasifikasi status kanker payudara menggunakan bahasa pemrograman Python dengan pustaka *scikit-learn*, *pandas*, dan *NumPy*. Dataset hasil *preprocessing* dan seleksi fitur digunakan pada model *Logistic Regression* dan *Random Forest* dengan pembagian data rasio 80:20 serta *5-Fold Cross Validation*. Penilaian performa model dilakukan berdasarkan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *AUC-ROC*. Selain itu, metode SHAP diterapkan untuk menganalisis kontribusi fitur terhadap hasil prediksi sehingga meningkatkan transparansi keputusan model.

3.5. Pengujian

Pada tahap ini dilakukan pengujian terhadap model klasifikasi menggunakan data *testing* hasil pembagian dataset untuk mengetahui kemampuan model dalam mengklasifikasikan data kanker payudara. Evaluasi performa model dilakukan menggunakan *Confusion Matrix* dengan menghitung nilai *precision*, *recall*, *accuracy*, *specificity*, *F1-score*, *false positive rate (FPR)*, dan *Area Under Curve (AUC)*. Metrik evaluasi tersebut digunakan untuk mengukur tingkat ketepatan model dalam melakukan klasifikasi serta menentukan performa model terbaik. Perhitungan evaluasi model ditunjukkan pada Persamaan (3), (4), (5), (6), dan (7).

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (3)$$

$$Recall (TPR) = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (5)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

$$Area Under Curve (AUC) = \int_0^1 ROC(x) dx \quad (7)$$

Keterangan:

TP : Jumlah data positif (*defect*) yang berhasil diklasifikasikan dengan benar sebagai positif.

TN : Jumlah data negatif (*non-defect*) yang berhasil diklasifikasikan dengan benar sebagai negatif.

FN : Jumlah data positif (*defect*) yang salah diklasifikasikan sebagai negatif.

FP : Jumlah data negatif (*non-defect*) yang salah diklasifikasikan sebagai positif.

4. HASIL DAN PEMBAHASAN

Bab ini membahas hasil implementasi dan evaluasi model klasifikasi status kanker payudara. Pengujian dilakukan untuk menilai kemampuan model dalam melakukan klasifikasi serta mengevaluasi performanya. Penilaian kinerja model menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *Area Under Curve (AUC)* untuk menentukan model terbaik dan mengukur keandalan hasil klasifikasi.

4.1. Skenario Pengujian

Pengujian pada penelitian ini dilakukan menggunakan dataset kanker payudara yang telah melalui tahap *preprocessing*, meliputi *data cleaning*, *encoding*, dan *scaling*, sehingga diperoleh sebanyak 353 data dengan 12 fitur dan 1 variabel target. Variabel target dibagi ke dalam dua kelas, yaitu BC (0) dan Non-BC (1). Selanjutnya, dataset dipisahkan menjadi data latih dan data uji dengan rasio 80:20, sehingga diperoleh 282 data latih dan 71 data uji.

Sebelum proses klasifikasi dilakukan, diterapkan metode *feature selection* menggunakan ReliefF untuk memilih 7 fitur terbaik berdasarkan tingkat relevansinya terhadap target klasifikasi. Selanjutnya,

pengujian dilakukan pada dua model *machine learning*, yaitu *Logistic Regression* sebagai *baseline model* dan *Random Forest* untuk menangkap hubungan non-linear antar fitur. Evaluasi model dilakukan menggunakan skema *5-Fold Cross Validation* guna memperoleh hasil yang lebih konsisten dan mengurangi bias akibat pembagian data. Kinerja model selanjutnya dinilai berdasarkan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *Area Under Curve (AUC-ROC)*.

4.2. Hasil Pengujian

Bagian ini membahas hasil implementasi dan evaluasi model klasifikasi status kanker payudara yang telah dikembangkan. Analisis dilakukan secara bertahap meliputi hasil *preprocessing* data, seleksi fitur dilakukan menggunakan ReliefF, kemudian kinerja *Logistic Regression* dan *Random Forest* dievaluasi, sementara model dengan performa terbaik diinterpretasikan menggunakan SHAP untuk menjelaskan pengaruh setiap fitur terhadap prediksi.

Sebelum proses pemodelan, dataset terlebih dahulu melalui tahapan *preprocessing* yang meliputi pembersihan data, transformasi fitur kategorikal, dan normalisasi data. Dataset yang digunakan berjumlah 400 data dengan distribusi seimbang antara kelas BC dan Non-BC, masing-masing sebanyak 200 data. Distribusi data sebelum *preprocessing* ditampilkan pada Tabel 2.

Tabel 2. Data Sebelum Tahap *Preprocessing*

Komponen	Nilai
Total data awal	400
BC awal	200
Non-BC awal	200
Jumlah fitur	12

Tahap data *cleaning* dilakukan dengan menangani nilai yang hilang dan menghapus data yang terduplikasi untuk meningkatkan kualitas data dan mengurangi potensi bias pada model. Setelah proses *cleaning* dilakukan, jumlah data berkurang menjadi 353 data dengan total 12 fitur dan 1 variabel target. Variabel target kemudian ditransformasikan menjadi nilai numerik melalui metode *label encoding*, dengan pemetaan kelas BC sebagai 0 dan Non-BC sebagai 1. Hasil distribusi data setelah tahap *preprocessing* menunjukkan jumlah data BC sebanyak 168 data, sedangkan data Non-BC sebanyak 185 data. Hasil tahap *preprocessing* setelah proses data *cleaning*, *encoding*, dan *scaling* disajikan pada Tabel 3, yang menunjukkan perubahan jumlah data serta struktur dataset yang digunakan pada tahap pemodelan.

Tabel 3. Data Setelah Tahap *Preprocessing*

Komponen	Nilai
Total data setelah <i>cleaning</i>	353
BC setelah <i>cleaning</i>	168
Non-BC setelah <i>cleaning</i>	185

Dataset yang telah melalui proses *preprocessing* selanjutnya dipisahkan ke dalam data pelatihan dan data pengujian dengan proporsi 80% dan 20%. Pemisahan ini dilakukan untuk mengukur kemampuan model dalam melakukan generalisasi terhadap data yang belum pernah digunakan selama pelatihan. Dari proses tersebut diperoleh 282 data latih dan 71 data uji. Distribusi masing-masing subset disajikan pada Tabel 4.

Tabel 4. Pembagian Data Latih dan Data Uji (80:20)

Komponen	Nilai
Data latih	282
Data uji	71

Setelah *preprocessing*, dilakukan seleksi fitur menggunakan metode ReliefF untuk menentukan tingkat relevansi fitur terhadap klasifikasi status kanker payudara. Metode ReliefF menilai pentingnya suatu fitur berdasarkan kemampuannya membedakan data dari kelas yang sama (*nearest hit*) dan kelas berbeda (*nearest miss*). Nilai bobot yang lebih besar menunjukkan bahwa fitur tersebut memiliki pengaruh yang lebih kuat dalam proses klasifikasi.

Hasil ReliefF menunjukkan bahwa setiap fitur memiliki tingkat relevansi yang berbeda. Fitur *Breastfeeding*, *Marital Status*, *Highfat*, *Menopause*, dan *Age (years)* memperoleh bobot tertinggi, sedangkan *BMI* dan *Education* memiliki bobot rendah. Hasil pembobotan fitur ditunjukkan pada Tabel 5.

Tabel 5. Tabel Hasil Seleksi Fitur

Ranking	Fitur	Bobot
1	<i>Breastfeeding</i>	908.0
2	<i>Marital Status</i>	824.0
3	<i>Highfat</i>	794.0

Tabel 6. Hasil Evaluasi 5-Fold Cross Validation

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
<i>Logistic Regression</i>	0.7838	0.9156	0.6487	0.7580	0.8275
<i>Random Forest</i>	0.7447	0.8605	0.6152	0.7159	0.7841

Hasil evaluasi menunjukkan bahwa Logistic Regression memberikan performa yang lebih unggul dibandingkan model lainnya karena memperoleh nilai *F1-score* dan *AUC-ROC* tertinggi. Pemilihan model dilakukan berdasarkan hasil validasi silang untuk menghindari bias evaluasi terhadap data uji. Pada tahap pengujian menggunakan data *testing*, model *Logistic Regression* menghasilkan nilai *accuracy* sebesar 0,8310, *precision* sebesar 0,8788, *recall*

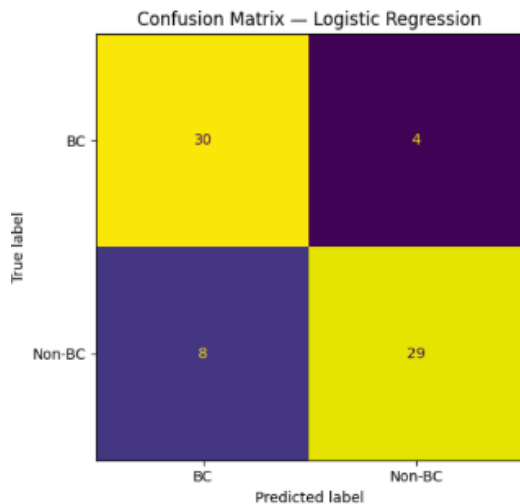
Ranking	Fitur	Bobot
4	<i>Menopause</i>	746.0
5	<i>Age (Years)</i>	722.0
6	<i>Parity</i>	694.0
7	<i>Ethnicity</i>	414.0
8	<i>First Pregnancy</i>	364.0
9	<i>Working Status</i>	294.0
10	<i>Menarche (Years)</i>	250.0
11	<i>BMI</i>	-32.0
12	<i>Education</i>	-250.0

Berdasarkan hasil pemeringkatan, dipilih tujuh fitur terbaik, yaitu *Breastfeeding*, *Marital Status*, *Highfat*, *Menopause*, *Age (years)*, *Parity*, dan *Ethnicity*, yang digunakan sebagai input pelatihan model. Pemilihan fitur ini bertujuan mengurangi dimensionalitas data serta meningkatkan efisiensi dan performa model. Kinerja model dievaluasi menggunakan skema *5-Fold Cross Validation* dengan membagi data pelatihan ke dalam lima subset pada algoritma *Logistic Regression* dan *Random Forest*. Pendekatan ini diterapkan untuk memperoleh hasil evaluasi yang lebih konsisten sekaligus meminimalkan bias akibat pembagian data. Berdasarkan hasil evaluasi, *Logistic Regression* menunjukkan performa yang lebih unggul dibandingkan *Random Forest*, khususnya pada metrik *F1-score* dan *AUC-ROC*. *Logistic Regression* menghasilkan nilai *accuracy*, *precision*, *recall*, *F1-score*, dan *AUC-ROC* masing-masing sebesar 0,7838; 0,9156; 0,6487; 0,7580; dan 0,8275. Sementara itu, *Random Forest* memperoleh nilai sebesar 0,7447; 0,8605; 0,6152; 0,7159; dan 0,7841 untuk metrik yang sama. Ringkasan hasil evaluasi kedua model pada tahap *5-Fold Cross Validation* disajikan pada Tabel 7.

sebesar 0,7838, *F1-score* sebesar 0,8286, dan *AUC-ROC* sebesar 0,8394, yang menunjukkan kemampuan model dalam melakukan klasifikasi status kanker payudara dengan baik pada data yang belum pernah dilihat sebelumnya. Hasil evaluasi model menggunakan data *testing* disajikan pada Tabel 8, yang menunjukkan performa akhir model pada data yang belum pernah dilihat selama proses pelatihan.

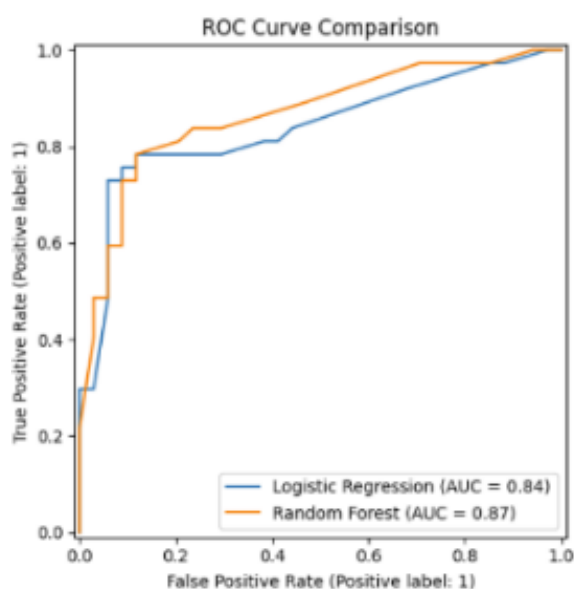
Tabel 7. Hasil Evaluasi Model Pada Data Testing

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
<i>Logistic Regression</i>	0.8310	0.8788	0.7838	0.8286	0.8394
<i>Random Forest</i>	0.8310	0.8788	0.7838	0.8286	0.8672



Gambar 3. Confusion Matrix Logistic Regression

Berdasarkan hasil *confusion matrix* pada model *Logistic Regression*, model menunjukkan kemampuan klasifikasi yang cukup baik pada kedua kelas. Sebanyak 30 data BC berhasil diklasifikasikan dengan benar sebagai BC (*true positive*), sedangkan 4 data BC salah diklasifikasikan sebagai Non-BC (*false negative*). Sebanyak 29 data pada kelas Non-BC berhasil diprediksi dengan benar sebagai Non-BC (*true negative*), sementara 8 data Non-BC terklasifikasi sebagai BC (*false positive*). Distribusi hasil pada *confusion matrix* mengindikasikan bahwa model mampu mengklasifikasikan sebagian besar data secara akurat. Hal tersebut didukung oleh capaian *accuracy* sebesar 83,10%, *precision* sebesar 87,88%, *recall* sebesar 78,38%, dan *F1-score* sebesar 82,86% pada data pengujian.

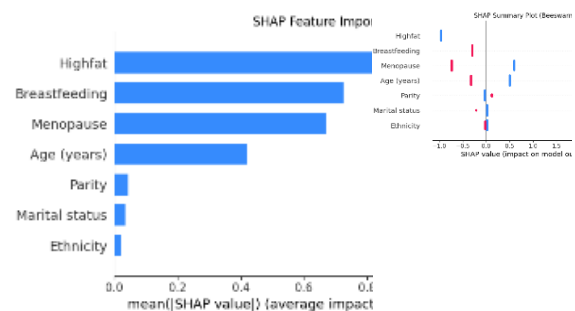


Gambar 4. Kurva AUC-ROC

Kemampuan model dalam membedakan kelas BC dan Non-BC selanjutnya dievaluasi menggunakan kurva *Receiver Operating Characteristic (ROC)*,

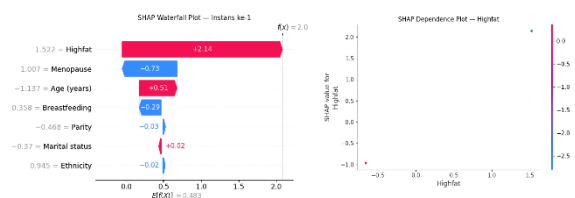
sebagaimana ditunjukkan pada Gambar 4. Hasil evaluasi menunjukkan bahwa model memperoleh nilai *AUC-ROC* sebesar 0,8394, yang mencerminkan kemampuan yang baik dalam membedakan kelas BC dan Non-BC. Nilai tersebut mengindikasikan bahwa model memiliki probabilitas sebesar 83,94% untuk memberikan skor prediksi yang lebih tinggi pada sampel BC dibandingkan sampel Non-BC. Dengan demikian, model mampu mengenali pola klasifikasi kanker payudara secara cukup baik, termasuk pada data yang tidak digunakan selama proses pelatihan.

Untuk memperoleh interpretasi terhadap keputusan model, dilakukan analisis menggunakan metode SHAP (*SHapley Additive exPlanations*) melalui *summary plot (beeswarm)*, *bar plot*, *waterfall plot*, dan *dependence plot*. Berdasarkan *summary plot* dan *bar plot* pada Gambar 5, fitur *Highfat* menunjukkan pengaruh terbesar terhadap proses klasifikasi status kanker payudara, diikuti oleh fitur lain yang memberikan kontribusi berbeda terhadap hasil prediksi model.



Gambar 5. SHAP Summary Plot & Feature Importance Bar

Selanjutnya, *waterfall plot* dan *dependence plot* pada Gambar 6 digunakan untuk menjelaskan kontribusi fitur pada prediksi individual serta hubungan antara nilai fitur dengan kontribusinya terhadap hasil prediksi. Hasil visualisasi menunjukkan bahwa kontribusi fitur dapat berbeda pada setiap sampel dan dipengaruhi oleh interaksi antar fitur, sehingga keputusan model menjadi lebih transparan dalam mendukung klasifikasi status kanker payudara.



Gambar 6. SHAP Waterfall Plot & Dependence Plot

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, metode *machine learning* dengan penerapan seleksi fitur *ReliefF* berhasil digunakan untuk klasifikasi status kanker payudara. Dari total 12 fitur, metode *ReliefF* menghasilkan 7 fitur terbaik, yaitu *Breastfeeding*, *Marital Status*, *Highfat*,

Menopause, *Age (years)*, *Parity*, dan *Ethnicity*, yang dinilai paling relevan terhadap proses klasifikasi. Berdasarkan hasil evaluasi *5-Fold Cross Validation*, model *Logistic Regression* menunjukkan performa lebih baik dibandingkan *Random Forest*, sehingga dipilih sebagai model terbaik. Pada tahap pengujian menggunakan data testing, model *Logistic Regression* memperoleh nilai *accuracy* sebesar 83,10%, *precision* sebesar 87,88%, *recall* sebesar 78,38%, *F1-score* sebesar 82,86%, dan *AUC-ROC* sebesar 83,94%, yang menunjukkan kemampuan model yang baik dalam membedakan kelas BC dan Non-BC. Selain menghasilkan performa klasifikasi yang baik, penerapan SHAP juga mampu menjelaskan pengaruh masing-masing fitur terhadap prediksi model, sehingga meningkatkan transparansi dan kemudahan interpretasi hasil klasifikasi kanker payudara.

Untuk penelitian selanjutnya, disarankan penggunaan dataset dengan jumlah dan keragaman data yang lebih besar guna meningkatkan kemampuan generalisasi model. Selain itu, penelitian dapat dikembangkan dengan mengevaluasi algoritma klasifikasi maupun metode seleksi fitur lainnya untuk memperoleh performa yang lebih optimal. Kajian interpretabilitas model juga dapat diperluas melalui penerapan metode *Explainable Artificial Intelligence* (XAI) lain sebagai pembandingan terhadap SHAP.

DAFTAR PUSTAKA

- [1] World Health Organization, "Breast Cancer," 2026. <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>
- [2] A. Ghasemi and A. S. Nejad, "Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review," *Computational and Applied Intelligence*, 2024, doi:10.1002/cai2.136
- [3] P. C. Austin, F. E. Harrell Jr., and E. W. Steyerberg, "Predictive performance of machine and statistical learning methods: Impact of data-generating processes on external validity," *Statistics in Medicine*, vol. 40, no. 26, pp. 5444–5462, 2021, doi:10.1177/09622802211002867
- [4] M. A. Ahmed et al., "Predicting cancer risk using machine learning on lifestyle and genetic data," *Scientific Reports*, 2025, doi:10.1038/s41598-025-15656-8
- [5] K. Kourou et al., "Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 5546–5555, 2021, doi:10.1016/j.csbj.2021.10.006
- [6] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021, doi:10.1109/TNNLS.2020.3027314
- [7] T. Hulsén, "Explainable artificial intelligence (XAI): Concepts and challenges in healthcare," *Applied Sciences*, vol. 14, no. 3, 2024, doi:10.3390/app14030334
- [8] X. Wang and T. Kee, "Integrating Shapley Value and Least Core Attribution for Robust Explainable AI in Rent Prediction," *Buildings*, vol. 15, no. 17, 2025, doi:10.3390/buildings15173133
- [9] Y. Hasan et al., "Improving breast cancer diagnosis using grammatical evolution-based feature selection," *SN Computer Science*, 2025, doi:10.1007/s42979-025-03840-6
- [10] A. Tijjani et al., "Interpretable machine learning approach for breast cancer classification," *Human-Centric Intelligent Systems*, vol. 5, no. 3, pp. 308–322, 2025, doi:10.1007/s44230-025-00111-8
- [11] S. Sahana and H. K. Madhu, "Breast cancer prediction using machine learning," *IARJSET*, vol. 11, no. 7, pp. 170–173, 2024, doi:10.17148/IARJSET.2024.11723
- [12] J. Kamwele et al., "Explainable AI for breast cancer diagnosis: Comparative analysis of ML models using Random Forest feature selection and SHAP interpretability," 2025 (preprint)
- [13] G. Kallah-Dagadu et al., "Breast cancer prediction based on gene expression data using interpretable machine learning techniques," *Scientific Reports*, 2025
- [14] V. Borisov et al., "Deep Neural Networks and Tabular Data: A Survey," 2022
- [15] Q. An et al., "A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges," 2023
- [16] J. Sen, *Machine Learning*, vol. 7, Springer, 2023
- [17] OpenStax, *Principles of Data Science*, OpenStax (Rice University), 2023
- [18] T. Islam et al., "Predictive modeling for breast cancer classification using explainable AI," *Scientific Reports*, 2024, doi:10.1038/s41598-024-57740-5
- [19] J. Sulaksono et al., "Analisis perbandingan model machine learning untuk prediksi penyakit jantung," *STAINs Proceedings*, 2026
- [20] L. Barreñada et al., "Understanding overfitting in random forest for probability estimation: a visualization and simulation study," *Diagnostic and Prognostic Research*, 2024, doi:10.1186/s41512-024-00177-1
- [21] J. Li et al., "Self-attention random forest for breast cancer image classification," *Frontiers in Oncology*, 2023, doi:10.3389/fonc.2023.1043463
- [22] M. Buyukkececi, "A comprehensive review of feature selection and feature selection stability in machine learning," *Gaziantep University Journal of Science*, vol. 36, no. 4, pp. 1506–1520, 2023
- [23] S. Pau et al., "An evaluation of feature selection robustness on class noisy data," *Information*, vol. 14, no. 8, 2023

- [24] K. M. Rajabi, "Penerapan algoritma KNN dengan fitur Relief-F dalam penentuan status stunting," *Innovative Journal*, 2023
- [25] A. A. Noor and M. A. Qureshi, "Unveiling explainable AI in healthcare: Current trends, challenges, and future directions," *WIREs Data Mining and Knowledge Discovery*, 2025, doi:10.1002/widm.70018
- [26] X. Wang and T. Kee, "Integrating Shapley Value and Least Core Attribution for Robust Explainable AI in Rent Prediction," *Buildings*, 2025
- [27] M. Wanyonyi et al., "Machine Learning Models for Predicting Stroke Risk Among Patients with Coronary Heart Disease," *medRxiv*, 2026 (preprint)
- [28] R. Dana et al., "Dataset of Indonesian women's reproductive, high-fat diet and body mass index risk factors for breast cancer," *Data in Brief*, vol. 36, p. 107107, 2021, doi:10.1016/j.dib.2021.107107
- [29] H. Bichri et al., "Investigating the impact of train/test split ratio on the performance of pre-trained models with custom datasets," 2024 (preprint)