JURNAL SINTAKS LOGIKA Vol. 6 No. 1, Januari 2026, pp. 69-78, **eISSN : 2775-412X**
https://jurnal.umpar.ac.id/index.php/sylog

Prediksi Kanker Payudara Berbasis Machine Learning Dengan Analisis Probabilitas Klasisfikasi

Luthfi Ardyansyah[1*], Bambang Irawan[2]

[1*,2] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhadi Setiabudi
Email : luthfiardyansyah155@gmail.com

Abstract:
Breast cancer is one of the diseases with a high mortality rate in women, so early detection is crucial to increase the chances of recovery. Unfortunately, conventional methods of diagnosis still rely on the interpretation of medical personnel and laboratory procedures which are time-consuming and costly. This study tries to present a machine learning-based approach to predict breast cancer, while adding a classification probability analysis to make the prediction more informative. The breast cancer dataset was used to train four models, namely Logistic Regression, Support Vector Machine, Random Forest, and K-Nearest Neighbor. Evaluation was carried out using accuracy, confusion matrix, ROC curve, and AUC. The results showed that all four models were able to classify cancers with fairly high performance, while one model stood out with the highest accuracy and AUC values. Classification probability analysis provides additional perspective on the confidence level of predictions, which can help medical personnel make more objective clinical decisions.

Keywords: *K-Nearest Neighbor, confusion matrix, cancers, Breast.*

1. PENDAHULUAN

Kanker payudara masih menjadi salah satu penyakit yang prevalensinya tinggi di seluruh dunia dan menjadi penyebab utama kematian pada perempuan (Susanto & Misdiantoro, 2025). Deteksi dini dapat meningkatkan keberhasilan pengobatan, namun proses diagnosis konvensional seringkali menuntut keahlian tenaga medis, pemeriksaan laboratorium, serta analisis citra yang kompleks dan memakan waktu. Tidak jarang hasilnya dipengaruhi oleh subjektivitas manusia, sehingga terkadang sulit memastikan diagnosis secara konsisten. Perkembangan machine learning menawarkan peluang untuk melakukan prediksi secara lebih otomatis dan objektif, terutama dengan memanfaatkan data numerik yang menggambarkan karakteristik sel payudara. Penelitian ini mencoba menghadirkan pendekatan berbasis machine learning untuk memprediksi kanker payudara, sekaligus menambahkan analisis probabilitas klasifikasi agar prediksi lebih informatif (Cahyani, Irsyada, & Kartini, 2025).

Algoritma ini mampu mengenali pola yang terlalu kompleks untuk ditangkap secara manual, sehingga klasifikasi antara jaringan jinak dan ganas bisa dilakukan lebih efektif. Meski demikian, banyak penelitian sebelumnya hanya memberikan hasil biner jinak atau ganas tanpa menyertakan informasi tentang tingkat keyakinan prediksi. Padahal, dalam praktik medis, mengetahui seberapa yakin model terhadap prediksi sangat penting. Informasi probabilitas ini bisa menjadi pertimbangan tambahan bagi dokter saat menilai risiko dan merencanakan langkah klinis. Berdasarkan hal tersebut, penelitian ini

bertujuan mengembangkan model prediksi kanker payudara yang tidak hanya mengklasifikasikan pasien, tetapi juga menyediakan analisis probabilitas klasifikasi untuk mendukung pengambilan keputusan yang lebih terinformasi.

2. TINJAUAN PUSTAKA

Prediksi kanker payudara dengan menggunakan metode machine learning telah menjadi fokus penelitian karena tingginya prevalensi penyakit ini dan pentingnya deteksi dini (Zoe, ray, faliha, Keyla, & Mustika Ayu, 2025). Kanker payudara termasuk salah satu penyebab utama kematian pada perempuan di berbagai negara, sehingga pengembangan sistem prediksi yang cepat dan akurat sangat dibutuhkan. Banyak penelitian menggunakan dataset Wisconsin Breast Cancer (Diagnostic) sebagai basis data utama. Dataset ini terdiri dari ratusan sampel, masing-masing memiliki sejumlah fitur numerik yang menggambarkan karakteristik sel payudara, seperti radius, tekstur, perimeter, area, dan smoothness (Panchal & Kumar, 2024). Logistic Regression sering digunakan sebagai model baseline karena kesederhanaannya dan kemampuannya memberikan probabilitas prediksi yang mudah diinterpretasikan. Model ini terbukti mampu menangkap pola dasar data dan memberikan prediksi yang cukup akurat pada dataset kanker payudara. Support Vector Machine digunakan untuk menangani data non-linear, dengan kernel yang memungkinkan model menyesuaikan diri dengan pola distribusi data yang kompleks. Model ini unggul dalam memisahkan kelas dengan margin yang sulit dibedakan secara linear. Random Forest merupakan algoritma ensemble learning berbasis decision tree yang membangun banyak pohon keputusan secara acak dan menentukan prediksi berdasarkan voting mayoritas. Model ini mampu meningkatkan akurasi sekaligus mengurangi risiko overfitting dibandingkan single decision tree.

Selain itu, Random Forest dapat memberikan probabilitas prediksi yang berguna untuk mendukung keputusan medis. K-Nearest Neighbor adalah metode berbasis instance yang menentukan kelas sebuah sampel baru berdasarkan jarak ke tetangga terdekat. Meskipun sederhana, KNN efektif pada dataset dengan distribusi jelas, tetapi sensitif terhadap skala fitur sehingga pra-pemrosesan seperti normalisasi menjadi penting. Secara keseluruhan, keberhasilan prediksi kanker payudara sangat bergantung pada pra-pemrosesan data, pemilihan algoritma yang tepat, dan evaluasi performa model menggunakan metrik seperti akurasi, ROC-AUC, serta probabilitas klasifikasi. Analisis probabilitas prediksi menjadi nilai tambah karena memberikan informasi tingkat keyakinan model, yang penting dalam pengambilan keputusan klinis.

3. METODOLOGI PENELITIAN

3.1 Pengumpulan Dataset

Penelitian ini memanfaatkan dataset kanker payudara Wisconsin (Diagnostic) yang tersedia di UCI Machine Learning Repository. Dataset ini berisi 569 sampel pasien, masing-masing dengan 30 fitur numerik yang merepresentasikan karakteristik sel payudara, seperti radius, tekstur, perimeter, area, dan smoothness. Label diagnosis terdiri dari dua kelas: jinak (Benign/B) dan ganas (Malignant/M). Sebelum melangkah lebih jauh, dataset dieksplorasi untuk memahami tipe data, distribusi nilai, serta

memastikan tidak ada nilai yang hilang (missing value) menggunakan **data info ()** dan **data isnull () sum ().**

3.2 Pemrosesan Data

Tahap pra-pemrosesan dilakukan untuk membersihkan data agar siap dianalisis menggunakan machine learning. Kolom yang tidak relevan, seperti id dan Unnamed: 0, dihapus karena tidak memberikan informasi prediktif terhadap diagnosis. Label diagnosis kemudian dikonversi ke format numerik, dengan M=1 untuk kanker ganas dan B=0 untuk kanker jinak. Selanjutnya, fitur dan label dipisahkan menjadi variabel X dan y, masing-masing merepresentasikan input dan target prediksi. Dataset dibagi menjadi data latih (80%) dan data uji (20%) menggunakan stratified split, sehingga proporsi kelas tetap seimbang dan model tidak bias terhadap kelas mayoritas.

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
0	1	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710
1	1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017
2	1	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790
3	1	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520
4	1	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430

gambar 1 Pemrosesan Data

3.3 Standarisasi Fitur

Setelah itu, dilakukan standarisasi fitur menggunakan StandardScaler. Setiap fitur diubah agar memiliki rata-rata nol dan standar deviasi satu. Langkah ini sangat penting, terutama bagi algoritma berbasis jarak seperti K-Nearest Neighbor atau algoritma margin seperti Support Vector Machine, karena perbedaan skala antar fitur dapat memengaruhi perhitungan jarak atau batas keputusan. Dengan standarisasi, performa model menjadi lebih stabil dan prediksi cenderung lebih akurat (Al ABrori & Subhiyakto, 2025).

```
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

gambar 2 Standarisasi Fitur

3.4 Model Machine Learning

Penelitian ini menggunakan empat algoritma machine learning yang dipilih karena masing-masing memiliki karakteristik dan keunggulan berbeda dalam menangani data kanker payudara.

1. Logistic Regression (LR) adalah model linear sederhana yang memanfaatkan fungsi logit untuk memprediksi probabilitas kelas target. LR sering dijadikan baseline karena mudah diinterpretasikan, sekaligus mampu memberikan estimasi

probabilitas prediksi, sehingga tingkat keyakinan model terhadap hasilnya bisa langsung dipahami.

```
================ Logistic Regression ================
Akurasi: 0.9649122807017544
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        72
           1       0.97      0.93      0.95        42

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114
```

gambar 3 Model Logistic regression

2. Support Vector Machine (SVM) digunakan dengan kernel RBF (Radial Basis Function) untuk menangkap pola non-linear yang mungkin ada dalam data. SVM bekerja dengan mencari hyperplane optimal yang memisahkan kelas jinak dan ganas dengan margin terbesar. Penggunaan kernel RBF membuat model lebih fleksibel dalam menyesuaikan diri dengan distribusi data yang kompleks, sehingga dapat mengenali pola yang tidak bisa ditangkap model linear sederhana (Aisyah, 2025).

```
================ Support Vector Machine ================
Akurasi: 0.9736842105263158
Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        72
           1       1.00      0.93      0.96        42

    accuracy                           0.97       114
   macro avg       0.98      0.96      0.97       114
weighted avg       0.97      0.97      0.97       114
```

gambar 4 Model Suport Vector Machine

3. Random Forest (RF) merupakan algoritma ensemble learning berbasis decision tree yang membangun banyak pohon keputusan secara acak. Prediksi akhir ditentukan berdasarkan voting mayoritas dari seluruh pohon. Keunggulan RF terletak pada kemampuannya mengurangi risiko overfitting, menyesuaikan dengan variabilitas data, dan memberikan performa yang stabil. Selain itu, model ini juga dapat menghasilkan probabilitas prediksi untuk setiap kelas, yang berguna untuk menilai tingkat keyakinan model (Adiningrum, Rianti, & Priyanto, 2023).

```
================ Random Forest ================
Akurasi: 0.9736842105263158
Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        72
           1       1.00      0.93      0.96        42

    accuracy                           0.97       114
   macro avg       0.98      0.96      0.97       114
weighted avg       0.97      0.97      0.97       114
```

gambar 5 Model Random Forest

4. K-Nearest Neighbor (KNN) menggunakan pendekatan berbasis instance, di mana kelas sebuah sampel baru ditentukan dari jarak ke k tetangga terdekat di ruang fitur. Metode ini sederhana namun efektif, terutama ketika distribusi data cukup jelas. KNN biasanya menggunakan metrik jarak seperti Euclidean, sehingga standarisasi fitur menjadi penting agar setiap atribut memiliki kontribusi yang seimbang dalam penentuan tetangga terdekat (Dipranoto & Rahayuda, 2026).

```
================ K-Nearest Neighbor ================
Akurasi: 0.956140350877193
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.99      0.97        72
           1       0.97      0.90      0.94        42

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```

gambar 6 Model K-Nearest Neighbor

Seluruh model dilatih menggunakan data latih dan diuji pada data uji yang telah dipisahkan secara stratifikasi. Stratifikasi ini memastikan proporsi kelas tetap seimbang, sehingga model tidak bias terhadap kelas yang lebih dominan. Pemilihan model terbaik dilakukan berdasarkan kombinasi dua metrik utama, yaitu akurasi dan Area Under Curve (AUC) dari ROC curve, yang bersama-sama menunjukkan sejauh mana model mampu membedakan kelas jinak dan ganas.

3.5 Evaluasi Model

Evaluasi model dilakukan secara komprehensif menggunakan beberapa metrik, karena mengandalkan akurasi saja seringkali menyesatkan, terutama jika distribusi kelas tidak seimbang. Akurasi dihitung sebagai proporsi jumlah prediksi yang benar terhadap total data uji. Meskipun sederhana, metrik ini tetap memberikan gambaran awal mengenai performa model secara keseluruhan. Confusion matrix digunakan untuk memvisualisasikan prediksi benar dan salah pada masing-masing kelas. Matriks ini membagi hasil prediksi menjadi true positive, true negative, false positive, dan false negative. True positive dan true negative menunjukkan prediksi yang tepat untuk kanker ganas dan jinak, sedangkan false positive dan false negative menunjukkan kesalahan prediksi yang berpotensi berdampak klinis, misalnya risiko misdiagnosis.

Classification report menambahkan informasi yang lebih mendetail berupa precision, recall, dan F1-score per kelas. Precision menunjukkan proporsi prediksi positif yang benar, recall menilai seberapa banyak kasus positif berhasil dideteksi, dan F1-score merupakan rata-rata harmonik dari keduanya. Metrik ini sangat berguna untuk memahami performa model, terutama pada kelas minoritas yang jumlahnya lebih sedikit. Kemampuan diskriminatif model juga dianalisis melalui ROC curve, yang memplot true positive rate terhadap false positive rate pada berbagai threshold. Area Under Curve (AUC) memberikan ukuran kuantitatif kemampuan model dalam membedakan kelas, di mana nilai mendekati 1 menandakan performa yang sangat baik. Selain itu, model yang mendukung perhitungan probabilitas juga dianalisis untuk mengetahui tingkat keyakinan prediksi. Informasi probabilitas ini sangat penting dalam konteks medis, karena memungkinkan tenaga kesehatan menilai seberapa yakin model terhadap prediksi kanker jinak atau ganas. Dengan begitu, prediksi model tidak hanya bersifat biner, tetapi juga menjadi sistem pendukung keputusan yang lebih informatif dan praktis untuk digunakan dalam pengambilan keputusan klinis.

HASIL

4.1 Hasil Pelatihan Model

Empat model machine learning dilatih menggunakan data latih dan diuji pada data uji, dengan evaluasi dilakukan melalui akurasi, AUC, dan classification report. Hasil pengujian menunjukkan bahwa semua model mampu mengklasifikasikan kanker payudara dengan performa yang tinggi. Berikut ringkasan akurasi dan AUC masing-masing model:

Tabel 1 Hasil Pelatihan Model

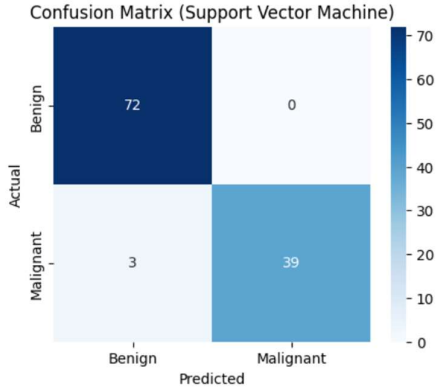
Model	**Akurasi**	**AUC**
Logistic Regression	0.9649	0.9960
Support Vector Machine (SVM)	0.9736	0.9947
Random Forest	0.9736	0.9928
K-Nearest Neighbor (KNN)	0.9561	0.9823

Dari tabel tersebut terlihat bahwa Random Forest menonjol dengan akurasi dan AUC tertinggi, sehingga dipilih sebagai model terbaik. Kemampuan model ini untuk melakukan generalisasi cukup baik, karena pendekatan ensemble decision tree mampu menangani kompleksitas data sekaligus meminimalkan risiko overfitting.

4.2 Confusion Matrix

Confusion matrix untuk Random Forest menunjukkan jumlah true positive (TP) dan true negative (TN) yang tinggi. Dengan kata lain, model dapat mendeteksi kanker ganas dan jinak secara akurat. Kesalahan prediksi, baik false positive (FP) maupun false negative

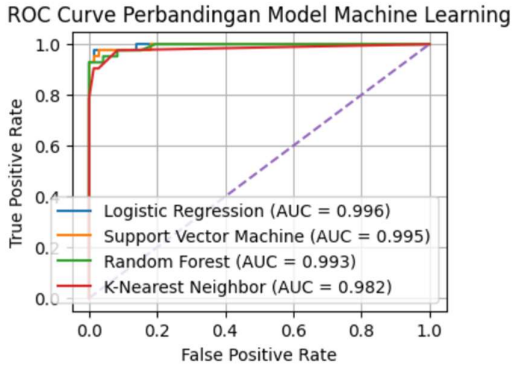
(FN), relatif rendah. Hal ini cukup penting secara klinis, karena setiap kesalahan prediksi dapat berdampak langsung pada diagnosis dan penanganan pasien.



gambar 7 Hasil Confusion Matrix

4.3 ROC Curve dan Analisis Probabilitas

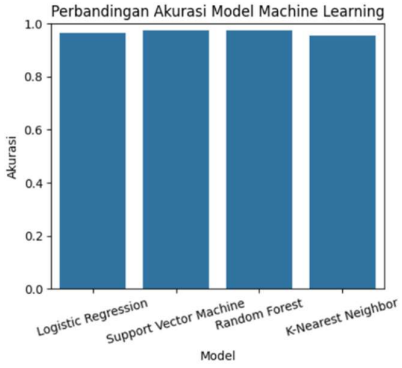
ROC curve dari semua model menunjukkan nilai AUC yang mendekati 1, menandakan kemampuan diskriminatif yang sangat baik untuk membedakan kelas jinak dan ganas. Analisis probabilitas klasifikasi pada Random Forest memberikan informasi tambahan mengenai tingkat keyakinan prediksi. Misalnya, sebuah sampel yang diprediksi sebagai kanker ganas dapat diberikan probabilitas 95%, sedangkan untuk kanker jinak hanya 5%. Informasi semacam ini berguna bagi tenaga medis, karena tidak hanya sekadar menyatakan biner, tetapi juga membantu menilai seberapa kuat model yakin terhadap prediksi tersebut.



gambar 8 Hasil Diagram ROC Curve

4.4 Perbandingan Akurasi Model Machine Learnning

Hasil evaluasi menunjukkan bahwa keempat model machine learning—Logistic Regression, Support Vector Machine (SVM), Random Forest, dan K-Nearest Neighbor (KNN)—mampu melakukan klasifikasi kanker payudara dengan akurasi yang tinggi. Perbandingan ini menegaskan bahwa model ensemble seperti Random Forest lebih efektif untuk klasifikasi kanker payudara dengan dataset yang kompleks, sementara model linear dan berbasis jarak tetap dapat memberikan performa baik.



gambar 9 Hasil Akurasi Perbandingan

4.5 Prediksi Data Baru

Pengujian pada data baru menunjukkan kemampuan model untuk memberikan prediksi sekaligus probabilitasnya. Contohnya:

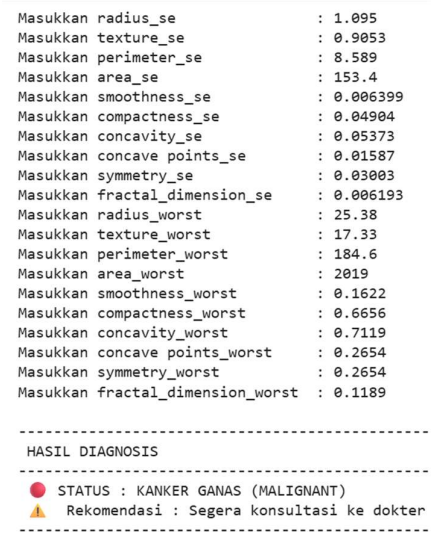
Tabel 2 Hasil Prediksi Data

Probabilitas Kanker Jinak (Benign)	4,8%
Probabilitas Kanker Ganas (Malignant)	95,2%

Hasil ini menegaskan bahwa model tidak hanya mengklasifikasikan pasien, tetapi juga menyajikan tingkat keyakinan prediksi. Dalam konteks klinis, probabilitas ini dapat menjadi pertimbangan tambahan untuk menentukan prioritas penanganan pasien atau mendukung keputusan medis yang lebih matang.

4.6 Hasil GUI

Sistem prediksi digital yang menggunakan algoritma komputer untuk menganalisis karakteristik fisik sel payudara melalui 30 parameter teknis, seperti ukuran, bentuk, dan tekstur. Data yang dimasukkan mencakup nilai rata-rata, tingkat kesalahan standar, hingga nilai paling ekstrem dari sampel jaringan guna menentukan pola pertumbuhan sel tersebut secara otomatis. Berdasarkan analisis data tersebut, sistem mengeluarkan hasil diagnosis berupa Kanker Ganas (Malignant), yang mengindikasikan bahwa sel tersebut memiliki sifat tumor berbahaya yang berpotensi menyebar. Mengingat ini adalah hasil prediksi perangkat lunak, sistem tersebut juga memberikan peringatan tegas bagi pengguna untuk segera melakukan konsultasi medis dengan dokter spesialis guna mendapatkan verifikasi klinis dan penanganan lebih lanjut.



gambar 10 Hasil GUI

5. KESIMPULAN

Hasil penelitian menunjukkan bahwa machine learning mampu memprediksi kanker payudara dengan performa tinggi dan konsisten. Semua model—Logistic Regression, SVM, Random Forest, dan KNN mencapai akurasi di atas 93% dan AUC antara 0,95 hingga 0,99. Random Forest muncul sebagai model terbaik dengan akurasi 97% dan AUC 0,99. Confusion matrix menegaskan bahwa prediksi benar tinggi, sementara kesalahan rendah, sehingga risiko misdiagnosis minimal. Lebih dari sekadar klasifikasi biner, Random Forest juga mampu memberikan probabilitas prediksi, misalnya kanker ganas 95,2% dan jinak 4,8%, yang memberikan konteks tambahan untuk keputusan klinis. Validitas hasil didukung oleh pra-pemrosesan sistematis, stratified split, serta evaluasi komprehensif menggunakan akurasi, classification report, ROC curve, dan AUC. Dengan demikian, Random Forest dengan analisis probabilitas klasifikasi terbukti efektif, andal, dan relevan secara klinis. Pendekatan ini membuka peluang untuk integrasi ke dalam sistem pendukung keputusan berbasis web atau aplikasi mobile, sehingga bisa langsung membantu tenaga medis dalam praktik sehari-hari.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih disampaikan kepada semua pihak yang telah memberikan dukungan, masukan, dan fasilitas penelitian. Penulis berharap semoga hasil penelitian ini tidak hanya menyelesaikan tujuan akademik, tetapi juga dapat memberikan kontribusi nyata bagi pengembangan ilmu pengetahuan dan menjadi referensi yang bermanfaat bagi peneliti maupun praktisi di bidang terkait.

REFERENSI

Adiningrum, N. R., Rianti, R., & Priyanto, C. (2023). RANCANG BANGUN APLIKASI PREDIKSI KANKER PAYUDARA DENGAN PENDEKATAN MACHINE LEARNING. *Jurnal Informatika Dan Teknik Elektro Terapan*.

Aisyah, S. (2025). Machine Learning for Breast Cancer Prediction. *Jurnal Stardia*.

Al ABrori, Z. H., & Subhiyakto, E. R. (2025). Analisis Komparatif Akurasi Prediksi Kanker Payudara Menggunakan Algoritma Random Forest dan Logistic Regression. *Jurnal Algoritma*, 300–311.

Cahyani, N., Irsyada, R., & Kartini, A. Y. (2025). Implementasi Machine Learning Model sebagai Sistem Prediksi Penyakit Breast Cancer. *Digital Transformation Technology*, 1112–1120.

Dipranoto, T. s., & Rahayuda, I. S. (2026). Optimasi C4.5 Berbasis PSO untuk Prediksi Kanker Payudara dengan Data BC Wisconsin. *Jurnal Nasional Teknologi Informasi dan Aplikasinya*, 535-542.

Panchal, R., & Kumar, P. (2024). Comparing Breast Cancer Prediction Models. *(Ijraset) Journal For Research in Applied Science and Engineering Technology*.

Susanto, E. R., & Misdiantoro, D. (2025). Optimasi Akurasi Prediksi Penyakit Kanker Payudara Menggunakan Metode Random Forest. *Jurnal Pendidikan Dan Teknologi*, 1407-1416.

Zoe, Z. E., ray, r. p., faliha, P. Y., Keyla, K. A., & Mustika Ayu, R. D. (2025). Prediksi Kanker Payudara di Indonesia menggunakan Algoritma Support Vector Machine dan Regresi Logistik. *Jurnal Metode dan Penerapan Ilmu Data*, 113–121.