Jurnal Kecerdasan Buatan dan Teknologi Informasi
Vol. 4, No. 1, January 2025, hlm. 33-42
e-ISSN: 2964-2922, p-ISSN: 2963-6191
DOI : https://doi.org/10.69916/jkbti.v4i1.191

PERBANDINGAN ALGORITMA *SUPPORT VECTOR MACHINE (SVM)* DAN *LOGISTIC REGRESSION* DALAM KLASIFIKASI KANKER PAYUDARA

Anita Desiani[*1], Des Alwine Zayanti[2], Indri Ramayanti[3], Faishal Fitra Ramadhan[4] , Giovillando[5]

[1,2,4,5]Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sriwijaya, Palembang, Indonesia
[3]Ilmu Kedokteran, Fakultas Kedokteran, Universitas Muhammadiyah Palembang, Palembang, Indonesia
Email: [1]anita_desiani@unsri.ac.id, [2]desalwinez@unsri.ac.id, [3]indri_ramayanti@um-palembang.ac.id, [4]faishalfitraramadhan01@gmail.com, [5]gio.villando11@.gmail.com

(Diterima : 17 November 2024, Direvisi : 27 Desember 2024, Disetujui : 2 Januari 2025)

Abstrak

Kanker payudara memberikan dampak fisik dan dampak psikologis pada pasien. Deteksi dini terhadap kanker payudara dibutuhkan pada pengidap yang berisiko mengidap kanker payudara. Salah satu solusi yang bisa dilakukan untuk deteksi dini penyakit kanker payudara yaitu dengan melakukan klasifikasi menggunakan pendekatan *data mining* menggunakan algoritma *Support Vector Machine (SVM)* dan *Algoritma Logistic regression (LR)* dengan teknik pengujian *Percentage Split* dan *K-Fold Cross Validation*. Penelitian ini bertujuan untuk mendapatkan hasil klasifikasi terbaik untuk mendeteksi penyakit kanker payudara dengan membandingkan kedua algoritma tersebut. Hasil Akurasi yang dihasilkan dari penelitian ini yaitu pada algoritma SVM diperoleh 96% dengan *Percentage Split* dan 98% pada metode *K-Fold Cross Validation*. Sementara pada *Algoritma Logistic regression* didapat hasil akurasi sebesar 96% pada metode *Percentage Split* dan 97% untuk *metode K-Fold Cross Validation*. Berdasarkan hasil akurasi, algoritma SVM dengan metode *K-Fold Cross Validation* merupakan algoritma terbaik dalam mengklasifikasi penyakit kanker payudara. Namun, hasil akurasi dari LR masih bisa dikatakan sangat baik karena menghasilkan akurasi lebih dari 90%.

Kata kunci: *data mining*, kanker payudara, klasifikasi, *logistic regression*, SVM.

COMPARISON OF SUPPORT VECTOR MACHINE (SVM) AND LOGISTIC REGRESSION ALGORITHMS IN BREAST CANCER CLASSIFICATION

Abstract

Breast cancer has both a physical and psychological impact on patients. Early detection of breast cancer is needed in people who are at risk of developing breast cancer. One solution that can be done for early detection of breast cancer is by classifying using a data mining approach using the Support Vector Machine (SVM) algorithm and the Logistic regression (ALR) algorithm with Percetange Split and K-Fold Cross Validation testing techniques. This research aims to get the best classification results for detecting breast cancer by comparing the two algorithms. The accuracy results generated from this study are in the SVM algorithm obtained 96% in the Percetange Split method and 98% in the K-Fold Cross Validation method. While the Logistic regression algorithm obtained an accuracy of 96% for the Percetange Split method and 97% for the K-Fold Cross Validation method. Based on the accuracy results, the SVM algorithm K-Fold Cross Validation method is the best algorithm in classifying breast cancer diseases. However, the accuracy results of ALR can still be said to be very good because it is more than 90%.

Keywords: *breast cancer, classification, data mining, logistic regression, SVM.*

1. PENDAHULUAN

Kanker adalah tumor ganas yang dapat menyerang jaringan tubuh dan menyebar ke organ lain. Penyakit ini merupakan salah satu penyebab utama kematian global, dengan tercatat Jumlah kasus baru akibat kanker sampai dengan tahun 2020 di dunia adalah 19,2 juta jiwa. Sedangkan jumlah kematian akibat kanker tahun 2020 di dunia mencapai 9,9 juta jiwa [1]. Salah satu jenis kanker yang paling sering menyerang perempuan adalah kanker payudara, yang menyumbang 30% dari seluruh kasus kanker pada perempuan dan merupakan jenis kanker paling dominan di Indonesia [2]. Kanker payudara adalah kanker pada jaringan payudara yang disebabkan oleh sel-sel ganas yang berasal dari komponen kelenjar seperti pembuluh darah, jaringan saraf, dan jaringan lemak di payudara

[3]. Meskipun penyebab pasti kanker payudara belum diketahui, beberapa faktor risiko telah diidentifikasi, termasuk kelemahan genetik pada sel tubuh, iritasi dan inflamasi kronis pada payudara, paparan radiasi sinar-X atau sinar matahari berlebihan, serta konsumsi senyawa karsinogenik dalam jumlah berlebihan [4]. Pada tahun 2020, terdapat 396.914 kasus kanker di Indonesia, dengan 68.858 kasus baru kanker payudara yang menyebabkan sekitar 22.000 kematian [5]. Untuk menekan jumlah kasus kanker payudara, deteksi dini sangat penting, terutama pada individu yang memiliki risiko tinggi. Salah satu pendekatan dalam deteksi dini adalah pemanfaatan metode *data mining*. *Data mining* menawarkan solusi efektif untuk analisis data kesehatan, terutama dalam deteksi kanker payudara [6]. Metode ini memungkinkan pengolahan data medis dalam jumlah besar secara sistematis untuk menemukan pola-pola yang mungkin tidak terlihat secara kasat mata [6]. Dengan memanfaatkan *data mining*, tenaga medis dapat membuat keputusan diagnosis yang lebih akurat dan objektif berdasarkan pembelajaran dari data historis pasien [7]. Pendekatan ini juga memungkinkan identifikasi faktor-faktor risiko yang signifikan dan dapat membantu dalam perencanaan strategi pencegahan yang lebih efektif [8].

Data mining adalah teknik yang digunakan untuk menemukan informasi tersembunyi dalam dataset yang telah terpilih [9]. Beberapa algoritma yang umum digunakan dalam *data mining* adalah *Support Vector Machine (SVM)* dan *Logistic regression* (LR). *Support Vector Machine (SVM)* adalah algoritma yang bekerja dengan mencari *hyperplane* optimal untuk memisahkan dua kelas pada ruang input menggunakan prinsip *Structural Risk Minimization* (SRM) [10]. Algoritma ini memiliki stabilitas dan kemampuan generalisasi yang baik dibandingkan algoritma lainnya [11]. Penelitian sebelumnya yang menunjukkan keunggulan SVM, seperti penelitian oleh Amalia [12] yang mengaplikasikan SVM pada klasifikasi penyakit ginjal kronis dan mencapai akurasi 95,16%. Apriyani [13] menggunakan SVM untuk klasifikasi diabetes melitus dengan akurasi 96,27%, sementara Budiman dan Niqotaini [14], mengaplikasikan SVM untuk prediksi penyakit kulit dengan akurasi 98,1%. Namun, algoritma SVM memiliki kelemahan pada dataset berukuran besar karena waktu pelatihan yang cukup lama [15]. Sebaliknya, *Logistic regression* (LR) memiliki keunggulan dalam memproses dataset besar dengan efisiensi tinggi karena kebutuhan komputasi yang rendah [16].

Logistic regression adalah metode analisis multivariat yang bertujuan memprediksi variabel dependen berdasarkan variabel independent [17]. Algoritma ini menganalisis data kategorikal dengan variabel respon biner (Y) dan variabel penjelas kontinu atau kategorikal (X) [18]. Parameter yang dihasilkan oleh algoritma ini memberikan wawasan mengenai pentingnya setiap fitur, sehingga hubungan antar fitur dapat diinterpretasikan [19]. Penelitian sebelumnya yang menunjukkan potensi *Algoritma Logistic regression*, seperti penelitian oleh Azhar [20] yang menggunakan algoritma LR untuk klasifikasi penyakit stroke dengan akurasi 98,63%. Penelitian oleh Ismafillah [21] juga menunjukkan akurasi 93,1% untuk klasifikasi stroke, sementara Nugroho [22], menggunakan LR untuk klasifikasi penyakit kardiovaskular. Meskipun algoritma ini cocok untuk dataset besar, Algoritma LR rentan terhadap masalah *underfitting* pada dataset dengan kelas yang tidak seimbang sehingga dapat menurunkan akurasi. Penelitian ini bertujuan membandingkan kinerja algoritma SVM dan *Logistic regression* untuk klasifikasi kanker payudara. Perbandingan antara algoritma SVM dan *Logistic regression* dipilih karena keduanya memiliki karakteristik yang saling melengkapi, di mana SVM unggul dalam stabilitas dan generalisasi sementara LR efisien dalam memproses dataset besar.

Penelitian ini memberikan beberapa kontribusi signifikan dalam pengembangan sistem klasifikasi kanker payudara. Pertama, penelitian ini menghadirkan evaluasi komparatif sistematis antara algoritma SVM dan LR, sehingga menghasilkan analisis komprehensif tentang performa kedua algoritma dalam konteks diagnostik kanker payudara [23]. Kedua, implementasi *dual-validation framework* yang mengintegrasikan *Percentage Split dan K-Fold Cross Validation* memberikan evaluasi model yang lebih *robust* dan *reliable*, meminimalisasi bias dalam pengukuran performa [24]. Ketiga, hasil perbandingan algoritma ini berkontribusi pada pengembangan *Computer-Aided Diagnosis* (CAD) *system* untuk deteksi kanker payudara, menyediakan fundamental metodologis bagi pengembangan sistem pendukung keputusan klinis yang lebih akurat [25]. Dua metode validasi yang digunakan pada penelitian ini adalah *Percentage Split* dan *K-Fold Cross Validation*. Pada kedua metode ini, dataset dibagi menjadi 80% data pelatihan dan 20% data pengujian, dengan nilai *K* pada *K-Fold Cross Validation* sebesar 10. Klasifikasi akan dilakukan untuk membedakan kanker payudara jinak (*Benign*, B) dan ganas (*Malignant*, M). Kinerja kedua algoritma akan dievaluasi berdasarkan metrik akurasi, presisi, dan *recall* untuk menentukan algoritma terbaik dalam mendeteksi kanker payudara.

2. METODE PENELITIAN

Metode Penelitian bertujuan untuk memperoleh hasil yang sesuai dengan gambaran garis besar. Dari proses analisis ini, digunakan 4 jenis tahapan. Gambaran alir proses penelitian dapat dilihat pada gambar 1 berikut ini:

Gambar 1. Alur Proses Penelitian (Sumber: Penulis)

2.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah dataset yang diperoleh dari situs *kaggle* yaitu https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset dengan format csv. Dataset yang diambil dari situs *kaggle* merupakan dataset Kanker Payudara yang terdiri dari 569 data serta memiliki 31 atribut dimana salah satu atribut adalah atribut diagnosa dan pada atribut terdapat label *Malignent* (M) yang merupakan kanker payudara ganas dengan jumlah 212 data dan *Benign* (B) merupakan kanker payudara jinak dengan jumlah 357 data. Adapun atribut-atribut pada dataset yang dapat dilihat pada tabel 1 berikut.

Tabel 1. *Input* Algoritma Dataset

No	Atribut	Nilai Atribut	Tipe Data
1	Diagnosis	M (-1), B (+1)	Kategorik
2	*Texture mean*	9.71 – 39.28	Numerik
3	*area mean*	143.5 - 2501	Numerik
4	*Perimeter mean*	43.79 – 188.5	Numerik
5	*Smoothness mean*	0.05263 – 0.1634	Numerik
6	*Concavity mean*	0 – 0.4268	Numerik
7	*Compactness mean*	0.01938 – 0.3454	Numerik
8	*Concave points mean*	0 – 0.2012	Numerik
9	*Fractal dimensxtureion mean*	0.04996 – 0.09744	Numerik
10	*Symmetry mean*	0.106 – 0.304	Numerik
11	*Texture se*	0.3602 – 4.885	Numerik
12	*Radius se*	0.1115 – 2.873	Numerik
13	*Area se*	6.802 – 542.2	Numerik
14	*Perimeter se*	0.757 – 21.98	Numerik
15	*Compactness se*	0.002252 – 0.1354	Numerik
16	*Smoothness se*	0.001713 – 0.03113	Numerik
17	*Concave points se*	0 – 0.05279	Numerik
18	*Concavity se*	0 – 0.396	Numerik
19	*Fractal dimension se*	0.0008949 – 0.02984	Numerik
20	*Symmetry se*	0.007882 – 0.07895	Numerik
21	*Texture worst*	12.02 – 49.54	Numerik
22	*Radius worst*	7.93 – 36.04	Numerik
23	*Area worst*	185.2 – 4254	Numerik
24	*Perimeter worst*	50.41 – 251.2	Numerik
25	*Compactness worst*	0.02729 – 1.058	Numerik
26	*Smoothness worst*	0.07117 – 0.2226	Numerik
27	*Concave points worst*	0 – 0.291	Numerik
28	*Concavity worst*	0 – 1.252	Numerik
29	*Fractal dimension worst*	0.05504 – 0.2075	Numerik
30	*Symmetry worst*	0.1565 – 0.6638	Numerik
31	*Radius mean*	6.981 – 28.11	Numerik

2.2. Pengolahan Data Awal

Pada pengolahan data awal, *range data* akan diperiksa pada masing masing atribut. Ditemukan 9 atribut yang memiliki range data yang harus dinormalisasi yaitu *Texture mean, area mean, Perimeter mean, Area se, Texture worst, Radius worst, Area worst, Perimeter worst,* dan *Radius mean*. Tujuan dilakukan normalisasi adalah untuk memperoleh hasil klasifikasi yang lebih baik dengan cara nilai rentang setiap atribut disamakan dengan skala skala dari 0 (*min*) hingga 1 (*max*) [26]. Adapun rumus dari normalisasi dituliskan pada persamaan (1) [27].

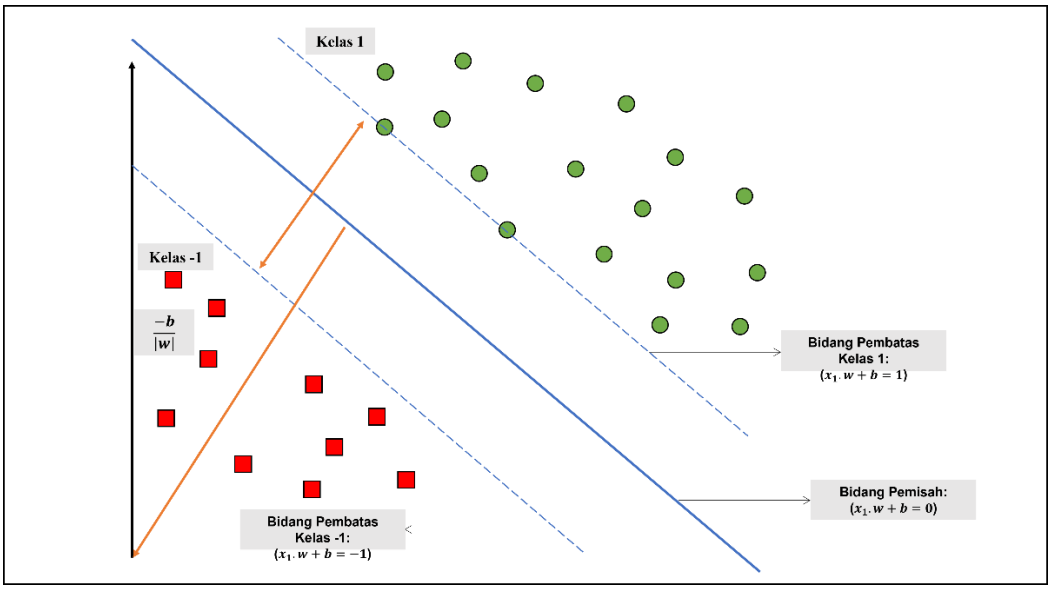
$$normalized\ (x) = \frac{\min Range + (x - \min Value)(\max Range - \min Range)}{\max Value - \min Value} \quad (1)$$

Pada tahap normalisasi, data bertipe kategorik diubah menjadi data yang bertipe numerik. Pada kasus ini atribut Diagnosis akan diubah labelnya dari M dan B menjadi -1 dan 1. Dalam pengolahan data awal juga digunakan beberapa teknik yaitu *K-Fold Cross Validation* dan *Percetange Split*. Tujuan penggunaan teknik pengujian adalah untuk memperoleh data yang berkualitas [28]. *K-Fold Cross Validation* merupakan metode yang berfungsi untuk memeriksa *overfitting* pada model [29]. Pada metode *K-Fold Cross Validation* digunakan nilai k dengan jumlah 10. Sementara itu, *Percetange Split* merupakan Algoritma yang dapat melakukan prediksi presentase data dengan cara mengevaluasi suatu algoritma [30]. Pada penelitian ini, digunakan komposisi sebesar 80% pada data *training* dan 20% data *testing*.

2.3. Penerapan Algoritma

1. *Support Vector Machine*

Algoritma SVM pada proses perhitungannya menggunakan prinsip *Structural Risk Minimization* (SRM) [31]. SVM merupakan algoritma yang digunakan untuk menemukan *hyperplane* optimal yang berfungsi sebagai pembatas/pemisah dua kelas yang berbeda serta untuk memaksimalkan margin antara dua kelas tersebut di ruang input [32]. Berikut merupakan gambar *Hyperplane* Dimensi yang dapat dilihat pada Gambar 2 [33].



Gambar 2. *Hyperplane* Dua Dimensi

Pada Gambar 2, setiap kelas terdefinisi oleh vektor w dan nilai b, yang bersama-sama mendefinisikan sebuah bidang *hyperplane* dalam ruang fitur [33]. Perubahan pada nilai b akan mengakibatkan pergeseran pada posisi *hyperplane* tersebut. Dalam kasus dua kelompok objek yang berasal dari kelas yang berbeda, terdapat sebuah *hyperplane* optimal yang memisahkan keduanya. Untuk menemukan *hyperplane* optimal, tujuan utama adalah memaksimalkan margin, yaitu jarak terbesar antara *hyperplane* dengan titik data terdekat dari kedua kelas. Titik data yang berada pada kelas -1 memenuhi persamaan $w \cdot x_i + b = -1$, sedangkan titik data yang berada pada kelas +1 memenuhi persamaan $w \cdot x_i + b = 1$. Setelah *hyperplane* ditemukan, langkah selanjutnya adalah menerapkan algoritma *Support Vector Machine (SVM)* untuk melakukan klasifikasi. Langkah-langkah dalam algoritma SVM untuk menemukan *hyperplane* optimal dan mengklasifikasikan data yaitu sebagai berikut [34].

a. Tentukan titik data $\{x_1, x_2, \ldots, x_n\}$ yang merupakan atribut pada data.
b. Menentukan kelas : $y_1\{-1, +1\}, -1$ sebagai *Malignent* dan +1 sebagai *Benign.*
c. Menentukan data kelas berdasarkan rumus $\{(x_i, y_i)\}_{i-1}^{N}$ dimana x_1 merupakan vektor baris fitur ke-i, y merupakan label kelas dari x_1, dan N merupakan banyaknya data.
d. Memaksimalkan fungsi berdasarkan persamaan (2).

$$Ld = \sum_{i=1}^{N} d_i - \sum_{i=1}^{N}\sum_{j=1}^{N} k_i\, k_j y_i y_j K(x_i, x_j),\ 0 \le k_i \le C \text{ dan } \sum_{i=1}^{N} k_i y_i = 0 \quad (2)$$

Dengan Ld yaitu dualitas *Langrange Multipier*, K yaitu nilai bobot setiap titik data, C yaitu Nilai Konstanta, N yaitu banyaknya data, dan d yaitu jarak antar setiap data ke *hyperplane*.

e. Menghitung nilai w dan b berdasarkan persamaan (3) dan (4).

$$w = \sum_{i=1}^{N} a_i y_{i.} x_i \quad (3)$$

$$b = \frac{-1}{2}(w.x^i + w.x) \quad (4)$$

dimana w adalah nilai yang berkaitan dengan margin, N adalah banyaknya data, a merupakan nilai angka terendah, x merepresentasikan vektor *input*, y merupakan $\{-1, +1\}$, dan b merupakan bias.

f. Menghitung fungsi keputusan klasifikasi $sign(f(x))$ menggunakan persamaan (5)

$$f(x) = w.x + b \quad (5)$$

atau dengan persamaan (6)

$$f(x) = \sum_{i=1}^{m} a_i y_i K(x, x_j) + b \quad (6)$$

dengan m adalah jumlah titik data yang dimiliki $a_i > 0$, dan $K(x, x_i)$ merupakan fungsi kernel. Pada penelitian ini kernel yang digunakan adalah kernel linier. Adapun pendefinisiannya pada persamaan (7).

$$K(x, y) = x.y \quad (7)$$

2. *Logistic regression*

Algoritma Logistic regression menggunakan maksimum *likelihood* dengan distribusi logistik. Adapun langkah pengerjaan *Algoritma Logistic regression* sebagai berikut [35].

a. Menentukan *dependent variabel* dan *independent variabel* yang merupakan atribut dari data.
b. Menentukan kelas : , -1 sebagai *Malignent* dan +1 sebagai *Benign*.
c. Menentukan nilai acak untuk $\beta_1, \beta_2, \ldots, \beta_n$ sebagai nilai asumsi dasar untuk menentukan *likelihood*.
d. Menghitung nilai Y prediksi menggunakan persamaan (8).

$$Y = a + \beta_{1.} X_{1j} + \beta_{2.} X_{2j} + \cdots + \beta_{n.} X_{nj} \quad (8)$$

dimana a merupakan *intercept*, $\beta_1, \ldots, \beta_n$ adalah atribut *independent* penurunan, n adalah jumlah atribut *independent*, dan j adalah jumlah *record* dalam dataset.

e. Mengeksponensialkan nilai Y prediksi yang didapat dari persamaan (8).
f. Menghitung nilai peluang $P(X)$ dengan menggunakan persamaan (9).

$$P(X) = \frac{e^Y}{1+e^Y} \quad (9)$$

dimana $P(X)$ merupakan peluang kejadian sukses dengan nilai probabilitas $0 \le P(X \le 1)$ dan Y merupakan nilai prediksi.

g. Menghitung maksimum *log likelihood* dengan persamaan (10).

$$maks = Y_i \times \ln\big(P(X) + (1 - Y_i)\big) \times \ln(1 - Y_i) \quad (10)$$

dimana Y_i merupakan nilai *independent variabel* berdasarkan kelas yang sudah ditentukan, dan $P(X)$ adalah nilai hasil yang diperoleh dari persamaan (9).

2.4. Evaluasi Hasil

Confusion Matrix merupakan matriks yang memiliki fungsi untuk menampilkan visualisasi kinerja dari algoritma, serta digunakan untuk menghitung kinerja performa dari suatu model algoritma dalam suatu prediksi aktual dengan bentuk *False Positif* (FP), *True Positif* (TP), *False Negative* (FN), dan *True Negative* (TN) dari informasi [36]. Penjelasan lengkap dari *confusion matrix* bisa dilihat di tabel 2.

Tabel 2. *Confusion Matrix* [36]

Kelas		**Nilai Aktual**	
		Positive	***Negative***
Nilai Prediksi	***Positive***	***True Positive (TP)***	***False Negative (FN)***
	Negative	***False Positive (FP)***	***True Negative (TN)***

3. HASIL DAN PEMBAHASAN

3.1. Algoritma SVM

Pada algoritma SVM metode *percetange split*, digunakan data *training* sebesar 80% dan data *testing* sebesar 20%. Kernel yang digunakan pada algoritma SVM adalah kernel linear. Metode *K-Fold Cross Validation* juga digunakan pada algoritma ini. Nilai *K-Fold* yang digunakan pada penelitian ini adalah 10. Diperoleh *confusion matrix* dari metode *Percetange Split* dan *K-Fold Cross Validation* algoritma SVM yang dapat dilihat pada tabel 3.

Tabel 3. *Confusion Matrix Percetange Split* dan *K-Fold Cross Validation* Algoritma SVM

Percentage Split				***K-Fold Cross Validation***			
Kelas		**Nilai Aktual**		**Kelas**		**Nilai Aktual**	
		-1	**1**			**-1**	**1**
Nilai Prediksi	**-1**	71	1	**Nilai Prediksi**	**-1**	352	5
	1	3	39		**1**	9	203

Pada tabel 3 terlihat bahwa di metode *percetange* Split terdapat 71 data yang diprediksi benar sebagai kategori -1 dan 39 data yang diprediksi benar sebagai kategori 1. Terdapat 1 data yang diprediksi sebagai -1 yang seharusnya adalah 1. Terdapat pula 3 data yang harusnya masuk ke dalam kategori 1, tetapi diprediksi sebagai kategori -1. Sementara itu, untuk metode *K-Fold Cross Validation* 352 data diprediksi benar sebagai kategori -1 dan 203 data diprediksi benar sebagai kategori 1. Sementara itu, terdapat 5 data yang harusnya masuk ke kategori -1, tetapi diprediksi sebagai kategori 1, serta 9 data yang diprediksi sebagai -1 dan masuk ke kategori 1. Dari *Confusion Matrix* yang diperoleh, dihasilkan nilai akurasi, presisi dan *recall* dari masing-masing label yang ditunjukkan oleh tabel 4 berikut ini.

Tabel 4. Nilai Akurasi, Presisi dan *Recall* Metode *Percetange Split* dan *K-Fold Cross Validation* Algoritma SVM

Metode	**Label**	**Presisi**	***Recall***	**Akurasi**
Percetange Split	**-1**	93%	97%	96%
	1	99%	96%	
K-Fold Cross Validation	**-1**	98%	96%	98%
	1	98%	99%	

Tabel 4 menunjukkan nilai akurasi, presisi, dan *recall* dari kedua metode pengujian algoritma SVM pada klasifikasi penyakit kanker payudara. Terlihat bahwa metode *K-Fold Cross Validation* menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan metode *Percetange Split* walaupun nilai presisi dan *recall* dari masing-masing label tidak berbeda jauh.

3.2. *Algoritma Logistic regression*

Hasil klasifikasi *Algoritma Logistic regression* pada penyakit kanker payudara juga menggunakan metode *Percetange Split* dan *K-Fold Cross Validation* dimana tingkat keberhasilannya diukur menggunakan *confusion matrix*. Adapun *Confusion Matrix* dari metode *Percetange Split* dan *K-Fold Cross Validation Algoritma Logistic regression* dapat dilihat pada tabel 5.

Tabel 5. *Confusion Matrix Percetange Split* Algoritma LR

Percentage Split				***K-Fold Cross Validation***			
Kelas		**Nilai Aktual**		**Kelas**		**Nilai Aktual**	
		-1	**1**			**-1**	**1**
Nilai Prediksi	**-1**	72	5	**Nilai Prediksi**	**-1**	355	2
	1	0	37		**1**	14	198

Pada tabel 5, pada metode *Percetange Split* terdapat 72 data yang diprediksi benar sebagai kategori -1, 37 data yang diprediksi benar sebagai kategori 1, 0 data yang diprediksi sebagai -1 yang seharusnya adalah 1, dan 5 data yang harusnya masuk ke dalam kategori 1, tetapi diprediksi sebagai kategori -1. Sementara pada metode *K-Fold Cross Validation* terdapat 355 data diprediksi benar sebagai kategori -1 dan 198 data diprediksi benar sebagai kategori 1. Terdapat pula 2 data yang diprediksi di kategori 1 yang harusnya masuk ke kategori -1, serta 14 data yang harusnya masuk ke dalam kategori 1 tetapi diprediksi sebagai kategori -1. Dari *Confusion Matrix* yang diperoleh, didapat nilai akurasi, presisi, dan *recall* dari masing-masing label yang dituliskan pada tabel 6.

Tabel 6. Nilai Akurasi, Presisi dan *Recall* Metode *Percetange Split* dan *K-Fold Cross Validation* Algoritma LR

Metode	**Label**	**Presisi**	***Recall***	**Akurasi**
Percetange Split	-1	91%	100%	96%

	1	100%	95%	
K-Fold Cross Validation	-1	99%	93%	97%
	1	96%	99%	

Tabel 6 menunjukkan nilai akurasi, presisi, dan *recall* dari kedua metode pengujian *Algoritma Logistic regression* pada klasifikasi penyakit kanker payudara. Pada *Algoritma Logistic regression*, metode *K-Fold Cross Validation* Hasil evaluasi *K-Fold Cross Validation* menunjukkan bahwa presisi untuk label -1 meningkat menjadi 99% dibandingkan dengan *Percentage Split*, meskipun *recall* sedikit menurun menjadi 93%. Untuk label 1, presisi tercatat sebesar 96%, dengan *recall* yang lebih tinggi, yaitu 99%. Akurasi keseluruhan pada metode ini mencapai 97%, menunjukkan bahwa metode *K-Fold Cross Validation* memberikan performa yang lebih konsisten dibandingkan dengan metode *Percentage Split*.

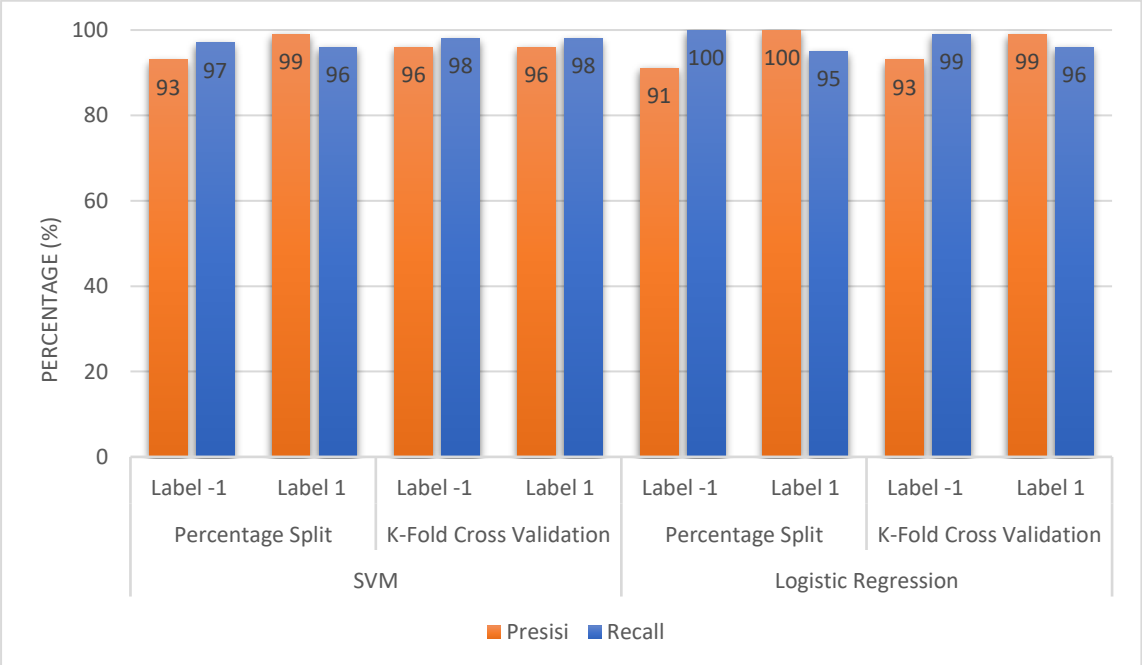
3.3. Perbandingan Kedua Algoritma

Nilai akurasi dari algoritma SVM dan *Algoritma Logistic regression* dengan metode *percetange split* dan *K-Fold Cross Validtion* pada penyakit kanker payudara menunjukkan bahwa algoritma SVM dan ALR dengan kedua metode tersebut sangat baik dalam mengklasifikasi penyakit kanker payudara. Adapun perbandingan nilai akurasi, *recall*, dan presisi dari kedua algoritma, baik metode *percetange split* maupun metode *K-Fold Cross Validation* dapat dilihat pada tabel 7 sebagai berikut.

Tabel 7. Perbandingan Nilai Akurasi, *Recall*, dan Presisi Algoritma SVM dan LR

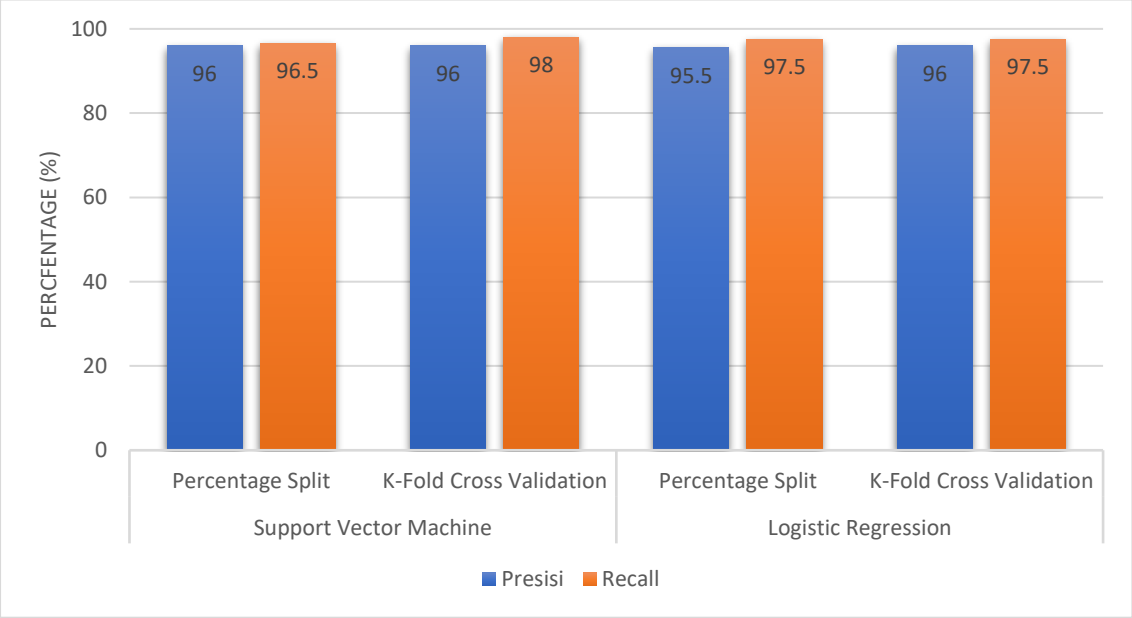
Algoritma	**Pemodelan**	**Label**	**Presisi**	***Recall***	**Akurasi**
SVM	*Percetange Split*	**-1**	93%	97%	96%
		1	99%	96%	
	K-Fold Cross Validation	**-1**	96%	98%	98%
		1	96%	98%	
Logistic regression	*Percetange Split*	**-1**	91%	100%	96%
		1	100%	95%	
	K-Fold Cross Validation	**-1**	93%	99%	97%
		1	99%	96%	

Pada tabel 7, terlihat bahwa pada algoritma SVM metode *K-Fold Cross Validation* memiliki tingkat akurasi yang lebih tinggi dibanding metode *Percetange Split*. Sedangkan pada *Algoritma Logistic regression*, metode *K-Fold Cross Validation* menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan metode *Percetange Split*. Meskipun demikian, kedua algoritma menghasilkan nilai akurasi, *recall*, dan presisi yang tinggi dengan baik menggunakan metode *Percetange Split* maupun metode *K-Fold Cross Validation*. Adapun perbandingan dari nilai *recall* dan presisi dari kedua algoritma menggunakan metode *Percetange Split* dan *K-Fold Cross Validation* bisa dilihat lebih jelas dalam Gambar 3.



Gambar 3. Hasil *Percetange Split* dan *K-Fold Cross Validation*

Pada gambar 3 diatas dapat dilihat bahwa nilai presisi dan *recall* algoritma SVM dan LR tidak berbeda jauh walaupun memiliki sedikit perbandingan. Untuk melihat perbandingan yang lebih baik perlu dilakukan perhitungan keseluruhan hasil *recall*, presisi dan akurasi dari kedua algoritma. Perhitungan nilai rata-rata presisi dan *recall* dari algoritma SVM dan *Logistic regression* menggunakan metode *Percetange Split* dan *K-Fold Cross Validation* dapat dilihat lebih jelas pada gambar 4 berikut ini.



Gambar 4. Perbandingan Rata-Rata Nilai Presisi dan *Recall*

Dari gambar 4 dapat dilihat bahwa nilai presisi dan *recall* yang dihasilkan oleh metode *K-Fold Cross Validation* lebih baik dibandingkan dengan *Percetange Split* pada algoritma SVM dan *Logistic regression*. Hasil *K-Fold Cross Validation* yang diperoleh menggunakan algoritma SVM lebih baik dibanding *Algoritma Logistic regression*. Dari keterangan gambar 4, maka dapat disimpulkan bahwa algoritma SVM lebih baik dibandingkan dengan *Logistic regression* dalam melakukan klasifikasi kanker payudara

4. KESIMPULAN

Berdasarkan penelitian dan pembahasan yang dilakukan untuk mengklasifikasi penyakit kanker payudara dari *dataset* yang diperoleh, proses klasifikasi dilakukan dengan menggunakan teknik pengujian *K-Fold Cross Validation* dan *Percetange Split*. Dari teknik pengujian tersebut diperoleh nilai rata-rata diatas 90%, baik menggunakan metode *Percetange Split* maupun metode *K-Fold Cross Validation*. Dari nilai rata-rata tersebut, dapat disimpulkan bahwa penggunaan teknik pengujian *K-Fold Cross Validation* dan *Percetange Split* pada algoritma *Support Vector Machine (SVM)* dan *Algoritma Logistic regression* bekerja dengan sangat baik. Namun, dalam hasil rata-rata nilai akurasi, presisi dan *recall* yang diperoleh dengan menggunakan metode *K-Fold Cross Validation*, algoritma SVM lebih baik dibandingkan dengan *Algoritma Logistic regression* dalam mengklasifikasi penyakit kanker payudara. Sehingga dapat disimpulkan bahwa algoritma *Suppport Vector Machine* (SVM) merupakan algoritma terbaik dalam mengklasifikasi penyakit kanker payudara dibandingkan dengan *Algoritma Logistic regression*.

DAFTAR PUSTAKA

[1] J. Ferlay *et al.*, "Global cancer observatory: cancer today," *Lyon Int. agency Res. cancer*, vol. 20182020, 2020.

[2] D. Alfiani, M. P. Putri, and W. Widayanti, "Literature Study: Obesitas Sebagai Faktor Risiko Pada Kanker Payudara Triple Negative," in *Bandung Conference Series: Medical Science*, 2022, pp. 326–329.

[3] S. M. Bachtiar, *Penurunan Intensitas Nyeri Pasien Kanker Payudara dengan Teknik Guided Imagery*. Penerbit NEM, 2022.

[4] G. A. T. Dewi and L. Y. Hendrati, "Breast Cancer Risk Analysis by the Use of Hormonal Contraceptives and Age of Menarche," *J. Berk. Epidemiol.*, vol. 3, no. 1, p. 12, 2016, doi: 10.20473/jbe.v3i12015.12-23.

[5] Kemenkes, "Kanker Payudara Paling Banyak di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan."

[6] R. Selvan, B. Pepin, C. Igel, G. Samuel, and E. B. Dam, "Equity through Access: A Case for Small-scale Deep Learning," *arXiv Prepr. arXiv2403.12562*, 2024.
[7] S. Syam *et al.*, *Data mining: Teori dan Penerapannya dalam Berbagai Bidang*. PT. Sonpedia Publishing Indonesia, 2024.
[8] M. M. Islam, A. Rahaman, and M. R. Islam, "Development of Smart Healthcare Monitoring System in IoT Environment," *SN Comput. Sci.*, vol. 1, pp. 1–11, 2020.
[9] Yuli Mardi, "*Data mining* : Klasifikasi Menggunakan Algoritma C4 . 5 *Data mining* merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika," *J. Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2019.
[10] F. Rahutomo, P. Y. Saputra, and M. A. Fidyawan, "Implementasi Twitter Sentiment Analysis Untuk Review Film Menggunakan Algoritma Support Vector Machine," *J. Inform. Polinema*, vol. 4, no. 2, p. 93, 2018, doi: 10.33795/jip.v4i2.152.
[11] D. Kurniawan and D. C. Supriyanto, "Optimasi Algoritma *Support Vector Machine (SVM)* Menggunakan Adaboost Untuk Penilaian Risiko Kredit," *J. Teknol. Inf.*, vol. 9, no. 1, pp. 1414–9999, 2013.
[12] H. Amalia, "Perbandingan Metode *Data mining* SVM Dan NN Untuk Klasifikasi Penyakit Ginjal Kronis," *J. Pilar Nusa Mandiri*, vol. 14, no. 1, p. 1, 2018.
[13] H. Apriyani and K. Kurniati, "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus," *J. Inf. Technol. Ampera*, vol. 1, no. 3, pp. 133–143, 2020, doi: 10.51519/journalita.volume1.isssue3.year2020.page133-143.
[14] Dian Prajarini, "Perbandingan Algoritma Klasifikasi *Data mining* Untuk Prediksi Penyakit Kulit," *Informatics J.*, vol. 15, no. 3, pp. 1–5, 2021.
[15] Trivusi, "Penjelasan Lengkap Algoritma *Support Vector Machine (SVM)*."
[16] AWS, "Apa itu regresi logistik?," 2023.
[17] A. P. Wicaksono, T. Badriyah, and A. Basuki, "*Data mining* Studi Perbandingan Prediksi Penyakit Diabetes dengan menggunakan *Logistic regression* dan Decision Trees," *J. semnaskit*, pp. 66–69, 2015.
[18] H. Kusuma, "Desentralisasi Fiskal dan Pertumbuhan Ekonomi di Indonesia," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2015.
[19] T. Zulhaq Jasman, E. Hasmin, C. Susanto, and W. Musu, "Perbandingan *Logistic regression*, Random Forest, dan Perceptron pada Klasifikasi Pasien Gagal Jantung," *CRSID J.*, vol. 14, no. 3, pp. 271–286, 2022.
[20] Y. Azhar, A. K. Firdausy, and P. J. Amelia, "Perbandingan Algoritma Klasifikasi *Data mining* Untuk Prediksi Penyakit Stroke," *SINTECH (Science Inf. Technol. J.*, vol. 5, no. 2, pp. 191–197, 2022, doi: 10.31598/sintechjournal.v5i2.1222.
[21] D. Ismafillah, T. Rohana, and Y. Cahyana, "Implementasi Model Support Vector Machine dan *Logistic regression* Untuk Memprediksi Penyakit Stroke," *JURIKOM (Jurnal Ris. Komputer)*, vol. 10, no. 1, pp. 2407–389, 2023, doi: 10.30865/jurikom.v10i1.5478.
[22] Adi Nugroho, Agustinus Bimo Gumelar, Adri Gabriel Sooai, Dyana Sarvasti, and Paul L Tahalele, "Perbandingan Performansi Kinerja Algoritma Pengklasifikasian Terpandu Untuk Kasus Penyakit Kardiovaskular," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 998–1006, 2020, doi: 10.29207/resti.v4i5.2316.
[23] B. S. Abunasser, M. R. J. AL-Hiealy, I. S. Zaqout, and S. S. Abu-Naser, "Literature review of breast cancer detection using machine learning algorithms," in *AIP Conference Proceedings*, AIP Publishing, 2023.
[24] D. M. U. Atmaja, A. R. Hakim, D. Haryadi, and N. Suwaryo, "Penerapan Algoritma K-Nearest Neighbor Untuk Prediksi Pengelompokkan Tingkat Risiko Penyebaran COVID-19 Jawa Barat," in *Prosiding Seminar Nasional Teknologi Energi dan Mineral*, 2021, pp. 1218–1226.
[25] K. Loizidou, R. Elia, and C. Pitris, "Computer-aided Breast Cancer Detection And Classification in Mammography: A Comprehensive Review," *Comput. Biol. Med.*, vol. 153, p. 106554, 2023.
[26] Trivusi, "Normalisasi Data: Pengertian, Tujuan, dan Metodenya."
[27] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.
[28] A. Rohman and M. Rochcham, "Komparasi Metode Klasifikasi *Data mining* Untuk Prediksi Kelulusan Mahasiswa," *J. Neo Tek*, vol. 5, no. 1, pp. 34–40, 2019, doi: 10.37760/neoteknika.v5i1.1379.
[29] M. R. A. Nasution and M. Hayaty, "Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter," *J. Inform.*, vol. 6, no. 2, pp. 226–235, 2019, doi: 10.31311/ji.v6i2.5129.
[30] D. Sartika and D. I. Sensuse, "Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian," *Jatisi*, vol. 1, no. 2, pp. 151–161, 2017.
[31] O. Arifin and T. B. Sasongko, "Analisa perbandingan tingkat performansi metode support vector machine dan naïve bayes classifier," *J. Semin. Nas. Teknol. Inf. dan Multimed. 2018*, vol. 6, no. 1, pp. 67–72, 2018.

[32] A. Pratama, R. C. Wihandika, and D. E. Ratnawati, "Implementasi Algoritme *Support Vector Machine (SVM)* untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa | Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 4, pp. 1704–1708, 2018.

[33] Shedriko, "Perbandingan Algoritma SVM dan KNN Dalam Mengklasifikasi Kelulusan Mahasiswa Pada Suatu Mata Kuliah," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 6, no. 2, pp. 115–122, 2021.

[34] N. Nurajijah and D. Riana, "Algoritma Naïve Bayes, Decision Tree, dan SVM untuk Klasifikasi Persetujuan Pembiayaan Nasabah Koperasi Syariah," *J. Teknol. dan Sist. Komput.*, vol. 7, no. 2, pp. 77–82, 2019, doi: 10.14710/jtsiskom.7.2.2019.77-82.

[35] S. Agustina, A. Agoestanto, and P. Hendikawati, "Klasifikasi Tingkat Kesejahteraan Keluarga Jawa Tengah Tahun 2015 Menggunakan Metode Regresi Logistik Ordinal," *Unnes J. Math.*, vol. 6, no. 1, pp. 59–69, 2017.

[36] J. Liang, "Confusion Matrix: Machine learning," *POGIL Act. Clear.*, vol. 3, no. 4, 2022.