Journal of Artificial Intelligence and Software Engineering
Vol. 5, No. 3, September 2025, pp. 1285~1295
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v5i3.7028

Comparative Analysis of Random Forest Algorithms, Artificial Neural Networks, and Logistic Regression in Breast Cancer Prediction with Machine Learning Approach

Rahmadi M.Ali[1], Nurdin[2], Al Khaidar[3*], Maghriza Azzanna[4] , Athirah Rusadi[5]
[1,2,3,4,5] Program Studi Magister Teknologi Informasi, Fakultas Teknik, Universitas Malikussaleh, Kota Lhokseumawe, 24355, Indonesia

Informasi Artikel

Diterima : 29 Mei 2025
Revisi : 9 Juni 2025
Publikasi : 30 September 2025

Kata Kunci:

Random Forest
Artificial Neural Networks
Logistic Regression
Breast Cancer Prediction
Machine Learning

ABSTRAK

Perkembangan teknologi informasi khususnya kecerdasan buatan dan machine learning, telah meningkatkan efektivitas deteksi dini penyakit seperti kanker payudara. Namun, tingginya angka kejadian dan kematian akibat kanker payudara di Indonesia masih menjadi tantangan besar, terutama karena rendahnya tingkat deteksi dini dan banyak pasien datang dalam stadium lanjut. Penelitian ini membandingkan performa tiga algoritma machine learning, yaitu Random Forest, Artificial Neural Network (ANN), dan Logistic Regression, dalam memprediksi diagnosis kanker payudara berdasarkan akurasi, efisiensi komputasi, dan kestabilan kinerja. Evaluasi dilakukan dengan classification report dan validasi silang 10-Fold Cross Validation. Data yang digunakan data medis dengan 10 parameter age ,meno, size, grade, nodes, pgr, er,hormon, rfstime dan status, data yang digunakan sebanyak 2510 data pasien. Hasil menunjukkan Logistic Regression memiliki akurasi rata-rata tertinggi sebesar 77,56% dan waktu eksekusi tercepat, yaitu 0,024897 detik, menandakan efisiensi dan kestabilan yang baik. Random Forest memberikan akurasi classification report 80% dan nilai AUC tertinggi 0,89, menunjukkan keunggulan dalam diskriminasi kelas. ANN memiliki performa terendah dengan akurasi validasi silang 74,64% dan recall rendah untuk kelas positif.

ABSTRACT

The development of information technology, especially artificial intelligence and machine learning, has increased the effectiveness of early detection of diseases such as breast cancer. However, the high incidence and mortality of breast cancer in Indonesia is still a major challenge, especially due to the low rate of early detection and many patients who come to the stadium further. This study compares the performance of three machine learning algorithms, namely Random Forest, Artificial Neural Network (ANN), and Logistic Regression, in predicting breast cancer diagnosis based on accuracy, computational efficiency, and performance stability. The evaluation was carried out with a classification report and 10-Fold Cross Validation cross validation. The data used were medical data with 10 parameters age, meno, size, grade, node, pgr, er, hormone, rfstime and status, the data used was 2510 patient data. The results showed that Logistic Regression had the highest average accuracy of 77.56% and the fastest execution time, which was 0.024897 seconds, indicating good efficiency and stability. Random Forest provided a classification report accuracy of 80% and the highest AUC value of 0.89, indicating superiority in class discrimination. ANN has the lowest performance with a cross-validation accuracy of 74.64% and low recall to the positive class.

Penulis Koresponden

Email: alkhaidarkutablang@gmail.com

Cara sitasi IEEE::

GR. M. Ali. A. Khaidar, M. Azanna & A. Rusadi, "Comparative Analysis of Random Forest Algorithms, Artificial Neural Networks, and Logistic Regression in Breast Cancer Prediction with Machine Learning Approach," *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 5, no. 3, pp. 1285-1295, September 2025, doi: 10.30811/jaise.v5i3.7028

1. PENDAHULUAN

Perkembangan teknologi informasi dan komputer telah mengalami kemajuan yang sangat pesat dalam beberapa dekade terakhir [1] [2]. Teknologi tidak lagi sekadar menjadi alat bantu, tetapi telah menjadi elemen integral dalam berbagai bidang kehidupan, termasuk sektor kesehatan. Salah satu cabang dari teknologi yang berkembang pesat adalah kecerdasan buatan *Artificial Intelligence*, khususnya machine learning, yang memungkinkan komputer untuk belajar dari data dan membuat prediksi atau keputusan secara otomatis tanpa pemrograman eksplisit [3] [4]. Perkembangan teknologi informasi dan komputasi telah membuka peluang baru dalam bidang kesehatan, khususnya dalam upaya deteksi dini penyakit.

Prediksi adalah proses memperkirakan suatu kejadian atau hasil yang akan terjadi di masa depan berdasarkan data, pola, atau pengalaman sebelumnya. Prediksi merupakan hasil dari analisis logis atau perhitungan matematis yang didasarkan pada data historis, teori, atau model tertentu untuk memperkirakan kondisi atau peristiwa yang akan datang [5] [6]. Prediksi adalah proses estimasi nilai atau kejadian di masa depan menggunakan algoritma dan metode statistik atau pembelajaran mesin (machine learning) berdasarkan data yang ada [7] [8]. Machine learning sebagai bagian dari kecerdasan buatan menawarkan pendekatan yang potensial dalam menganalisis data medis untuk prediksi dan diagnosis penyakit, termasuk kanker payudara.

Kanker payudara adalah neoplasma ganas yang berasal dari parenkim payudara, terutama berkembang di saluran susu atau lobulus sebagai penghasil air susu [9] [10] [11]. Kanker ini ditandai oleh pertumbuhan sel-sel abnormal yang cepat dan tidak terkendali sehingga membentuk tumor yang dapat teraba atau terdeteksi melalui mamografi [12] [13]. Meskipun lebih sering menyerang wanita, kanker payudara juga dapat terjadi pada pria, meskipun kasusnya sangat jarang. Berdasarkan data terbaru, jumlah pengidap kanker payudara di Indonesia mencapai 66.000 kasus, dengan lebih dari 48 persen di antaranya telah berada pada stadium lanjut [14]. Prevalensi kanker payudara di Indonesia tercatat sebesar 18 per 100.000 wanita, dengan jumlah kasus mencapai 61.682, dan menempati urutan kedua setelah kanker serviks [15]. Setiap tahunnya, diperkirakan terdapat sekitar 400.000 kasus baru kanker yang terdeteksi di Indonesia, dengan angka kematian mencapai 240.000 jiwa. Pada tahun 2022, jumlah kasus baru meningkat menjadi lebih dari 408.661, disertai dengan hampir 242.099 kematian akibat kanker. Kanker payudara menjadi jenis kanker paling umum di Indonesia, dengan 66.271 kasus baru tercatat pada tahun 2024 . Tanpa adanya intervensi yang efektif, jumlah kasus kanker secara keseluruhan diperkirakan akan meningkat lebih dari 70 persen pada tahun 2050.

Permasalahan yang terjadi tingginya angka kejadian dan kematian akibat kanker payudara di Indonesia diperparah oleh rendahnya kesadaran masyarakat untuk melakukan deteksi dini. Sebagian besar pasien datang dalam kondisi stadium lanjut, sehingga peluang keberhasilan pengobatan menurun dan biaya perawatan meningkat secara signifikan. Hambatan psikologis dan kurangnya kemauan untuk melakukan skrining menjadi salah satu faktor utama sulitnya menekan angka kasus kanker payudara [16]. Menghadapi tantangan tersebut teknologi informasi dan pendekatan machine learning mulai diadopsi untuk meningkatkan akurasi dan efisiensi deteksi dini kanker payudara. Berbagai algoritma machine learning seperti Random Forest, Jaringan Syaraf Tiruan, dan Regresi Logistik telah digunakan untuk memprediksi risiko dan diagnosis kanker payudara. Studi-studi sebelumnya menunjukkan bahwa algoritma-algoritma ini mampu memberikan tingkat akurasi yang tinggi dalam klasifikasi dan prediksi kanker payudara, bahkan melebihi kemampuan radiolog dalam beberapa kasus.

Random Forest adalah metode pembelajaran mesin berbasis ensemble yang membangun sejumlah pohon keputusan secara acak dengan menggunakan subset data berbeda melalui teknik bootstrap, kemudian hasilnya digabungkan melalui voting untuk klasifikasi atau rata-rata untuk regresi, sehingga meningkatkan akurasi dan mengurangi risiko overfitting [17] [18]. Jaringan Syaraf Tiruan adalah sistem komputasi yang meniru cara kerja otak dengan lapisan neuron buatan yang saling terhubung dan mampu mempelajari pola kompleks melalui penyesuaian bobot dan fungsi aktivasi, banyak digunakan untuk klasifikasi dan prediksi data besar [19] [20]. Regresi Logistik adalah metode statistik yang menghubungkan variabel dependen biner dengan

variabel independen menggunakan fungsi logit untuk menghitung probabilitas kejadian, sering dipakai dalam klasifikasi seperti diagnosis penyakit [21] .

Penelitian ini bertujuan untuk mengembangkan model prediksi kanker payudara berbasis machine learning guna mendukung deteksi dini yang lebih akurat dan terjangkau, khususnya di Indonesia. Pemilihan tiga algoritma, yaitu Random Forest, Jaringan Syaraf Tiruan, dan Regresi Logistik, didasarkan pada keunggulan masing-masing dalam pengolahan data medis. Random Forest dipilih karena kemampuannya dalam menjaga stabilitas model serta mengurangi risiko overfitting. Jaringan Syaraf Tiruan diprioritaskan untuk kemampuan dalam mengenali pola-pola kompleks pada data, sedangkan Regresi Logistik dipilih sebagai model yang sederhana dan mudah diinterpretasikan. Berbeda dengan penelitian-penelitian sebelumnya, studi ini melakukan perbandingan komprehensif terhadap ketiga algoritma tersebut serta melakukan optimasi model khususnya pada data populasi Indonesia. Selain itu, penelitian ini juga mempertimbangkan aspek implementasi praktis seperti efisiensi biaya dan kebutuhan komputasi, sehingga hasil yang diperoleh tidak hanya memiliki akurasi yang tinggi, tetapi juga siap diadopsi dalam sistem pelayanan kesehatan yang nyata dan berkelanjutan.

2. METODE

Pada bab ini akan dijelaskan metode-metode yang digunakan dalam penelitian untuk memperoleh hasil prediksi yang akurat dalam analisis data kanker payudara. Metode yang digunakan meliputi Random Forest, Jaringan Syaraf Tiruan, dan Logistic Regression, yang masing-masing memiliki keunggulan dalam klasifikasi dan prediksi. Selain itu, bab ini juga memaparkan alur penelitian serta perancangan sistem sebagai dasar implementasi metode-metode tersebut. Penjelasan ini bertujuan memberikan gambaran yang jelas mengenai pendekatan teknis yang digunakan dalam penelitian.

2.1 Metode Random Forest

Breiman memperkenalkan metode Random Forest pada tahun 2001 sebagai teknik yang dapat digunakan untuk klasifikasi maupun prediksi suatu permasalahan. Metode ini didasarkan pada penggunaan sejumlah pohon keputusan secara bersamaan. Secara sederhana, Random Forest merupakan gabungan dari banyak pohon keputusan yang bekerja dengan cara menerima data input dari akar di bagian atas dan mengolahnya hingga mencapai daun di bagian bawah [22]. Untuk klasifikasi, hasil dari metode ini berupa struktur pohon-pohon yang terbentuk, sementara untuk prediksi, hasil diperoleh dengan mengambil rata-rata dari output seluruh pohon keputusan yang ada [23].

Metode Random Forest merupakan pengembangan dari metode Classification and Regression Tree (CART) yang menggunakan teknik agregasi bagging atau bootstrap serta pemilihan fitur secara acak. Bagging sendiri adalah salah satu teknik yang dapat meningkatkan performa algoritma klasifikasi dengan dasar konsep ensemble. Berdasarkan algoritma Random Forest dijalankan melalui beberapa tahapan sebagai berikut.

1. Pengambilan Sampel Bootstrap
   Ambil n sampel dari dataset asli dengan pengembalian (resampling bootstrap).
2. Pembuatan Pohon Klasifikasi
   Bangun pohon keputusan pada setiap sampel bootstrap dengan memilih secara acak m variabel prediktor sebagai kandidat pemisah, dimana nilai m dapat dihitung sebagai:
   $$m = \sqrt{m} \text{ atau } m = \log_2(M) + 1 \quad (1)$$
   dengan M adalah jumlah total variabel prediktor.
3. Pengulangan Pembuatan Pohon
   Ulangi proses pembuatan pohon hingga terbentuk k pohon klasifikasi.
4. Prediksi Tiap Pohon
   Lakukan prediksi klasifikasi menggunakan setiap pohon yang terbentuk.
5. Agregasi Hasil Prediksi
   Tentukan hasil akhir klasifikasi berdasarkan voting mayoritas dari seluruh pohon:
   $$\hat{y} = mode\{h_1(x), h_2(x), ..., h_k(x)\} \quad (2)$$
   dimana $h_i(x)$ adalah prediksi pohon ke-i.

2.2 Metode Jaringan Saraf Tiruan

Jaringan Syaraf Tiruan merupakan salah satu metode pembelajaran mesin yang meniru cara kerja otak manusia dalam memproses informasi. JST terdiri dari sejumlah unit pemrosesan sederhana yang disebut neuron atau node, yang tersusun dalam lapisan-lapisan (input, tersembunyi, dan output) dan saling terhubung melalui bobot (weights) [24]. Proses belajar pada JST dilakukan dengan menyesuaikan bobot koneksi antar neuron berdasarkan data latih menggunakan algoritma pembelajaran tertentu, seperti algoritma backpropagation. Model ini mampu mengenali pola yang kompleks dan non-linear sehingga banyak digunakan dalam klasifikasi, prediksi, dan pengenalan pola [25]. Berikut rumus metode jaringan saraf tiruan.

1. Output neuron (linear kombinasi input dan bobot)

$$z = \sum_{i=0}^{n} \binom{n}{k} wi\ xi^{k} + b \quad (3)$$

Dimana:

wi = bobot pada koneksi ke neuron,
xi = input ke neuron,
b = bias,
n = jumlah input.

2. Fungsi aktivasi

$$a\ \frac{1}{1+e^{-z}} \quad (4)$$

Dimana:

a = output neuron setelah fungsi aktivasi,
e = basis logaritma natural.

2.3 Metode Logistic Regression

Logistic Regression merupakan salah satu metode statistik yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen dengan variabel dependen yang bersifat biner (dikotomi), seperti ya/tidak atau 1/0 [26]. Berbeda dengan regresi linear yang menghasilkan output berupa nilai kontinu, Logistic Regression menghasilkan output dalam bentuk probabilitas [27]. Model ini menggunakan fungsi logistik (sigmoid) untuk memetakan nilai input ke rentang antara 0 dan 1, yang selanjutnya dapat ditafsirkan sebagai probabilitas suatu kejadian. Logistic Regression banyak diterapkan dalam bidang kesehatan, keuangan, dan klasifikasi risiko.

1. Fungsi logit (model linier):

$$z = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta nxn \quad (5)$$

2. Fungsi sigmoid (fungsi logistik):

$$P\ (\gamma = 1\ x) = \frac{1}{1+e^{-z}} \quad (6)$$

Dimana:

z = kombinasi linear dari variabel input,
β = koefisien regresi (parameter model),
x = variabel input,
$P\ (\gamma = 1\ x)$ = probabilitas hasil output bernilai 1

2.4 Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari *Kaggle*, yang merupakan salah satu platform berbagi dataset populer untuk keperluan analisis dan pembelajaran mesin (*machine learning*). Dataset ini berisi informasi medis pasien yang digunakan untuk mendeteksi risiko kanker payudara. Adapun parameter yang digunakan dapat dilihat pada Tabel 1 berikut.

Tabel 1. Parameter Prediksi Kanker Payudara

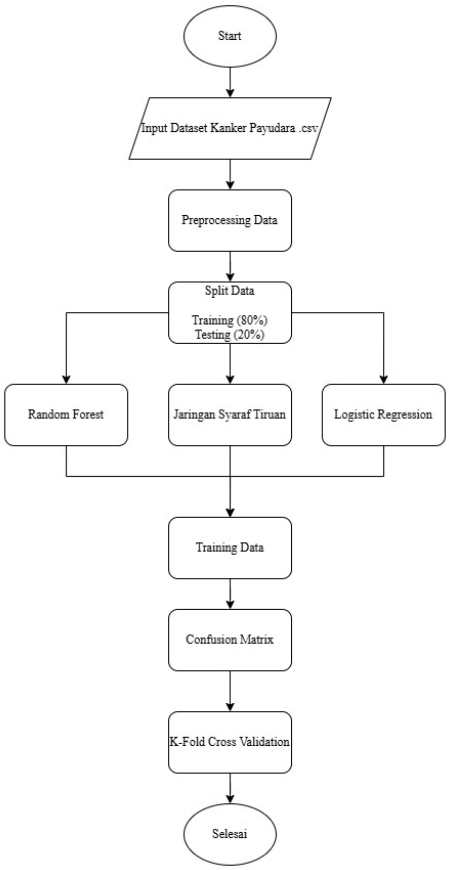
No	Nama Atribut	Deskripsi
1	age	Usia pasien
2	meno	Status menopause
3	size	Ukuran tumor
4	grade	Tingkat keparahan histologis kanker
5	nodes	Jumlah kelenjar getah bening yang terlibat
6	pgr	Status reseptor progesteron
7	er	Status reseptor estrogen
8	hormon	Status terapi hormon
9	rfstime	Waktu bebas penyakit (recurrence-free survival time)
10	Status	Label target, yaitu status pasien apakah terkena kanker payudara kembali (1) atau tidak (0)

Atribut status merupakan label atau kelas yang digunakan dalam proses klasifikasi. Nilai 1 menunjukkan bahwa pasien mengalami kekambuhan kanker payudara, sedangkan nilai 0 menunjukkan bahwa pasien tidak

mengalami kekambuhan. Oleh karena itu, data ini termasuk dalam kategori data klasifikasi biner (*binary classification*), yang digunakan untuk membangun model prediksi menggunakan algoritma pembelajaran mesin. Dataset ini sangat sesuai digunakan untuk tujuan penelitian dalam mengklasifikasikan risiko kekambuhan kanker payudara berdasarkan karakteristik medis pasien.

2.5 Perancangan Sistem

Perancangan sistem merupakan tahapan penting dalam penelitian ini yang bertujuan untuk menggambarkan alur kerja, struktur, dan komponen-komponen utama dari sistem yang akan dibangun. Tahapan ini mencakup identifikasi kebutuhan sistem, pemodelan proses, serta perancangan arsitektur sistem guna memastikan bahwa implementasi metode yang digunakan. Adapun perancangan sistem pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Perancangan Sistem

Berdasarkan Gambar 1 proses dimulai dengan menginput dataset kanker payudara dalam format .csv, yang selanjutnya dilakukan tahap preprocessing data. Tahapan ini mencakup pembersihan data seperti penghapusan atribut yang tidak relevan, penanganan data yang hilang, serta proses normalisasi data. Normalisasi bertujuan untuk menyamakan skala nilai dari setiap fitur numerik agar tidak terjadi dominasi oleh fitur tertentu dalam proses pelatihan model, terutama pada algoritma yang sensitif terhadap perbedaan skala seperti Jaringan Syaraf Tiruan dan Regresi Logistik. Setelah preprocessing selesai, data dibagi menjadi dua bagian, yaitu data latih sebesar 80% dan data uji sebesar 20%. Selanjutnya, ketiga algoritma klasifikasi Random Forest, Jaringan Syaraf Tiruan, dan Regresi Logistik dilatih menggunakan data latih. Untuk mengevaluasi kinerja masing-masing model, dilakukan pengujian menggunakan confusion matrix guna mengukur akurasi, sensitivitas, dan spesifisitas, serta K-Fold Cross Validation sebagai metode validasi untuk mengukur stabilitas dan generalisasi model secara lebih menyeluruh. Evaluasi ini bertujuan untuk menentukan model yang paling optimal dalam memprediksi status kanker payudara berdasarkan data yang tersedia.

3. HASIL DAN PEMBAHASAN

Bab ini menguraikan hasil yang diperoleh dari penerapan metode klasifikasi terhadap data yang terdiri dari berbagai parameter, seperti patient ID (pid), age (usia pasien), menopause status (meno), tumor size (size), tumor grade (grade), number of lymph nodes (nodes), serta status biomarker seperti progesterone receptor (pgr), estrogen receptor (er), dan hormone therapy (hormon). Selain itu, terdapat juga variabel rfstime yang menunjukkan waktu bertahan pasien, serta status yang merupakan label kelas untuk prediksi.Data ini telah melalui tahap preprocessing, termasuk pembersihan data, transformasi, dan pembagian menjadi data latih dan data uji. Selanjutnya, dilakukan penerapan tiga metode klasifikasi yaitu Random Forest, Jaringan Saraf Tiruan (Artificial Neural Network), dan Regresi Logistik. Hasil dari masing-masing metode dianalisis berdasarkan akurasi, precision, recall, dan F1-score, untuk mengetahui kinerja model dalam memprediksi status pasien berdasarkan parameter yang diberikan.Pembahasan dalam bab ini juga mencakup evaluasi terhadap kekuatan masing-masing metode, interpretasi hasil prediksi, serta relevansinya terhadap studi literatur dan kondisi riil dalam bidang medis, khususnya yang berkaitan dengan prediksi keberlangsungan hidup pasien kanker payudara berdasarkan data klinis.

3.1 Evaluasi Model Confusion Matrix

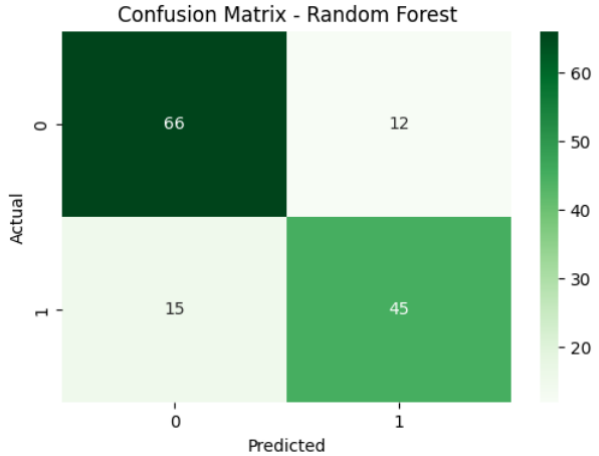
3.1.1 Model Random Forest

Model Random Forest merupakan algoritma ensemble yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi. Evaluasi kinerja model ini dilakukan melalui classification report dan confusion matrix. Adapun hasil evaluasi model random forest dapat dilihat pada Tabel 1.

Tabel 2. Classification Report Model Random Forest

Class	Precision	Recall	F1-Score	Support
0	0.81	0.85	0.83	78
1	0.79	0.75	0.77	60
accuracy			0.80	138
macro avg	0.80	0.80	0.80	138
weighted avg	0.80	0.80	0.80	138

Berdasarkan Tabel 1, hasil evaluasi model Random Forest menunjukkan bahwa algoritma ini memiliki tingkat akurasi sebesar 80% dalam memprediksi status pasien kanker payudara. Untuk kelas 0 (negatif), model memperoleh precision sebesar 0,81, recall sebesar 0,85, dan f1-score sebesar 0,83, sedangkan untuk kelas 1 (positif), precision yang diperoleh adalah 0,79, recall sebesar 0,75, dan f1-score sebesar 0,77. Nilai rata-rata makro (macro average) dan rata-rata tertimbang (weighted average) untuk precision, recall, dan f1-score semuanya sebesar 0,80, yang mengindikasikan bahwa model memiliki performa yang seimbang dalam mengklasifikasikan kedua kelas. Hasil ini menunjukkan bahwa Random Forest merupakan model yang cukup andal dalam memprediksi data diagnosis kanker payudara pada dataset yang digunakan.



Gambar 2. Confusion Matrix Random Forest

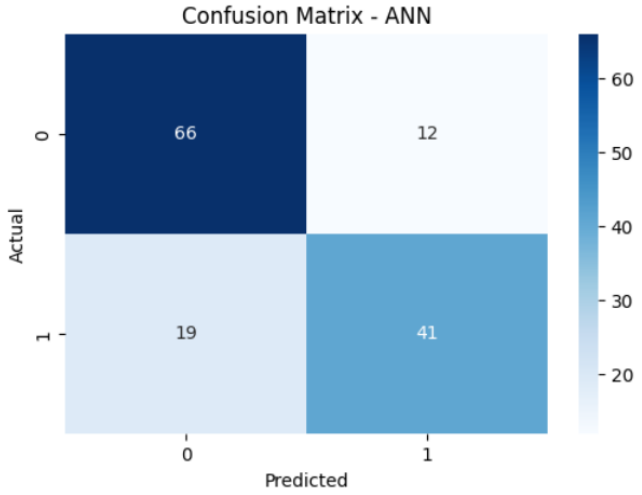
3.1.2 Model Artificial Neural Network

Model Jaringan Saraf Tiruan (Artificial Neural Network) digunakan untuk mengklasifikasikan data dengan memanfaatkan arsitektur lapisan-lapisan neuron. Evaluasi performa model dilakukan melalui classification report dan confusion matrix. Adapun hasil evaluasi model jaringan saraf tiruan dapat dilihat pada Tabel 2.

Tabel 3. Classification Report Model Jaringan Saraf Tiruan

Class	Precision	Recall	F1-Score	Support
0	0.78	0.85	0.81	78
1	0.77	0.68	0.73	60
accuracy			0.78	138
macro avg	0.78	0.76	0.77	138
weighted avg	0.78	0.78	0.77	138

Berdasarkan Tabel 2, model Jaringan Saraf Tiruan (JST) menghasilkan akurasi sebesar 78% dalam mengklasifikasikan status pasien kanker payudara. Untuk kelas 0, model mencatat precision sebesar 0,78, recall 0,85, dan f1-score sebesar 0,81, sedangkan pada kelas 1, precision sebesar 0,77, recall 0,68, dan f1-score 0,73. Nilai rata-rata makro untuk precision, recall, dan f1-score masing-masing sebesar 0,78, 0,76, dan 0,77, sedangkan nilai rata-rata tertimbang menunjukkan hasil yang relatif serupa. Hal ini menunjukkan bahwa meskipun model JST memiliki performa yang baik, terutama dalam mengklasifikasikan kelas negatif, namun masih terdapat kelemahan dalam mengenali kelas positif secara akurat.



Gambar 3. Confusion Matrix Jaringan Saraf Tiruan

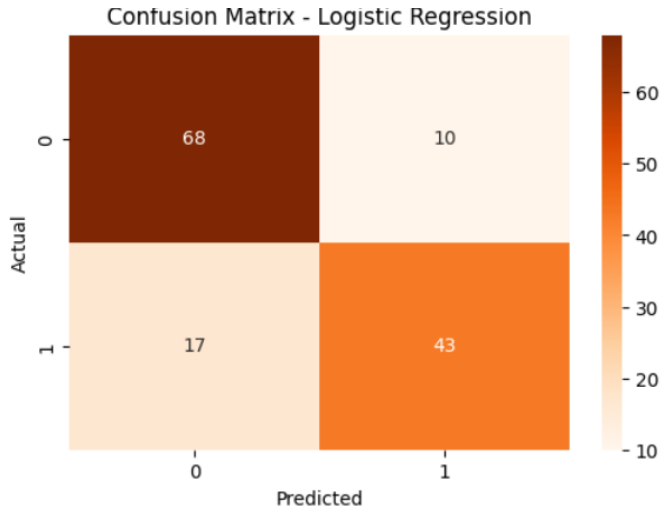
3.1.3 Model Logistic Regression

Model Logistic Regression digunakan sebagai pendekatan linier untuk memprediksi klasifikasi biner pada data kanker payudara. Hasil evaluasi disajikan dalam bentuk classification report dan confusion matrix. Adapun hasil evaluasi model *Logistic Regression* dapat dilihat pada Tabel 3.

Tabel 4. Classification Report Model Logistic Regression

Class	Precision	Recall	F1-Score	Support
0	0.80	0.87	0.83	78
1	0.81	0.72	0.76	60
accuracy			0.80	138
macro avg	0.81	0.79	0.80	138
weighted avg	0.80	0.80	0.80	138

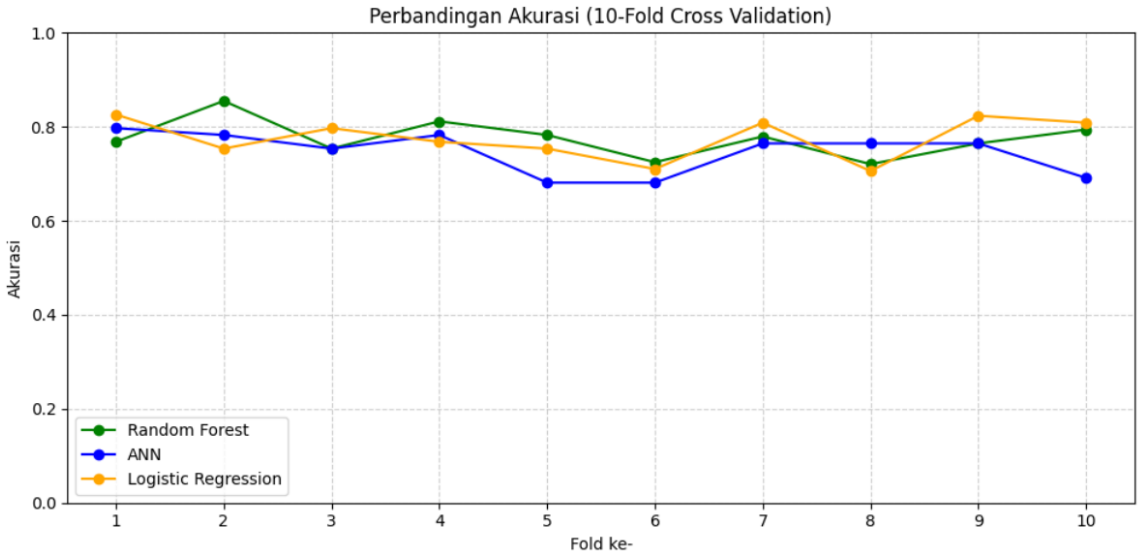
Berdasarkan Tabel 3, model Logistic Regression mampu mencapai akurasi sebesar 80% dalam mengklasifikasikan status pasien kanker payudara. Pada kelas 0, model menunjukkan precision sebesar 0,80, recall 0,87, dan f1-score 0,83. Sementara itu, untuk kelas 1, nilai precision sebesar 0,81, recall 0,72, dan f1-score 0,76. Rata-rata makro untuk precision, recall, dan f1-score masing-masing adalah 0,81, 0,79, dan 0,80. Nilai rata-rata tertimbang juga konsisten dengan hasil tersebut. Hal ini mengindikasikan bahwa Logistic Regression mampu memberikan kinerja prediksi yang cukup seimbang antara kedua kelas, dengan performa yang sedikit lebih unggul dalam mengidentifikasi kasus negatif (kelas 0).



Gambar 4. Confusion Matrix Logistic Regression

3.2 Evaluasi Hasil Perbandingan K- Fold Cross Validation

Subbab ini menyajikan evaluasi hasil perbandingan kinerja tiga model klasifikasi, yaitu Random Forest, Artificial Neural Network (ANN), dan Logistic Regression, dengan menggunakan metode 10-Fold Cross Validation. Evaluasi ini bertujuan untuk menilai konsistensi, keandalan, serta stabilitas performa ketiga model dalam melakukan prediksi terhadap diagnosis kanker payudara secara lebih menyeluruh dan objektif. Adapun hasil perbandingan K-Fold dapat dilihat pada Gambar 5.



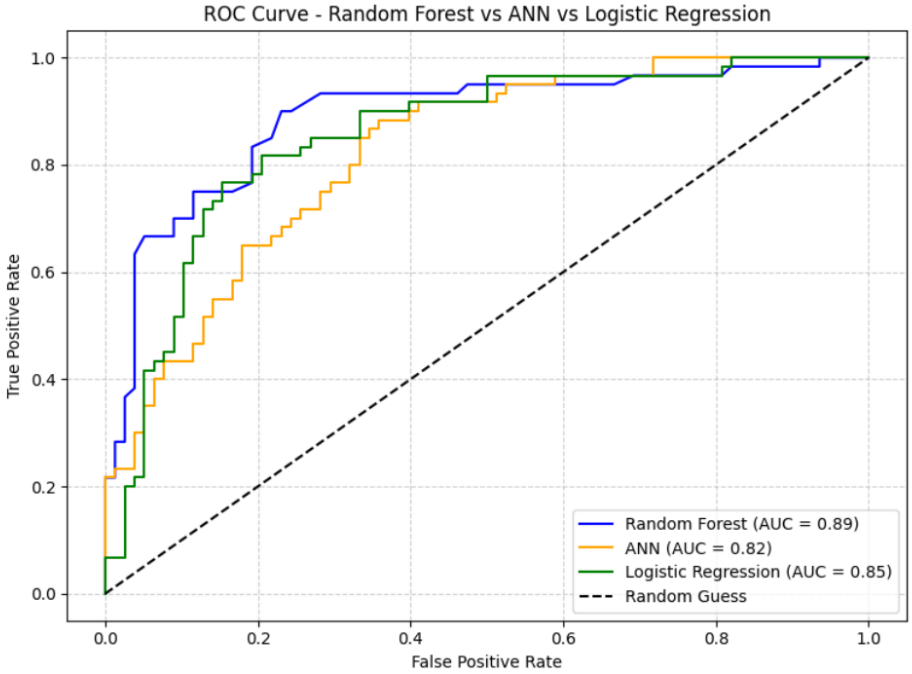
Gambar 5. Evaluasi Hasil Perbandingan K- Fold Cross Validation

Berdasarkan Gambar 5 yang menunjukkan hasil evaluasi rata-rata akurasi dari proses validasi menggunakan metode 10-Fold Cross Validation, terlihat bahwa model Logistic Regression memperoleh nilai akurasi tertinggi sebesar 0,7756, sedikit lebih tinggi dibandingkan dengan model Random Forest yang memiliki akurasi rata-rata sebesar 0,7754. Meskipun perbedaan tersebut sangat kecil dan dapat dianggap tidak signifikan secara statistik, hal ini menunjukkan bahwa kedua model memiliki performa

yang relatif setara dalam mengklasifikasikan data. Sementara itu, model Artificial Neural Network (ANN) menunjukkan akurasi terendah, yaitu sebesar 0,7464, yang mengindikasikan bahwa model ini kurang optimal dalam menangani dataset yang digunakan jika dibandingkan dengan dua model lainnya.

3.3 Kurva Receiver Operating Characteristic

Kurva Receiver Operating Characteristic (ROC) merupakan salah satu metode evaluasi yang digunakan untuk mengukur kinerja model klasifikasi dengan membandingkan nilai true positive rate (TPR) terhadap false positive rate (FPR) pada berbagai ambang batas klasifikasi. Kurva ini memberikan gambaran visual mengenai kemampuan model dalam membedakan antara kelas positif dan negatif, di mana semakin mendekati sudut kiri atas grafik, semakin baik performa model. Adapun hasik Kurva ROC dapat dilihat pada Gambar 6.



Gambar 6. Perbandingan Model dengan Kurva Receiver Operating Characteristic

Berdasarkan Gambar 6 yang menampilkan kurva Receiver Operating Characteristic (ROC), ketiga model Random Forest, Artificial Neural Network (ANN), dan Logistic Regression yang menunjukkan performa yang lebih baik dibandingkan baseline random guess. Nilai Area Under the Curve (AUC) masing-masing model tercatat sebesar 0.89 untuk Random Forest, 0.82 untuk ANN, dan 0.85 untuk Logistic Regression. Random Forest memperoleh nilai AUC tertinggi, yang mengindikasikan kemampuannya yang paling unggul dalam membedakan antara kasus positif dan negatif. Posisi ini diikuti oleh Logistic Regression dan ANN dengan nilai AUC yang sedikit lebih rendah.

3.4 Running Time Classifer

Evaluasi terhadap running time classifier dilakukan untuk mengukur efisiensi komputasi dari setiap model dalam memproses data. Berdasarkan hasil pengujian, dapat diamati perbedaan waktu eksekusi yang signifikan antara algoritma Random Forest, Artificial Neural Network (ANN), dan Logistic Regression, yang mencerminkan kompleksitas dan beban komputasi masing-masing pendekatan. Adapun hasilnya dapat dilihat pada Tabel

Tabel 5. Running Time Classifer

Classifier	Running Time
Random Forest	0.459078 detik
Artificial Neural Network	1.727734 detik
Logistic Regression	0.024897 detik

3.5 Implikasi Analisis Model

Hasil evaluasi dari tiga model klasifikasi—Random Forest, Artificial Neural Network (ANN), dan Logistic Regression—menunjukkan variasi performa dari segi akurasi, efisiensi waktu eksekusi, serta kemampuan diskriminasi antar kelas yang ditunjukkan melalui nilai AUC. Beberapa implikasi penting dari hasil ini dapat dianalisis sebagai berikut:

1. Efisiensi Eksekusi Model
   Model Logistic Regression menunjukkan waktu eksekusi tercepat, yaitu sebesar 0,024897 detik, jauh lebih cepat dibandingkan Random Forest (0,459078 detik) dan ANN (1,727734 detik). Hal ini dapat dijelaskan oleh kompleksitas algoritma:
   a) *Logistic Regression* merupakan model linier dengan struktur matematis yang relatif sederhana. Proses pelatihan hanya memerlukan estimasi parameter melalui metode optimisasi seperti gradient descent tanpa harus membentuk struktur seperti pohon atau jaringan.
   b) *Random Forest* merupakan ensemble dari banyak pohon keputusan yang harus dibentuk dan diuji secara bersamaan. Ini membutuhkan lebih banyak waktu dan memori komputasi.
   c) *Artificial Neural Network* memiliki arsitektur yang lebih kompleks dengan banyak neuron dan bobot yang harus dilatih iteratif, sehingga waktu prosesnya cenderung paling tinggi. Implikasinya, Logistic Regression sangat sesuai untuk sistem prediksi real-time atau deployment pada perangkat dengan sumber daya terbatas karena efisiensinya, tanpa harus mengorbankan akurasi secara signifikan.
2. Kinerja Klasifikasi
   Logistic Regression merupakan model linier, performanya dalam hal akurasi (80%) dan AUC (0.85) berada pada tingkat yang sangat kompetitif dibandingkan Random Forest (akurasi 80%, AUC 0.89). Ini menunjukkan bahwa:
   a) Dataset kanker payudara yang digunakan memiliki pola klasifikasi yang relatif linier atau cukup sederhana untuk dipetakan dengan pendekatan linier.
   b) Logistic Regression dapat memberikan hasil prediksi yang seimbang tanpa perlu arsitektur kompleks, selama data tidak terlalu non-linear.
   c) Random Forest tetap unggul secara keseluruhan dalam hal kemampuan membedakan kelas (nilai AUC tertinggi 0.89) dan f1-score yang lebih seimbang antara kelas 0 dan Hal ini membuat Random Forest cocok untuk kasus yang membutuhkan klasifikasi yang lebih teliti dan presisi tinggi.
3. Risiko Overfitting dan Sensitivitas
   Model Artificial Neural Network menunjukkan performa yang cenderung lebih rendah (akurasi 78%, AUC 0.82) meskipun secara teori memiliki kemampuan representasi lebih tinggi. Hal ini mengindikasikan adanya kemungkinan overfitting atau sensitivitas terhadap hyperparameter (jumlah neuron, learning rate, epoch) yang tidak optimal pada proses pelatihan.
   a) ANN sangat sensitif terhadap konfigurasi arsitektur dan parameter, serta memerlukan jumlah data besar untuk mencapai performa optimal.
   b) Dalam kasus ini, ukuran dataset yang terbatas (n=138) tidak cukup untuk melatih ANN secara efektif, yang menyebabkan hasil prediksi pada kelas minoritas (positif) menjadi kurang akurat (recall hanya 0.68).
   c) Random Forest memiliki keunggulan dalam mengurangi overfitting melalui mekanisme random sampling dan ensemble voting, sedangkan Logistic Regression cenderung stabil jika data tidak terlalu kompleks.

4. KESIMPULAN

Berdasarkan hasil evaluasi terhadap tiga model Random Forest, Artificial Neural Network (ANN), dan Logistic Regression dapat disimpulkan bahwa secara umum ketiganya menunjukkan performa yang cukup baik dalam mengklasifikasikan data diagnosis kanker payudara, dengan akurasi yang berkisar antara 74,64% hingga 80%. Model Random Forest menghasilkan akurasi sebesar 80% berdasarkan classification report dan 77,54% dari rata-rata 10-Fold Cross Validation. Model Artificial Neural Network (ANN) memperoleh akurasi 78% pada classification report dan 74,64% dari validasi silang. Sementara itu, model Logistic Regression mencatat akurasi 80% berdasarkan classification report dan memiliki akurasi rata-rata tertinggi sebesar 77,56% dari 10-Fold Cross Validation. Logistic Regression juga menunjukkan performa paling konsisten, dengan nilai f1-score yang seimbang untuk kedua kelas. Random Forest unggul dalam kemampuan diskriminasi antara kelas positif dan negatif dengan nilai AUC tertinggi sebesar 0,89. Di sisi lain, ANN menunjukkan performa yang relatif lebih rendah, terutama dalam mengenali kelas positif, yang tercermin dari nilai recall sebesar 0,68. Dari segi efisiensi komputasi, Logistic Regression memiliki waktu eksekusi tercepat, yaitu 0,024897 detik, jauh lebih efisien dibandingkan Random Forest (0,459078 detik) dan ANN (1,727734 detik). Dengan demikian, Logistic Regression dapat dipertimbangkan sebagai model yang paling seimbang dari segi akurasi, efisiensi, dan kestabilan kinerja, sedangkan Random Forest dapat menjadi pilihan alternatif apabila prioritas utama adalah kemampuan klasifikasi yang lebih tinggi.

REFERENSI

[1] A. A. Fauzi *Dkk.*, "No Title," In *Pemanfaatan Teknologi Informasi Di Berbagai Sektor Pada Masa Society 5.0*, Pt. Sonpedia Publishing Indonesia, 2023.

[2] H. Najwa, "Analisis Penerapan Trust Network Access (Ztna) Dengan Penggunaan Captcha Pada Website Umum," *Technol. Sci. Insights J.*, Vol. 1, No. 2, Hal. 76–80, 2024.

[3] E. Anggraini, "Masa Depan Internet Of Things Dimulai Dari Rumah," 2017. [Daring]. Tersedia Pada: Https://Www.Cnnindonesia.Com/Teknologi/20170105174130-185-

[4] M. Mukhsin, "Peranan Teknologi Informasi Dan Komunikasi Menerapkan Sistem Informasi Desa Dalam Publikasi Informasi Desa Di Era Globalisasi," *Teknokom*, Vol. 3, No. 1, Hal. 7–15, 2020.

[5] M. Suhaidi, "Penerapan Internet Of Things (Iot) Dalam Perancangan Aplikasi Pengaman Sepeda Motor Berbasis Android," 2019. Doi: 10.14257/Ijca.2014.7.12.07.

[6] M. B. Yel Dan M. K. Nasution, "Keamanan Informasi Data Pribadi Pada Media Sosial," *J. Inform. Kaputama (Jik*, Vol. 6, No. 1, Hal. 92–101, 2022.

[7] A. Khaidar, M. Arhami, Dan M. Abdi, "Application Of The Random Forest Method For Ukt Classification At Politeknik Negeri Lhokseumawe," *J. Artif. Intell. Softw. Eng.*, Vol. 4, No. 2, Hal. 94–103, 2024.

[8] A. Rusadi, Z. Ardian, Dan N. Nurdin, *Apl. Pencarian Guru Les Priv. Terdekat Menggunakan Metod. Haversine Formula. J. Informatics Comput. Sci.*, Vol. 10, No. 2, Hal. 75–80, 2024.

[9] A. Afdalia Dan N. A. Manaf, "Hubungan Penggunaan Kontrasepsi Hormonal Dengan Kejadian Kanker Payudara Di Rumah Sakit Grestelina Makassar," 2020.

[10] L. Pratiwi *Dkk.*, *Mengenal Mencegah*. Kanker Payudara: Sudut Pandang Teori & Penelitian. Cv Jejak (Jejak Publisher, 2024.

[11] H. A. Amalia Wahab, "Prevalensi Dan Karakteristik Penderita Kanker Payudara Di Poliklinik Bedah Onkologi Rsup Dr," 2024.

[12] U. Muhidin, "Faktor Yang Berhubungan Dengan Upaya Pencegahan Kanker Payudara Dengan Metode Sadari Pada Siswi Smkn 5 Enrekang= Factors Related To Efforts To Prevent Breast Cancer By Sadari Method The Regency Of Smk 5 Enrekang Students," 2022.

[13] N. Izzah, "Faktor Yang Berhubungan Dengan Upaya Pencegahan Kanker Payudara Di Madrasah Aliyah Negeri 2 Kota Makassar," 2024.

[14] L. N. Martini, "Hubungan Penggunaan Kontrasepsi Hormonal Dengan Kejadian Kanker Payudara Di Rsud Kabupaten Buleleng," 2024.

[15] F. A. Azlina Dan R. Firdausi, "Mengenal Kanker Serviks Dan Upaya Dalam Meningkatkan Deteksi Dini," 2025.

[16] E. Pujiati, "Health Education On Early Detection Of Breast Cancer Through Audio Visual Breast Self Examination In Women Of Rolling Age: Study Cross Sectional," *Menara J. Heal. Sci.*, Vol. 3, No. 1, Hal. 190–201, 2024.

[17] G. T. Reddy *Dkk.*, "February). An Ensemble Based Machine Learning Model For Diabetic Retinopathy Classification," In *2020 International Conference On Emerging Trends In Information Technology And Engineering*, Ieee, 2020, Hal. 1–6.

[18] B. M. Hermawan, M. A. Hakim, R. Arifin, Dan N. Puspitasari, "Pemanfaatan Artificial Intellegence, Khususnya Mechine Learning Dan Deep Learning System Dalam Pendidikan," In *Prosiding Seminar Nasional Amikom Surakarta*, Vol. 2, 2024, Hal. 345–354.

[19] M. O. Okwu, L. K. Tartibu, M. O. Okwu, Dan L. K. Tartibu, "Artificial Neural Network," In *Metaheuristic Optimization: Nature-Inspired Algorithms Swarm And Computational Intelligence, Theory And Applications*, 2021, Hal. 133–145.

[20] R. Dastres Dan M. Soori, "Artificial Neural Network Systems," *Int. J. Imaging Robot. (Ijir*, Vol. 21, No. 2, Hal. 13–25, 2021.

[21] R. Gomila, "Logistic Or Linear? Estimating Causal Effects Of Experimental Treatments On Binary Outcomes Using Regression Analysis," *J. Exp. Psychol. Gen.*, Vol. 150, No. 4, Hal. 700, 2021.

[22] C. Z. V Junus, T. Tarno, Dan P. Kartikasari, "Klasifikasi Menggunakan Metode Support Vector Machine Dan Random Forest Untuk Deteksi Awal Risiko Diabetes Melitus," *J. Gaussian*, Vol. 11, No. 3, Hal. 386–396, 2023.

[23] D. Ismafillah, T. Rohana, Dan Y. Cahyana, "Analisis Algoritma Pohon Keputusan Untuk Memprediksi Penyakit Diabetes Menggunakan Oversampling Smote," *Infotech J. Inform. Teknol.*, Vol. 4, No. 1, Hal. 27–36, 2023.

[24] I. Suhardin, A. Patombongi, Dan A. M. Islah, "Mengidentifikasi Jenis Tanaman Berdasarkan Citra Daun Menggunakan Algoritma Convolutional Neural Network," *Simtek J. Sist. Inf. Dan Tek. Komput.*, Vol. 6, No. 2, Hal. 100–108, 2021.

[25] D. Adhito Dwi, "Rancang Bangun Sistem Electronic Nose (E-Nose," In *Berbasis Multisensor Menggunakan Jaringan Syaraf Tiruan Metode Backpropagation Dengan Sampel Uji Boraks*, 2023.

[26] A. I. Sofiyat, A. Tjalla, Dan M. Mahdiyah, "Pemodelan Regresi Logistik Biner Terhadap Penerimaan Pegawai Di Pt Xyz Jakarta," *Mat. Sains*, Vol. 1, No. 1, Hal. 1–11, 2023.

[27] A. Salam, L. Azhari, R. S. Septarini, Dan N. Heriyani, "Pendekatan Hybrid K-Means Smote Dan Logistic Regression Untuk Deteksi Dini Diabetes Mellitus Pada Imbalanced Data," *Bull. Comput. Sci. Res.*, Vol. 5, No. 3, Hal. 219–227, 2025.