

## OPTIMASI K-MEANS CLUSTERING PSO UNTUK PENENTUAN JUMLAH CLUSTER OPTIMAL PADA DATA KANKER PAYUDARA

Adhitya Purboyo<sup>a\*</sup>, Laksamana Rajendra Haidar Azani Fajri<sup>b</sup>, Imam Syafii<sup>c</sup>

<sup>a</sup> [adhityapurboyo@uinsuku.ac.id](mailto:adhityapurboyo@uinsuku.ac.id), Universitas Islam Negeri Sunan Kudus; Jalan Conge Ngembalrejo Kudus

<sup>b</sup> [laksamanarajendra@uinsuku.ac.id](mailto:laksamanarajendra@uinsuku.ac.id), Universitas Islam Negeri Sunan Kudus; Jalan Conge Ngembalrejo Kudus

<sup>c</sup> [imamsyafii@uinsuku.ac.id](mailto:imamsyafii@uinsuku.ac.id), Universitas Islam Negeri Sunan Kudus; Jalan Conge Ngembalrejo Kudus

\* Penulis Korespondensi: Adhitya Purboyo

### ABSTRACT

Breast cancer is one of the most dangerous diseases and a leading cause of death among women worldwide. Clustering methods can assist in diagnosing breast cancer to determine the best course of treatment. K-Means is a widely used clustering algorithm known for its ability to handle large datasets efficiently with fast computational time. However, K-Means has a significant weakness: the number of clusters is determined randomly, resulting in suboptimal clustering outcomes. To overcome this limitation, Particle Swarm Optimization (PSO) is applied for automatic determination of the optimal number of clusters. PSO was selected due to its advantages, including requiring few parameters, ease of implementation, fast convergence, and low computational cost. This study uses the breast cancer dataset from the UCI Machine Learning Repository, consisting of 699 records and 10 attributes. The proposed PSO–K-Means method was evaluated using the Silhouette Coefficient and Davies-Bouldin Index. The results show that the optimal number of clusters is  $k = 2$ , achieving a Silhouette Coefficient of 0.92 and a Davies-Bouldin Index of 1.374. These results demonstrate that the PSO–K-Means method significantly outperforms standard K-Means by directly producing optimal clustering results without the need for conducting repeated experiments.

**Keywords:** K-Means, Particle Swarm Optimization, Clustering, Breast Cancer

### Abstrak

Kanker payudara merupakan salah satu penyakit berbahaya dan menjadi penyebab utama kematian pada wanita di seluruh dunia. Metode clustering dapat membantu dalam mendiagnosis kanker payudara sehingga tindakan terbaik dapat ditentukan. K-Means merupakan algoritma clustering yang banyak digunakan karena kemampuannya dalam mengelompokkan data besar dengan waktu komputasi yang cepat dan efisien. Namun, K-Means memiliki kelemahan utama yaitu jumlah cluster ditentukan secara acak sehingga hasil clustering menjadi kurang optimal. Untuk mengatasi kelemahan tersebut, Particle Swarm Optimization (PSO) diterapkan untuk menentukan jumlah cluster yang optimal secara otomatis. PSO dipilih karena memiliki beberapa kelebihan yaitu membutuhkan beberapa parameter, mudah diimplementasikan, konvergensi yang cepat, dan komputasi yang efisien. Penelitian ini menggunakan dataset breast cancer dari UCI Machine Learning Repository yang terdiri dari 699 record dan 10 atribut. Metode PSO–K-Means yang diusulkan dievaluasi menggunakan Silhouette Coefficient dan Davies-Bouldin Index. Hasil pengujian menunjukkan bahwa jumlah cluster optimal adalah  $k = 2$  dengan nilai Silhouette Coefficient sebesar 0.92 dan nilai Davies-Bouldin Index sebesar 1.374. Hasil tersebut membuktikan bahwa metode PSO–K-Means lebih unggul dibandingkan K-Means standar karena dapat menghasilkan cluster yang optimal tanpa perlu melakukan percobaan berulang kali.

**Kata Kunci:** K-Means, Particle Swarm Optimization, Clustering, Kanker Payudara.

## 1. PENDAHULUAN

Kanker merupakan salah satu penyakit yang paling berbahaya dan menjadi ancaman serius bagi kesehatan masyarakat di seluruh dunia. Kanker terjadi akibat pertumbuhan sel abnormal yang merusak jaringan di sekitarnya dan dapat menyebar ke bagian tubuh lain yang letaknya jauh. Pertumbuhan sel yang tidak terkendali ini bersifat parasit dan berkembang dengan mengorbankan tubuh manusia sebagai inangnya. Salah satu jenis kanker yang paling banyak menyerang adalah kanker payudara atau breast cancer. Kanker payudara merupakan kanker yang paling sering ditemukan pada wanita, dengan sekitar 16% dari seluruh kasus kanker pada wanita adalah kanker payudara. Pada tahun 2004, diperkirakan sebanyak 519000 perempuan meninggal dunia akibat penyakit ini. Meskipun kanker payudara sering dikaitkan dengan negara maju, namun sebagian besar kematian akibat kanker payudara justru terjadi di negara berkembang [1], [2].

Permasalahan utama dalam penanganan kanker payudara adalah kurangnya kesadaran masyarakat untuk melakukan pemeriksaan sejak dini. Pada tahap awal, kanker payudara sebenarnya sudah menunjukkan tanda-tanda dan gejala, namun sering kali tidak disadari sehingga baru diketahui setelah memasuki stadium lanjut. Pada stadium lanjut, kanker payudara telah menyebar ke organ tubuh lainnya seperti paru-paru, hati, tulang, bahkan otak. Kondisi ini menyebabkan angka kematian akibat kanker payudara menjadi sangat tinggi dan pengobatan yang dilakukan hanya bertujuan untuk meringankan gejala serta memperpanjang usia pasien. Oleh karena itu, pemeriksaan dini sangat diperlukan sebagai upaya pencegahan dan untuk mengetahui tingkat stadium kanker payudara sedini mungkin [3].

Salah satu pendekatan yang dapat digunakan untuk membantu diagnosis kanker payudara adalah dengan menerapkan teknik clustering dalam data mining. Clustering merupakan metode pengelompokan data yang memiliki kemiripan satu sama lain ke dalam kelompok-kelompok tertentu tanpa data latih. Dengan clustering, data pasien kanker payudara dapat dikelompokkan berdasarkan karakteristik yang serupa sehingga dapat membantu dalam menentukan tindakan medis terbaik yang perlu dilakukan [2], [4].

Di antara berbagai algoritma clustering yang tersedia seperti Hierarchical Clustering dan Fuzzy C-Means, algoritma K-Means merupakan metode yang paling populer dan banyak digunakan. K-Means dipilih karena kemampuannya dalam mengelompokkan data dalam jumlah besar dengan waktu komputasi yang cepat dan efisien, mudah diimplementasikan, sangat fleksibel, serta menggunakan prinsip yang sederhana [4], [5]. Namun demikian, K-Means memiliki kelemahan yang cukup mendasar, yaitu jumlah cluster ditentukan secara acak sehingga hasil clustering yang dihasilkan sering kali tidak optimal. Sulitnya menentukan jumlah cluster yang tepat dalam sebuah dataset secara akurat menjadi kendala utama yang dihadapi pengguna dalam menggunakan algoritma K-Means [6], [7].

Untuk mengatasi kelemahan tersebut, diperlukan sebuah metode optimasi yang mampu menentukan jumlah cluster secara otomatis dan akurat. Terdapat beberapa metode optimasi yang dapat digunakan untuk menentukan jumlah cluster, antara lain Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Elbow, Bootstrap, Cross-Validation, dan Gap Statistic. Dari berbagai metode tersebut, PSO dipilih karena memiliki beberapa keunggulan dibandingkan metode lainnya, yaitu hanya membutuhkan sedikit parameter, mudah diterapkan, memiliki konvergensi yang cepat, lebih efisien karena membutuhkan sedikit komputasi, dan bersifat sederhana [8], [9]. PSO merupakan algoritma evolusi yang terinspirasi dari perilaku kawanan burung dan kawanan ikan dalam mencari sumber makanan. Metode PSO akan sangat membantu proses clustering dengan melakukan optimasi dalam menentukan jumlah cluster yang tepat sehingga dapat meningkatkan akurasi dan kualitas hasil cluster yang dihasilkan [10], [11].

Berdasarkan latar belakang tersebut, penelitian ini mengusulkan penerapan metode PSO-K-Means, yaitu kombinasi antara algoritma K-Means dengan Particle Swarm Optimization untuk menentukan jumlah cluster yang optimal pada dataset breast cancer. Tujuan penelitian ini adalah untuk membuktikan bahwa metode PSO-K-Means mampu menghasilkan clustering yang lebih baik dan lebih optimal dibandingkan dengan metode K-Means standar, tanpa harus melakukan percobaan berulang kali dalam menentukan jumlah cluster yang tepat [12], [13]. Evaluasi hasil clustering dilakukan menggunakan Silhouette Coefficient dan Davies-Bouldin Index sebagai ukuran kualitas cluster yang dihasilkan [14].

## 2. TINJAUAN PUSTAKA

### 2.1. Kanker Payudara

Kanker payudara adalah jenis kanker yang paling umum pada wanita dan menjadi salah satu penyebab utama kematian akibat kanker di seluruh dunia. Kanker terjadi akibat pertumbuhan sel abnormal yang tidak terkendali dan dapat menyebar ke bagian tubuh lain. Sekitar 16% dari seluruh kasus kanker yang tercatat pada wanita adalah kanker payudara [1]. Deteksi dini sangat penting karena pada stadium awal kanker payudara masih dapat ditangani dengan efektif, sedangkan pada stadium lanjut kanker telah menyebar ke organ vital sehingga penanganannya jauh lebih sulit [3].

## 2.2 Clustering

Clustering adalah teknik dalam data mining yang mengelompokkan data berdasarkan kemiripan karakteristiknya tanpa memerlukan data latih. Data yang memiliki kemiripan akan masuk ke dalam satu kelompok yang sama, sedangkan data yang berbeda dipisahkan ke kelompok lain. Teknik ini telah banyak digunakan dalam bidang kesehatan, salah satunya untuk membantu analisis data kanker payudara [2], [4].

## 2.3 Algoritma K-Means

K-Means adalah algoritma clustering yang paling banyak digunakan karena sederhana, mudah diimplementasikan, dan mampu memproses data dalam jumlah besar secara efisien [4], [5]. Algoritma ini bekerja dengan cara mengelompokkan data ke dalam k cluster berdasarkan jarak terdekat antara data dengan centroid, kemudian centroid diperbarui secara iteratif hingga tidak ada perubahan. Perhitungan jarak menggunakan metode Euclidean Distance, sedangkan centroid awal ditentukan dengan rumus jumlah data dibagi jumlah kelas ditambah satu [6], [15]. Kelemahan utama K-Means adalah jumlah cluster harus ditentukan secara acak oleh pengguna sebelum proses dimulai, sehingga hasil clustering sering tidak optimal [6], [7].

## 2.4 Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) adalah algoritma optimasi yang terinspirasi dari perilaku kawanan burung dalam mencari sumber makanan. Setiap partikel bergerak dalam ruang pencarian menuju posisi terbaik dirinya masing-masing (Personal Best/Pbest) dan posisi terbaik kelompok (Global Best/Gbest) melalui pembaruan kecepatan dan posisi pada setiap iterasi [9], [16]. PSO dipilih karena hanya membutuhkan sedikit parameter, mudah diterapkan, konvergensi cepat, dan efisien secara komputasi [8]. Kemampuan PSO dalam mencari solusi optimal menjadikannya cocok untuk menentukan jumlah cluster yang tepat secara otomatis [10], [11].

## 2.5 Evaluasi Hasil Clustering

Kualitas hasil clustering diukur menggunakan metode evaluasi internal. Dalam penelitian ini digunakan dua metode evaluasi, yaitu Silhouette Coefficient dan Davies-Bouldin Index (DBI). Silhouette Coefficient mengukur seberapa baik sebuah objek ditempatkan dalam clusternya dengan nilai antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan kualitas cluster yang baik. Sebaliknya, DBI mengukur kualitas cluster berdasarkan kohesi dan separasi, di mana semakin kecil nilai DBI maka semakin baik kualitas cluster yang dihasilkan [14], [7].

## 3 METODOLOGI PENELITIAN

### 3.1 Dataset

Penelitian ini menggunakan dataset publik breast cancer yang bersumber dari UCI Machine Learning Repository (<https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original>). Dataset tersebut dipilih karena telah banyak digunakan sebagai benchmark dalam penelitian di bidang data mining dan machine learning. Karakteristik dataset yang digunakan disajikan pada Tabel 1.

Tabel 1. Dataset breast cancer

Dataset	Atribut Numerik	Jumlah Record	Missing Value
Breast Cancer	10	699	16

Dataset breast cancer terdiri atas 699 record dengan 10 atribut numerik dan memiliki 16 missing value. Atribut yang digunakan dalam penelitian ini meliputi Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, dan Mitoses. Atribut Sample Code Number tidak digunakan dari proses analisis karena tidak memiliki relevansi terhadap proses clustering.

### 3.2 Preprocessing Data

Sebelum data diproses lebih lanjut, dilakukan tahap preprocessing untuk memastikan kualitas data yang digunakan. Tahap ini mencakup dua proses utama, yaitu penanganan missing value terhadap 16 record yang tidak lengkap, dan penghapusan atribut Sample Code Number yang tidak relevan terhadap analisis clustering. Preprocessing bertujuan menghasilkan data yang bersih, lengkap, dan siap digunakan, sehingga dapat meningkatkan akurasi dan reliabilitas hasil clustering.

### 3.3 Tahapan Penelitian

Penelitian ini dilaksanakan melalui empat tahapan utama yang saling berkesinambungan, yaitu pengumpulan data, preprocessing data, penentuan jumlah cluster menggunakan PSO, dan proses clustering menggunakan algoritma K-Means.

#### 3.3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari UCI Machine Learning Repository, yaitu dataset breast cancer yang merupakan dataset publik dan telah banyak digunakan sebagai standar pengujian dalam penelitian data mining dan machine learning.

#### 3.3.2 Preprocessing Data

Tahap preprocessing dilakukan untuk mempersiapkan data sebelum masuk ke dalam proses clustering. Proses ini meliputi penanganan missing value pada 16 record yang tidak lengkap serta penghapusan atribut yang tidak relevan. Preprocessing bertujuan memastikan data yang digunakan berkualitas sehingga menghasilkan output clustering yang lebih optimal.

#### 3.3.3 Penentuan Jumlah Cluster dengan PSO

Berbeda dengan metode K-Means konvensional yang menentukan jumlah cluster (k) secara manual, penelitian ini menggunakan PSO untuk menentukan nilai k secara otomatis dan optimal. Adapun prosedur PSO yang diterapkan adalah sebagai berikut.

Langkah 1: Inisialisasi jumlah partikel serta posisi awal setiap partikel secara acak dalam ruang pencarian.

Langkah 2: Inisialisasi kecepatan awal setiap partikel dengan nilai nol (0).

Langkah 3: Hitung nilai fitness dari setiap partikel, kemudian tetapkan nilai Personal Best (Pbest) dan Global Best (Gbest).

Langkah 4: Perbarui kecepatan setiap partikel menggunakan persamaan berikut [9], [16]:

$$v_{k+1}^i = v_k^i + c_1 r_1 (p_k^i - x_k^i) + c_2 r_2 (p_k^g - x_k^i) \quad (1)$$

Langkah 5: Perbarui posisi setiap partikel berdasarkan kecepatan yang telah diperbarui:

$$x_{k+1}^i = x_k^i + v_{k+1}^i \quad (2)$$

Langkah 6: Hitung nilai fitness terbaru, kemudian perbarui nilai Pbest dan Gbest. Nilai Pbest diperbarui apabila nilai fitness posisi terkini lebih baik dari nilai Pbest sebelumnya. Nilai terbaik di antara seluruh Pbest ditetapkan sebagai Gbest baru.

Langkah 7: Ulangi Langkah 4 hingga Langkah 6 hingga kondisi konvergensi terpenuhi, sehingga dihasilkan nilai k optimal untuk digunakan pada proses clustering K-Means.

#### 3.3.4 Proses Clustering dengan Algoritma K-Means

Setelah nilai k optimal diperoleh dari tahap PSO, proses clustering dilaksanakan menggunakan algoritma K-Means dengan prosedur sebagai berikut.

Langkah 1: Tetapkan jumlah cluster (k) sesuai nilai optimal yang dihasilkan oleh PSO.

Langkah 2: Tentukan pusat cluster (centroid) awal menggunakan rumus berikut [15]:

$$\text{Centroid} = \frac{\text{JumlahData}}{\text{Jumlah Class}+1} \quad (3)$$

Langkah 3: Hitung jarak setiap data terhadap masing-masing centroid menggunakan Euclidean Distance [6]:

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (4)$$

Langkah 4: Tetapkan setiap objek data ke dalam cluster dengan jarak terdekat terhadap centroid.

Langkah 5: Perbarui nilai centroid dengan menghitung rata-rata seluruh anggota dalam masing-masing cluster:

$$V_{ij} = \sum_{k=0}^{N_i} X_{kj} \quad (5)$$

Langkah 6: Ulangi Langkah 3 hingga Langkah 5 sampai nilai centroid tidak mengalami perubahan atau mencapai kondisi konvergen.

### 3.4 Evaluasi Hasil Clustering

Evaluasi dilakukan untuk mengukur kualitas hasil clustering yang dihasilkan oleh metode PSO–K-Means. Penelitian ini menggunakan dua metode evaluasi internal, yaitu Silhouette Coefficient dan Davies-Bouldin Index (DBI).

#### 3.4.1 Silhouette Coefficient

Silhouette Coefficient merupakan metode evaluasi internal yang mengukur seberapa tepat suatu objek ditempatkan dalam clusternya. Metode ini menggabungkan ukuran cohesion dan separation. Nilai Silhouette Coefficient berkisar antara  $-1$  hingga  $1$ ; semakin mendekati nilai  $1$ , semakin baik kualitas cluster yang dihasilkan [14].

Hitung jarak rata-rata objek  $i$  terhadap seluruh objek dalam cluster yang sama ( $A$ ):

$$a(i) = \frac{1}{|A|} \sum_{j \in A, j \neq i} d(i, j) \quad (6)$$

Hitung jarak rata-rata objek  $i$  terhadap seluruh objek pada setiap cluster lain, kemudian ambil nilai minimum:

$$b(i) = \min_{C \neq A} d(i, C) \quad (7)$$

Hitung nilai Silhouette Coefficient untuk setiap objek:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (8)$$

#### 3.4.2 Davies-Bouldin Index (DBI)

Davies-Bouldin Index (DBI) merupakan metode evaluasi internal yang mengukur kualitas cluster berdasarkan nilai kohesi dan separasi. Kohesi mengacu pada tingkat kedekatan data terhadap centroid clusternya, sedangkan separasi mengacu pada jarak antar centroid cluster. Semakin kecil nilai DBI yang diperoleh, semakin baik kualitas cluster yang dihasilkan [7], [14]. Nilai DBI dihitung menggunakan persamaan berikut.

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} (R_{i,j}) \quad (9)$$

dengan nilai rasio  $R_{ij}$  dihitung sebagai berikut:

$$R_{ij} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (10)$$

## 4 HASIL DAN PEMBAHASAN

### 4.1 Hasil Preprocessing Data

Sebelum dilakukan proses clustering, dataset breast cancer terlebih dahulu melalui tahap preprocessing. Dataset yang digunakan terdiri dari 699 record dengan 10 atribut numerik. Dari hasil identifikasi data ditemukan sebanyak 16 missing value yang perlu ditangani agar tidak mempengaruhi kualitas hasil clustering. Selain itu, atribut Sample Code Number juga dihapus karena tidak relevan terhadap proses analisis clustering. Setelah proses preprocessing selesai dilakukan, data yang telah bersih dan lengkap telah siap digunakan untuk proses selanjutnya.

### 4.2 Hasil Penentuan Jumlah Cluster dengan PSO

Pada penelitian ini, penentuan jumlah cluster tidak dilakukan secara acak seperti pada metode K-Means standar, melainkan ditentukan secara otomatis menggunakan metode Particle Swarm Optimization (PSO).

Proses PSO dilakukan dengan menginisialisasi posisi dan kecepatan awal setiap partikel, kemudian menghitung nilai fitness menggunakan Silhouette Coefficient pada setiap iterasi hingga kondisi konvergen terpenuhi.

Hasil dari proses optimasi PSO menunjukkan bahwa jumlah cluster optimal yang diperoleh secara otomatis adalah  $k = 2$ . Nilai  $k = 2$  ini kemudian digunakan sebagai input pada proses clustering K-Means untuk menghasilkan pengelompokan data yang optimal. PSO mampu secara otomatis menemukan jumlah cluster yang optimal tanpa memerlukan percobaan berulang dengan nilai  $k$  yang berbeda-beda seperti yang biasa dilakukan pada metode K-Means standar.

#### 4.3 Hasil Clustering PSO–K-Means

Setelah jumlah cluster optimal diperoleh dari proses PSO, selanjutnya dilakukan proses clustering menggunakan algoritma K-Means dengan  $k = 2$ . Percobaan dilakukan sebanyak 2 kali dengan pusat cluster yang berbeda untuk memvalidasi konsistensi hasil yang diperoleh. Hasil clustering kemudian dievaluasi menggunakan Silhouette Coefficient dan Davies-Bouldin Index.

#### 4.4 Evaluasi Silhouette Coefficient

Evaluasi pertama dilakukan menggunakan Silhouette Coefficient untuk mengukur seberapa baik pengelompokan data yang dihasilkan. Hasil perbandingan nilai Silhouette Coefficient antara metode K-Means standar dengan metode PSO–K-Means pada berbagai jumlah cluster ditunjukkan pada Tabel 2.

Tabel 2. Perbandingan Nilai Silhouette Coefficient

Jumlah Cluster ( k )	Percobaan 1	Percobaan 2
k = 2 (PSO–K-Means)	0.822	0.920
k = 3 (K-Means)	0.739	0.822
k = 4 (K-Means)	0.621	0.406

Berdasarkan Tabel 2 dapat diketahui bahwa metode PSO–K-Means dengan jumlah cluster  $k = 2$  menghasilkan nilai Silhouette Coefficient tertinggi yaitu sebesar 0.822 pada percobaan pertama dan 0.920 pada percobaan kedua. Nilai ini jauh lebih tinggi dibandingkan dengan metode K-Means standar yang menggunakan  $k = 3$  dengan nilai Silhouette Coefficient sebesar 0.739 dan 0.822, maupun K-Means dengan  $k = 4$  yang menghasilkan nilai Silhouette Coefficient sebesar 0.621 dan 0.406.

Hasil tersebut menunjukkan bahwa semakin bertambahnya jumlah cluster pada metode K-Means standar, nilai Silhouette Coefficient cenderung semakin menurun dan semakin rendah mendekati nilai negatif. Hal ini membuktikan bahwa penentuan jumlah cluster secara acak pada K-Means standar menghasilkan kualitas clustering yang tidak konsisten dan tidak optimal. Sebaliknya, metode PSO–K-Means mampu secara menemukan jumlah cluster yang optimal sehingga menghasilkan nilai Silhouette Coefficient yang paling mendekati 1, yang berarti kualitas pengelompokan data yang dihasilkan adalah yang terbaik.

#### 4.5 Evaluasi Davies-Bouldin Index

Evaluasi kedua dilakukan menggunakan Davies-Bouldin Index (DBI) untuk mengukur kualitas cluster berdasarkan nilai kohesi dan separasi. Berbeda dengan Silhouette Coefficient, pada DBI semakin kecil nilai yang diperoleh maka semakin baik kualitas cluster yang dihasilkan. Hasil perbandingan nilai DBI antara metode K-Means standar dengan metode PSO–K-Means pada berbagai jumlah cluster ditunjukkan pada Tabel 3.

Tabel 3. Perbandingan Nilai Davies-Bouldin Index

Jumlah Cluster ( k )	Percobaan 1	Percobaan 2
k = 2 (PSO–K-Means)	1.375	1.374
k = 3 (K-Means)	2.349	1.470
k = 4 (K-Means)	3.800	2.035

Berdasarkan Tabel 3 dapat diketahui bahwa metode PSO–K-Means dengan jumlah cluster  $k = 2$  menghasilkan nilai Davies-Bouldin Index terendah yaitu sebesar 1.375 pada percobaan pertama dan 1.374 pada percobaan kedua. Nilai ini jauh lebih rendah dibandingkan dengan metode K-Means standar yang menggunakan  $k = 3$  dengan nilai DBI sebesar 2.349 dan 1.470, maupun K-Means dengan  $k = 4$  yang menghasilkan nilai DBI sebesar 3.800 dan 2.035.

Hasil tersebut menunjukkan bahwa semakin bertambahnya jumlah cluster pada metode K-Means standar, nilai Davies-Bouldin Index cenderung semakin meningkat. Hal ini berarti kualitas cluster yang dihasilkan semakin buruk karena jarak antar data dalam satu cluster semakin besar dan pemisahan antar cluster semakin tidak jelas. Sebaliknya, metode PSO–K-Means secara konsisten menghasilkan nilai DBI yang paling rendah pada kedua percobaan, yang membuktikan bahwa cluster yang dihasilkan memiliki kohesi yang tinggi dan separasi antar cluster yang baik.

#### 4.6 Pembahasan

Berdasarkan hasil evaluasi menggunakan Silhouette Coefficient dan Davies-Bouldin Index, dapat dilakukan perbandingan secara menyeluruh antara metode K-Means standar dengan metode PSO–K-Means yang diusulkan. Rekapitulasi hasil perbandingan kedua metode tersebut disajikan pada Tabel 4.

Tabel 4. Perbandingan Metode K-Means dan PSO–K-Means

Metode	Jumlah Cluster ( k )	Silhouette Coefficient	DBI
k = 2 (PSO–K-Means)	k = 2	0.920	1.374
k = 3 (K-Means)	k = 3	0.822	1.470
k = 4 (K-Means)	k = 4	0.621	2.035

Dari Tabel 4 terlihat dengan jelas bahwa metode PSO–K-Means unggul secara signifikan dibandingkan metode K-Means standar, baik dari segi nilai Silhouette Coefficient maupun Davies-Bouldin Index. Metode PSO–K-Means menghasilkan nilai Silhouette Coefficient tertinggi sebesar 0.920 yang mendekati nilai ideal 1, sekaligus menghasilkan nilai Davies-Bouldin Index terendah sebesar 1.374 yang mendekati nilai ideal 0. Kedua hasil evaluasi ini secara konsisten menunjukkan bahwa metode PSO–K-Means menghasilkan cluster yang lebih baik, lebih kompak, dan lebih terpisah antar clusternya dibandingkan dengan metode K-Means standar.

Hasil penelitian ini membuktikan bahwa penerapan Particle Swarm Optimization untuk mengoptimalkan jumlah cluster pada algoritma K-Means terbukti efektif dalam meningkatkan kualitas hasil clustering pada dataset breast cancer. Keunggulan metode PSO–K-Means dapat dijelaskan dari beberapa aspek. Pertama, dari aspek penentuan jumlah cluster, metode K-Means standar menentukan jumlah cluster secara acak sehingga memerlukan banyak percobaan untuk menemukan nilai k yang optimal. Sebaliknya, PSO mampu secara otomatis menemukan jumlah cluster yang paling optimal tanpa perlu melakukan percobaan berulang kali [12], [13]. Kedua, dari aspek kualitas clustering, metode PSO–K-Means menghasilkan cluster dengan nilai Silhouette Coefficient yang lebih tinggi dan nilai Davies-Bouldin Index yang lebih rendah dibandingkan K-Means standar [14], [7]. Ketiga, dari aspek efisiensi komputasi, PSO memiliki keunggulan dalam hal konvergensi yang cepat dan hanya membutuhkan sedikit parameter sehingga proses optimasi dapat dilakukan secara efisien [8], [9].

Namun demikian, penelitian ini juga menemukan bahwa metode PSO–K-Means masih memiliki kelemahan yaitu hasil clustering dapat menjadi kurang optimal apabila penentuan pusat cluster awal masih dilakukan secara acak. Hal ini dikarenakan PSO hanya melakukan optimasi pada jumlah cluster saja, sementara penentuan pusat cluster awal pada K-Means masih dilakukan secara random. Oleh karena itu, untuk penelitian selanjutnya perlu dipertimbangkan pengembangan metode yang juga mengoptimalkan penentuan pusat cluster awal sehingga dapat menghasilkan clustering yang lebih optimal secara keseluruhan.

## 5 KESIMPULAN DAN SARAN

Penelitian ini membuktikan bahwa PSO telah terbukti efektif mengatasi kelemahan utama K-Means dalam menentukan jumlah cluster, yang selama ini dilakukan secara acak. Dengan PSO, jumlah cluster optimal dapat ditentukan secara otomatis tanpa percobaan berulang. Hasilnya menunjukkan bahwa k = 2 adalah jumlah cluster optimal untuk dataset breast cancer, dengan Silhouette Coefficient 0.92 dan Davies-Bouldin Index 1.374, jauh lebih baik dibandingkan dengan K-Means standar yang kualitasnya cenderung menurun seiring bertambahnya jumlah cluster. Meski begitu, metode ini masih memiliki keterbatasan karena PSO hanya mengoptimalkan jumlah cluster, sementara penentuan pusat cluster awal masih dilakukan secara acak.

Pengembangan PSO–K-Means ke depannya sebaiknya diarahkan agar tidak hanya mengoptimalkan jumlah cluster, tetapi juga penentuan pusat cluster awal secara bersamaan. Metode ini selanjutnya perlu diuji pada dataset yang lebih beragam, baik dari sisi tipe maupun karakteristik data, serta dibandingkan dengan metode optimasi lain seperti Genetic Algorithm atau Simulated Annealing. Pada akhirnya, metode ini berpotensi

diintegrasikan ke dalam sistem pendukung keputusan klinis untuk membantu tenaga medis dalam mendiagnosis kanker payudara secara lebih cepat dan akurat.

#### DAFTAR PUSTAKA

- [1] S. Aamir *et al.*, “Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques,” *Comput. Math. Methods Med.*, vol. 2022, no. 1, 2022, doi 10.1155/2022/5869529
- [2] S. Salloum, “K-Means clustering and classification of breast cancer images using histogram of oriented gradients features and convolutional neural network models,” *JMIR Med. Informatics*, vol. 3, p. e71974, 2025, doi: 10.2196/71974
- [3] F. J. Fernández-Ovies, E. S. Alférez-García, E. J. De Andrés-Galiana, A. Cernea, Z. Fernández-Muñiz, and J. Klinger, “Detection of breast cancer using infrared thermography and unsupervised clustering algorithms,” in *Lecture Notes in Bioengineering*, 2023, pp. 3–12. doi: 10.1007/978-3-031-17024-9\_1
- [4] S. M. Miraftebzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, “K-Means and alternative clustering methods in modern power systems,” *IEEE Access*, vol. 11, pp. 119596–119633, 2023, doi: 10.1109/ACCESS.2023.3327640
- [5] A. Zhu, Z. Hua, Y. Shi, Y. Tang, and L. Miao, “An improved K-Means algorithm based on evidence distance,” *Entropy*, vol. 23, no. 11, p. 1550, 2021, doi: 10.3390/e23111550
- [6] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, “Improved the performance of the K-Means cluster using the sum of squared error (SSE) optimized by using the elbow method,” *J. Phys. Conf. Ser.*, vol. 1361, no. 1, p. 012015, 2019, doi: 10.1088/1742-6596/1361/1/012015
- [7] C. S. Wahono and N. Suryana, “Optimasi penentuan sentroid awal pada K-Means untuk meningkatkan hasil evaluasi Davies-Bouldin Index,” *J. Inform. Teknol. dan Sains*, vol. 5, no. 1, pp. 112–119, 2023, doi: 10.51401/jinteks.v5i1.3873
- [8] H. Harliana, M. H. Bhakti, O. S. Bachri, and F. . Efendi, “Optimasi K-Means dengan Particle Swarm Optimization pada pengelompokan daerah stunting,” *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 3, no. 02, pp. 95–101, 2021, doi: 10.46772/intech.v3i02.457
- [9] H. Hayashi, Y. Endo, and S. Miyamoto, “A new PSO-based clustering algorithm with adaptive parameter tuning for high dimensional data,” *J. Adv. Comput. Intell. Intell. Informatics*, vol. 25, no. 5, pp. 756–764, 2021, doi: 10.20965/jaciii.2021.p0756
- [10] A. Bello, S. . Ng, and M. F. Leung, “A PCA-based residual clustering analysis for hybrid PSO-K-Means optimization,” *Appl. Sci.*, vol. 13, no. 3, pp. 1–18, 2023, doi: 10.3390/app13031681
- [11] Q. Pu, J. Gan, L. Qiu, Z. Cai, and R. Li, “An efficient hybrid approach based on PSO, ABC and K-Means for cluster analysis,” *Appl. Multimed. Tools*, vol. 81, pp. 19321–19339, 2022, doi: 10.1007/s11042-021-11016-6
- [12] D. Surya, R. Hidayat, A. Ramadhan, and R. Putra, “Optimasi K-Means untuk pengelompokan jumlah sekolah di Riau menggunakan Particle Swarm Optimization,” *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 3, no. 1, pp. 37–46, 2024.
- [13] J. Ortiz-Bejar, J. Zarate-Orihuela, J. Tellez-Velazquez, and M. V. Chávez-Báez, “SA-PSO-GK++: A new hybrid clustering approach for analyzing medical data,” *IEEE Access*, vol. 12, pp. 1–15, 2024, doi: 10.1109/ACCESS.2024.3351719
- [14] O. D. Nurhayati and R. Permata, “Analysis of elbow, silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index evaluation on K-Means algorithm for classifying flood-affected areas in Jakarta,” *JAIC J. Artif. Intell. Capsul. Networks*, vol. 7, no. 1, pp. 95–103, 2023, doi: 10.36548/jaicn.2023.1.008
- [15] I. Arfiani, H. Yuliansyah, and M. D. Suratin, “Implementasi bee colony optimization pada pemilihan centroid dalam algoritma K-Means,” *Build. Informatics, Technol. Sci.*, vol. 3, no. 4, pp. 756–763, 2022, doi: 10.47065/bits.v3i4.1446
- [16] H. Yue, H. Zhang, and Y. Dai, “Application of PSO-integrated K-Means algorithm in resident digital portrait classification,” *PLoS One*, vol. 20, no. 8, p. e0329123, 2025, doi: 10.1371/journal.pone.0329123