**Journal of Science and Social Research** ISSN 2615 – 4307 (Print)
Oct 2026, IX (5): 9395 – 9403 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

# PERBANDINGAN KINERJA SUPPORT VECTOR MACHINE, RANDOM FOREST, DAN LOGISTIC REGRESSION DALAM PREDIKSI KANKER PAYUDARA

**Rustianna Tumanggor[1], Septian Simatupang[2], Dior Manta Tambunan[3]**
**[1,3]Universitas Murni Teguh**
**[2]Politeknik Wilmar Bisnis Indonesia**
e-mail: [1]tumanggorrusti@gmail.com, [2]septian.simatupang@wbi.ac.id
[3]dior.endlessbay@gmail.com

***Abstract:*** *Breast cancer is one of the most common cancers among women, making early detection essential for improving treatment outcomes. Machine learning can support breast cancer classification by identifying patterns in medical data. This study compares Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) to identify the model with the best performance and stability. The study used a breast cancer dataset comprising 569 samples and 30 numerical features. The research process included data preprocessing, removal of irrelevant attributes, feature transformation and standardization, class imbalance handling using the Synthetic Minority Over-sampling Technique (SMOTE), feature selection using SelectKBest with an ANOVA F-test, and hyperparameter optimization. Model performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and cross-validation. Test results showed that SVM and LR achieved 98.25% accuracy, 97.62% precision, 97.62% recall, and a 97.62% F1-score. SVM achieved the highest ROC-AUC at 99.74%, followed by LR at 99.70%. RF obtained 95.61% accuracy and 90.48% recall. LR demonstrated the best cross-validation stability. Therefore, LR offers a balanced candidate for a decision support system prototype, while SVM provides superior discrimination on test data. External validation is required before clinical application.*

***Keyword:*** *Breast Cancer; Machine Learning; Support Vector Machine; Random Forest; Logistic Regression.*

**Abstrak:** Kanker payudara merupakan salah satu jenis kanker yang paling umum di kalangan wanita, sehingga deteksi dini sangat penting untuk meningkatkan hasil pengobatan. Pembelajaran mesin dapat mendukung klasifikasi kanker payudara dengan mengidentifikasi pola dalam data medis. Penelitian ini membandingkan Support Vector Machine (SVM), Random Forest (RF), dan Regresi Logistik (LR) untuk mengidentifikasi model dengan kinerja dan stabilitas terbaik. Penelitian ini menggunakan dataset kanker payudara yang terdiri dari 569 sampel dan 30 fitur numerik. Proses penelitian meliputi prapemrosesan data, penghapusan atribut yang tidak relevan, transformasi dan standarisasi fitur, penanganan ketidakseimbangan kelas menggunakan Teknik Synthetic Minority Over-sampling (SMOTE), seleksi fitur menggunakan SelectKBest dengan uji F ANOVA, serta optimisasi hiperparameter. Kinerja model dievaluasi menggunakan akurasi, presisi, recall, skor F1, AUC kurva ROC, dan validasi silang. Hasil pengujian menunjukkan bahwa SVM dan LR mencapai akurasi 98,25%, presisi 97,62%, recall 97,62%, dan skor F1 sebesar 97,62%. SVM mencapai nilai ROC-AUC tertinggi sebesar 99,74%, diikuti oleh LR sebesar 99,70%. RF memperoleh akurasi sebesar 95,61% dan recall sebesar 90,48%. LR menunjukkan stabilitas validasi silang terbaik. Oleh karena itu, LR menawarkan kandidat yang seimbang untuk prototipe sistem pendukung keputusan, sementara SVM memberikan kemampuan diskriminasi yang lebih unggul pada data uji. Validasi eksternal diperlukan sebelum penerapan klinis.

**Kata kunci:** Kanker Payudara; Machine Learning; Support Vector Machine; Random

Forest; Logistic Regression.

## PENDAHULUAN

Kanker payudara merupakan salah satu masalah kesehatan global dan termasuk jenis kanker yang paling sering didiagnosis pada perempuan. Statistik kanker global melaporkan sekitar 2,3 juta kasus baru kanker payudara dan 685.000 kematian di seluruh dunia pada tahun 2020, yang menunjukkan besarnya beban penyakit ini terhadap sistem kesehatan masyarakat [1]. Di Indonesia, kanker payudara juga menjadi salah satu jenis kanker yang paling umum pada perempuan [2]. Keterlambatan diagnosis dan penemuan kanker pada stadium lanjut dapat mempersulit proses pengobatan serta berdampak negatif terhadap luaran pasien [3]. Oleh karena itu, peningkatan akurasi dan keandalan klasifikasi kanker payudara menjadi penting untuk mendukung deteksi dini dan penilaian klinis yang tepat waktu.

Prosedur diagnosis kanker payudara konvensional, seperti mamografi, ultrasonografi, dan biopsi, memiliki peran penting dalam pemeriksaan klinis. Namun, proses diagnosis dapat menghadapi berbagai tantangan terkait aksesibilitas, biaya, dan interpretasi hasil pemeriksaan medis [4]. Perkembangan machine learning (ML) membuka peluang untuk menganalisis data medis dan mengidentifikasi pola yang dapat membantu membedakan tumor jinak (benign) dan tumor ganas (malignant). Sejumlah penelitian telah mengkaji penerapan ML dalam klasifikasi kanker payudara dan menunjukkan potensi pendekatan tersebut untuk mendukung proses diagnosis berbasis data [5–7]. Meskipun demikian, kinerja model dapat bervariasi bergantung pada karakteristik dataset, prosedur prapemrosesan, metode seleksi fitur, dan strategi evaluasi yang digunakan [8–12].

Tiga algoritma yang banyak digunakan dalam penelitian klasifikasi adalah Support Vector Machine (SVM), Random Forest (RF), dan Logistic Regression (LR). SVM membentuk batas keputusan untuk memisahkan kelas, sedangkan RF menggabungkan sejumlah pohon keputusan untuk menghasilkan prediksi. Sementara itu, LR memperkirakan probabilitas keanggotaan kelas dan menawarkan pendekatan pemodelan yang relatif mudah diinterpretasikan. Walaupun ketiga algoritma tersebut telah diterapkan pada klasifikasi kanker payudara, perbedaan rancangan eksperimen dan prosedur evaluasi dalam penelitian terdahulu menyulitkan perbandingan secara langsung. Oleh sebab itu, pemilihan algoritma yang sesuai tidak cukup hanya didasarkan pada nilai akurasi tertinggi, tetapi juga perlu mempertimbangkan kinerja berdasarkan berbagai metrik evaluasi dan konsistensi model selama proses validasi.

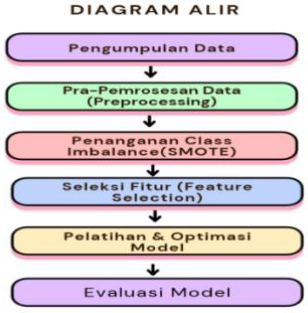
Aspek metodologis lain yang perlu diperhatikan adalah evaluasi model klasifikasi pada dataset dengan distribusi kelas yang tidak seimbang. Akurasi saja belum tentu menggambarkan kemampuan model dalam mengidentifikasi kasus ganas secara memadai, terutama karena kesalahan klasifikasi dapat memiliki konsekuensi yang berbeda dalam konteks pemeriksaan klinis. Presisi, recall, F1-score, dan luas area di bawah kurva karakteristik operasi penerima (Receiver Operating Characteristic–Area Under the Curve atau ROC-AUC) memberikan informasi yang saling melengkapi mengenai kinerja prediksi. Selain itu, validasi silang dapat digunakan untuk menilai konsistensi kinerja model pada beberapa pembagian data. Dengan demikian, perbandingan sistematis yang menggabungkan berbagai metrik tersebut dapat memberikan dasar yang lebih komprehensif dalam menentukan model yang layak dikaji lebih lanjut.

Penelitian ini bertujuan membandingkan kinerja SVM, RF, dan LR dalam klasifikasi kanker payudara

menggunakan dataset yang terdiri atas 569 sampel dan 30 fitur numerik. Tahapan eksperimen meliputi prapemrosesan data, transformasi dan standardisasi fitur, penanganan ketidakseimbangan kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE), seleksi fitur menggunakan SelectKBest dengan ANOVA F-test, serta optimasi hiperparameter. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, F1-score, ROC-AUC, dan validasi silang berstrata lima lipatan (stratified five-fold cross-validation). Kontribusi utama penelitian ini adalah menyajikan evaluasi komparatif dengan prosedur eksperimen yang konsisten serta mempertimbangkan kinerja prediksi dan stabilitas model, bukan hanya akurasi. Hasil penelitian diharapkan dapat mengidentifikasi kandidat model untuk pengembangan prototipe sistem pendukung keputusan sebagai proof of concept. Model yang dihasilkan masih merupakan prototipe penelitian dan tidak dimaksudkan untuk menggantikan diagnosis oleh tenaga medis profesional.

## METODE

Metode penelitian ini dirancang secara sistematis melalui beberapa tahapan utama, mulai dari pengumpulan data, pra-pemrosesan data, penanganan ketidakseimbangan kelas (class imbalance), seleksi fitur, pelatihan dan optimasi model, hingga evaluasi dan implementasi model ke dalam aplikasi berbasis web.




**Gambar 1 Diagram Alir Tahapan Penelitian**

### Sumber dan Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini adalah Wisconsin Breast Cancer Dataset (WBCD) yang diperoleh dari UCI Machine Learning Repository. Dataset ini terdiri dari 569 sampel citra hasil Fine Needle Aspiration (FNA) dari massa payudara. Dataset memiliki 30 fitur numerik sebagai prediktor dan 1 variabel target (diagnosis) yang terbagi menjadi dua kelas:
Benign (B / Jinak): 357 sampel (62,7%)
Malignant (M / Ganas): 212 sampel (37,3%)
Fitur-fitur numerik dikelompokkan menjadi tiga kategori utama (masing-masing 10 karakteristik sel):
Mean (_mean): Nilai rata-rata fitur inti sel.
Standard Error (_se): Standar eror pengukuran fitur.
Worst (_worst): Nilai terburuk/terbesar (rata-rata dari tiga inti sel terbesar).

### Pra-Pemrosesan Data (Data Preprocessing)

Pra-pemrosesan data dilakukan untuk memastikan kualitas masukan model dan mencegah data leakage:

Pembersihan Data (Data Cleaning): Kolom id dihapus karena tidak memiliki hubungan kausalitas dengan prediktor. Kolom Unnamed: 32 yang berisi nilai kosong (null) secara keseluruhan dihapus. Nilai pencilan (outliers) pada kelompok fitur _mean dan _worst tetap dipertahankan karena merupakan penanda diagnostik biologis yang valid untuk kasus Malignant.

Encoding Variabel Target: Variabel target diagnosis dikodekan menjadi nilai biner menggunakan LabelEncoder, yaitu 0 untuk Benign (Jinak) dan 1 untuk Malignant (Ganas).

Pembagian Dataset (Data Splitting): Dataset dibagi menjadi data latih (training set) sebesar 80% (455 sampel) dan data uji (test set) sebesar 20% (114 sampel) menggunakan Stratified Sampling (stratify=y) dengan random_state=42 untuk menjaga proporsi rasio kelas target tetap seimbang di kedua

subset data.

Skala Fitur (Feature Scaling): StandardScaler diterapkan pada 20 fitur kelompok _mean dan _worst yang terdistribusi mendekati normal untuk mengukur nilai z-score (μ=0, σ=1). PowerTransformer (Metode Yeo-Johnson) diterapkan pada 10 fitur kelompok _se yang memiliki distribusi miring ke kanan (right-skewed) guna mentransformasikan data mendekati distribusi Gaussian. Penyesuaian pemroses (scaler fit) hanya dilakukan pada training set, kemudian ditransformasikan ke training set dan testing set.

**Penanganan Ketidakseimbangan Kelas (Class Imbalance)**

Untuk mengatasi ketidakseimbangan kelas pada data latih (285 Benign vs 170 Malignant), diterapkan teknik SMOTE (Synthetic Minority Over-sampling Technique). SMOTE memunculkan sampel sintetis baru pada kelas minoritas (Malignant) melalui interpolasi k-nearest neighbors (k=5), sehingga jumlah data latih menjadi seimbang (285 Benign dan 285 Malignant). SMOTE hanya diterapkan pada data latih untuk mencegah data leakage pada evaluasi data uji.

**Seleksi Fitur (Feature Selection)**

Seleksi fitur bertujuan mengurangi dimensi data, mengurangi redundansi fitur yang berkorelasi tinggi (seperti radius, perimeter, dan area), serta mempercepat waktu komputasi. Metode yang digunakan adalah SelectKBest berbasis uji statistik ANOVA F-test (f_classif). Dari 30 fitur awal, dipilih k = 15 fitur terbaik dengan nilai rasio variansi antar-kelas (F-score) tertinggi. Seleksi fitur di-fit pada data latih yang telah diseimbangkan dengan SMOTE.

**Formulasi Matematis dan Pelatihan Model**

**Logistic Regression (LR)**

Logistic Regression adalah algoritma klasifikasi probabilistik linier yang memetakan kombinasi linier dari fitur terintegrasi ke dalam rentang probabilitas [0, 1] menggunakan fungsi sigmoid (σ). Fungsi prediksi probabilitas untuk variabel target biner $y \in \{0, 1\}$ didefinisikan sebagai:

$P(y = 1 \mid x) = \sigma(w^T x + b) = 1 / (1 + e^{-(w^T x + b)})$

Di mana $x = [x_1, x_2, ..., x_k]^T$ adalah vektor fitur hasil seleksi ANOVA (k = 15), $w = [w_1, w_2, ..., w_k]^T$ adalah vektor bobot (weights), dan b adalah nilai bias (intercept).

Berdasarkan optimasi hyperparameter, model ini menggunakan penalti $L_1$ (Lasso) dengan algoritma solver SAGA. Fungsi kerugian (loss function) yang dioptimalkan didefinisikan sebagai:

$\min_{w, b} [ - (1/N) \sum_{i=1}^{N} (y_i \log(P(y_i=1|x_i)) + (1-y_i) \log(1-P(y_i=1|x_i))) + (1/C) \sum_{i=1}^{k} |w_j| ]$

Di mana N adalah jumlah sampel data latih, dan C = 10 adalah parameter regularisasi invers.

**Support Vector Machine (SVM)**

Support Vector Machine bertujuan untuk menemukan hyperplane pemisah optimal yang memaksimumkan jarak marjin antara dua kelas (Benign dan Malignant). Dengan menggunakan kernel linear, persamaan hyperplane didefinisikan sebagai:

$f(x) = w^T x + b = 0$

Fungsi keputusannya dinyatakan sebagai:

$\hat{y} = \text{sign}(w^T x + b)$

Untuk menangani data yang tidak terseparasi secara sempurna di ruang asal, digunakan pendekatan soft-margin optimization dengan meminimalkan fungsi tujuan (objective function) berikut:

$\min_{w, b, \xi} ( (1/2) ||w||^2 + C \sum_{i=1}^{N} \xi_i )$

Dengan kendala (constraints): $y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i \in \{1, ..., N\}$

Di mana $y_i \in \{-1, +1\}$ adalah label kelas (Benign = -1, Malignant = +1), $\xi_i$ adalah slack variable untuk mengizinkan toleransi kesalahan klasifikasi pada sampel tertentu, dan C = 8.288 adalah parameter penalti margin kesalahan hasil

optimasi.

**Random Forest Classifier**

Random Forest adalah algoritma ensemble learning berbasis metode Bagging (Bootstrap Aggregating) yang menggabungkan banyak pohon keputusan (Decision Trees) secara independen.

Pembagian Node (Split Criterion): Setiap node pada pohon keputusan dievaluasi menggunakan indeks Gini Impurity (I_G) untuk mengukur tingkat ketidakmurnian sampel:

$I_G(p) = 1 - \sum_{i=1}^{c} (p_i)^2$

Di mana $p_i$ adalah proporsi sampel dari kelas i pada node tersebut (C_class = 2).

Agregasi Prediksi (Majority Voting): Prediksi akhir ensemble untuk sampel masukan x ditentukan melalui pengambil suara terbanyak (majority voting) dari seluruh pohon T = 369 (n_estimators):

$\hat{y} = \text{mode} \{ h_1(x), h_2(x), ..., h_T(x) \}$

Di mana h_t(x) adalah prediksi klasifikasi dari pohon keputusan ke-t. Pada setiap pemisahan node, fitur kandidat dipilih secara acak sebanyak $m = \log_2(k)$ fitur (max_features='log2', di mana k=15).

**Optimasi Hyperparameter**

Proses optimasi hyperparameter dilakukan menggunakan RandomizedSearchCV dengan 5-Fold Stratified Cross-Validation pada data latih. Metrik optimasi utama yang digunakan adalah F1-Score. Ruang pencarian hyperparameter terbaik yang diperoleh disajikan pada Tabel 1.

**Tabel 1 Parameter Hasil Optimasi Model**

| Algoritma | Hyperparameter Terbaik |
|---|---|
| Logistic Regression | solver='saga', penalty='l1', C=10, max_iter=5000 |
| Support Vector Machine | kernel='linear', C=8.288 |
| Random Forest | n_estimators=369, max_depth=5, max_features='log2', min_samples_split=5, min_samples_leaf=6, bootstrap=False |

**Evaluasi Model (Model Evaluation)**

Kinerja model dievaluasi pada test set (114 sampel: 72 Benign, 42 Malignant). Evaluasi dilakukan menggunakan metrik matriks konfusi (Confusion Matrix) dan standar evaluasi klasifikasi medis:

Accuracy: Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision: Precision = TP / (TP + FP)

Recall (Sensitivity): Recall = TP / (TP + FN)

F1-Score: F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

ROC-AUC: Luas area di bawah kurva Receiver Operating Characteristic untuk mengukur kemampuan diskriminasi probabilistik model.

## HASIL DAN PEMBAHASAN

Pembahasan mencakup analisis karakteristik data, pengaruh pra-pemrosesan data dan penanganan ketidakseimbangan kelas, hasil seleksi fitur, perbandingan performa ketiga algoritma (Logistic Regression, Support Vector Machine, dan Random Forest).

**Hasil Seleksi Fitur (ANOVA F-test)**

Seleksi fitur menggunakan ANOVA F-test (SelectKBest) berhasil mereduksi dimensi data dari 30 fitur awal menjadi 15 fitur paling kontributif. Langkah ini terbukti efektif menghilangkan redundansi tinggi antar-fitur geometris, seperti hubungan linier yang kuat antara radius, perimeter, dan area. Tabel 2 menampilkan daftar 15 fitur terbaik beserta kelompok kategorinya.

**Tabel 2 Fitur Hasil Seleksi ANOVA F-test (k=15)**

| No | Nama Fitur | Kelompok Fitur |
|---|---|---|

| | | |
|---|---|---|
| 1 | concave points_worst | Worst (Ekstrim) |
| 2 | perimeter_worst | Worst (Ekstrim) |
| 3 | concave points_mean | Mean (Rata-rata) |
| 4 | radius_worst | Worst (Ekstrim) |
| 5 | perimeter_mean | Mean (Rata-rata) |
| 6 | area_worst | Worst (Ekstrim) |
| 7 | radius_mean | Mean (Rata-rata) |
| 8 | area_mean | Mean (Rata-rata) |
| 9 | concavity_mean | Mean (Rata-rata) |
| 10 | concavity_worst | Worst (Ekstrim) |
| 11 | compactness_mean | Mean (Rata-rata) |
| 12 | compactness_worst | Worst (Ekstrim) |
| 13 | radius_se | Standard Error (Variabilitas) |
| 14 | perimeter_se | Standard Error (Variabilitas) |
| 15 | area_se | Standard Error (Variabilitas) |

**Evaluasi dan Perbandingan Performa Model**

Setiap model dievaluasi pada data uji (test set) independen yang terdiri dari 114 sampel (72 Benign dan 42 Malignant). Hasil pengujian dari ketiga algoritma yang telah dioptimasi disajikan pada Tabel 3.

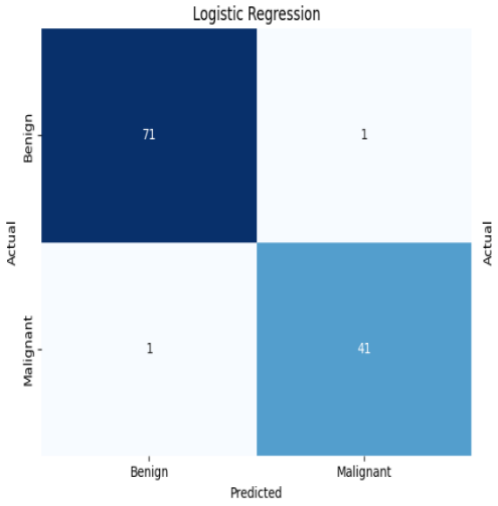
**Tabel 3 Perbandingan Kinerja Klasifikasi Model pada Data Uji**

| Model Algoritma | Akurasi | Presisi | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 98,25% | 97,62% | 97,62% | 97,62% | 0,9967 |
| Support Vector Machine | 97,37% | 95,35% | 97,62% | 96,47% | 0,9954 |
| Random Forest | 96,49% | 95,24% | 95,24% | 95,24% | 0,9921 |

**Analisis Matriks Konfusi (Confusion Matrix)**

Analisis terperinci terhadap matriks konfusi memberikan gambaran mendalam mengenai tingkat kesalahan prediktif pada ranah medis:

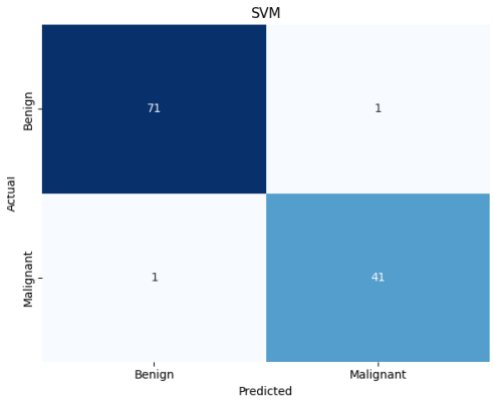
**Logistic Regression (Model Terbaik):** Memiliki matriks konfusi dengan True Negative (TN) = 71, False Positive (FP) = 1, False Negative (FN) = 1, dan True Positive (TP) = 41. Model ini hanya menghasilkan 1 kesalahan deteksi False Negative (kasus kanker ganas yang terprediksi jinak).




**Gambar 1 Matriks Konfusi Logistic Regression**

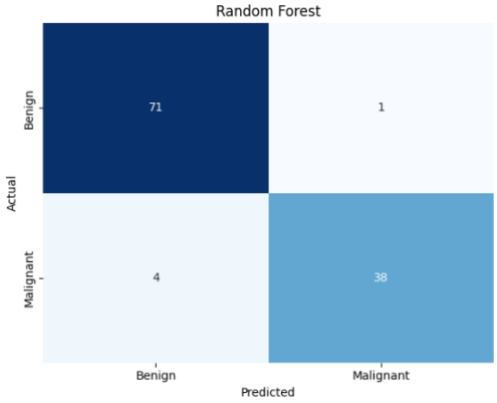
**Support Vector Machine:** Menghasilkan TN = 70, FP = 2, FN = 1, dan TP = 41. Meskipun memiliki Sensitivity/Recall yang sama tingginya dengan Logistic Regression (97,62%), SVM menghasilkan 2 kasus False Positive.

**Journal of Science and Social Research** ISSN 2615 – 4307 (Print)
Oct 2026, IX (5): 9395 – 9403 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR




**Gambar 2 Matriks Konfusi Support Vektor Machine**

**Random Forest:** Menghasilkan TN = 70, FP = 2, FN = 2, dan TP = 40. Model ensemble ini menghasilkan kesalahan deteksi False Negative terbanyak (2 sampel), yang membuat nilai Recall-nya berada di angka 95,24%.




**Gambar 3 Matriks Konfusi Random Forest**

**Pembahasan Medis dan Komputasional**
**Efektivitas Preprocessing Komposisional dan SMOTE**

Penerapan kombinasi StandardScaler untuk fitur berskala normal dan PowerTransformer (Yeo-Johnson) untuk fitur miring ke kanan (right-skewed) terbukti secara signifikan meningkatkan stabilitas konvergensi model linier (Logistic Regression dan SVM). Selain itu, penanganan ketidakseimbangan kelas menggunakan SMOTE pada data latih berhasil mencegah kecenderungan model memprediksi kelas mayoritas (Benign). Hal ini terbukti dari nilai Recall ketiga model yang mampu mencapai rentang 95,24% hingga 97,62%.

**Keunggulan Logistic Regression dengan Penalti L1**

Logistic Regression dengan penalti L1 (Lasso) dan solver SAGA mencapai kinerja terbaik (Akurasi 98,25% dan ROC-AUC 0,9967). Penalti L1 melakukan seleksi fitur secara implisit dengan memaksa bobot fitur yang kurang relevan menjadi tepat nol. Sifat model linier yang memisahkan boundary secara halus (smooth decision boundary) terbukti lebih resisten terhadap overfitting dibandingkan Random Forest pada dataset ukuran sedang seperti WBCD.

**Signifikansi Klinis Evaluasi Metrik Medis**

Dalam diagnosis kanker medis, meminimalkan kesalahan False Negative (FN) jauh lebih krusial dibandingkan False Positive (FP). Kesalahan FN berarti pasien pengidap kanker ganas tidak terdeteksi secara dini, yang berisiko fatal terhadap keselamatan pasien. Model Logistic Regression berhasil menekan kesalahan FN hingga hanya 1 kasus (Recall 97,62%), sekaligus menjaga presisi tinggi (97,62%).

## SIMPULAN

Berdasarkan penelitian yang telah dilakukan, dapat ditarik beberapa kesimpulan utama sebagai berikut:

1. Efektivitas Alur Pra-Pemrosesan dan Seleksi Fitur: Penerapan skala fitur berbasis karakteristik distribusi (StandardScaler untuk data berdistribusi normal dan PowerTransformer/Yeo-Johnson untuk data miring) bersama penanganan class imbalance menggunakan SMOTE pada data latih berhasil meningkatkan kualitas masukan model. Selain itu, seleksi fitur menggunakan ANOVA F-test (k=15) terbukti efektif mereduksi dimensi data hingga 50% tanpa

mengorbankan informasi diagnostik krusial.

2. Performa Model Terbaik: Algoritma Logistic Regression dengan regularisasi $L_1$ (Lasso) dan solver SAGA menghasilkan kinerja klasifikasi tertinggi dibandingkan Support Vector Machine (SVM) dan Random Forest. Model ini mencapai nilai Akurasi sebesar 98,25%, Presisi 97,62%, Recall (Sensitivity) 97,62%, F1-Score 97,62%, serta nilai ROC-AUC sebesar 0,9967 pada data uji.
3. Signifikansi Klinis Diagnostik: Dari perspektif medis, model Logistic Regression berhasil meminimalkan risiko False Negative hingga hanya 1 kasus dari 42 sampel tumor ganas (Malignant). Hal ini sangat penting untuk mencegah terjadinya kegagalan deteksi dini kanker payudara yang berpotensi fatal bagi keselamatan pasien.
4. Implementasi Sistem: Integrasi model Logistic Regression beserta artefak pipeline ke dalam antarmuka berbasis web Streamlit memberikan solusi alat bantu keputusan klinis (Clinical Decision Support System) yang interaktif, responsif, dan dapat memberikan estimasi risiko diagnosis tumor secara real-time bagi tenaga medis.

## DAFTAR PUSTAKA

Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast. 2022;66:15–23. doi: 10.1016/j.breast.2022.08.010. https://doi.org/10.1016/j.breast.2022.08.010

Momenimovahed Z, Salehiniya H. Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer (Dove Med Press). 2019;11:151–64. doi: 10.2147/BCTT.S176070.

Momenimovahed Z, Salehiniya H. Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer (Dove Med Press). 2019;11:151–64. doi: 10.2147/BCTT.S176070. https://doi.org/10.2147/BCTT.S176070

Patrício M, Pereira J, Crisóstomo J, Matafome P, Gomes M, Seiça R, et al. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer. 2018;18:29. doi: 10.1186/s12885-017-3877-1. https://doi.org/10.1186/s12885-017-3877-1

Ed-daoudy A, Maalmi K. Breast cancer classification with reduced feature set using association rules and support vector machine. Netw Model Anal Health Inform Bioinforma. 2020;9:34. doi: 10.1007/s13721-020-00237-8. https://doi.org/10.1007/s13721-020-00237-8

Vaka AR, Soni B, Reddy SK. Breast cancer detection by leveraging machine learning. ICT Express. 2020;6(4):320–4. doi: 10.1016/j.icte.2020.04.009. https://doi.org/10.1016/j.icte.2020.04.009

Shanbehzadeh M, Kazemi-Arpanahi H, Bolbolian Ghalibaf M, Orooji A. Performance evaluation of machine learning for breast cancer diagnosis: A case study. Informatics in Medicine Unlocked. 2022;31:101009. doi: 10.1016/j.imu.2022.101009. https://doi.org/10.1016/j.imu.2022.101009

Rashed AEE, Elmorsy AM, Atwa AEM. Comparative evaluation of automated machine learning techniques for breast cancer diagnosis. Biomed Signal Process Control. 2023;86:105016. doi: 10.1016/j.bspc.2023.105016. https://doi.org/10.1016/j.bspc.2023.105016

Naji MA, El Filali S, Aarika K, Benlahmar EH, Abdelouhahid RA,

Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Comput Sci. 2021;191:487–92. doi: 10.1016/j.procs.2021.07.062. https://doi.org/10.1016/j.procs.2021.07.062

Arooj S, Atta-ur-Rahman, Zubair M, Khan MF, Alissa K, Khan MA, et al. Breast cancer detection and classification empowered with transfer learning. Front Public Health. 2022;10:924432. doi: 10.3389/fpubh.2022.924432. https://doi.org/10.3389/fpubh.2022.924432

Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing. 2020;408:189–215. doi: 10.1016/j.neucom.2019.10.118.

Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput Sci. 2016;83:1064–9. doi: 10.1016/j.procs.2016.04.224. https://doi.org/10.1016/j.procs.2016.04.224

Andriansyah D, Fridayanthie EW. Optimization of Support Vector Machine and XGBoost methods using feature selection to improve classification performance. J Inform Telecommun Eng. 2023;6(2). doi: 10.31289/jite.v6i2.8373. https://doi.org/10.31289/jite.v6i2.837

Azis AIS, Idris ISK, Santoso B, Mustofa YA. Pendekatan machine learning yang efisien untuk prediksi kanker payudara. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi). 2019;3(3):458–69. doi: 10.29207/resti.v3i3.1347. https://doi.org/10.29207/resti.v3i3.13

Chairunisa R, Adiwijaya K, Astuti W. Perbandingan CART dan Random Forest untuk deteksi kanker berbasis klasifikasi data microarray. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi). 2020; 4(5): 805–12. doi: 10. 29207/ resti. v4i5.2083. https://doi.org/10.29207/resti.v4i5.2083

Resmiati R, Arifin T. Klasifikasi pasien kanker payudara menggunakan metode Support Vector Machine dengan Backward Elimination. Sistemasi. 2021;10(2):381. doi:10.32520/stmsi.v10i2.1238. https://doi.org/10.32520/stmsi.v10i2.1238

Kusuma J, Hayadi BH, Wanayumini W, Rosnelly R. Komparasi metode Multi Layer Perceptron (MLP) dan Support Vector Machine (SVM) untuk klasifikasi kanker payudara. MIND Journal. 2022;7(1):51–60. doi: 10.26760/mindjournal.v7i1.51-60. https://doi.org/10.26760/mindjournal.v7i1.51-60

Rahmayani OS, Permana RH, Witdiawati W. Early detection of breast cancer according to fertile age women. Media Keperawatan Indonesia. 2020;3(1):32–7. doi: 10.26714/mki.3.1.2020.32-37. https://doi.org/10.26714/mki.3.1.2020.32-37

Simatupang S, Ramadhansyah R., Tumanggor R., Tan ER., Fajar SA. Prediction of Heart Disease Risk Based on Patient Health History Using the Support Vector Machine (SVM) Algorithm. Zero : Jurnal Sains, Matematika, dan Terapan.

Rustianna Tumanggor, Theresia Anggriani H, Elshadddai Tabitha Siadari, Dita Togatorop, Dina Eudia Meha. Pelatihan Sadari Dan Pemeriksaan IVA Sebagai Upaya Deteksi Dini Kanker Di Wilayah Perdesaan: Studi Kasus Di Desa Tanjung Anom. Sevaka [Internet]. 2025 Aug. 4 [cited 2025 Dec. 29];3(3):1-10.