

# Implementasi Seleksi Fitur *Binary Particle Swarm Optimization* pada Algoritma K-NN untuk Klasifikasi Kanker Payudara

Rahmat Hidayat<sup>a1</sup>, Dwi Kartini<sup>a2</sup>, Muhammad Itqan Mazdadi<sup>a3</sup>, Irwan Budiman<sup>a4</sup>, Rahmat Ramadhani<sup>a5</sup>

<sup>a</sup>Universitas Lambung Mangkurat

Jl. Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714

<sup>1</sup>1711016110017@mhs.ulm.ac.id

<sup>2</sup>dwikartini@ulm.ac.id

<sup>3</sup>mazdadi@ulm.ac.id

<sup>4</sup>irwan.budiman@ulm.ac.id

<sup>5</sup>rahmat.ramadhani@ulm.ac.id

## Abstrak

Kanker Payudara adalah jenis kanker paling umum yang sering menyerang kalangan wanita di seluruh dunia. Diagnosa awal yang akurat dalam mendeteksi kanker payudara memainkan peran penting dalam pengobatan pasien karena semakin cepat kanker di diagnosa semakin cepat juga pengobatan dapat diberikan. Untuk menghasilkan diagnosa yang akurat terhadap pasien kanker payudara maka dilakukan penelitian dengan tujuan mendapatkan model klasifikasi yang dapat memberikan klasifikasi yang akurat terhadap penyakit kanker payudara. Algoritma klasifikasi yang sering digunakan dan cukup terkenal adalah *K-Nearest Neighbor* (K-NN). Algoritma K-NN melakukan klasifikasi menggunakan konsep tetangga terdekat berdasarkan pada data terdahulu, akan tetapi algoritma K-NN lemah terhadap data dengan jumlah fitur yang besar. Maka dari itu, untuk kasus dataset dengan banyak fitur dapat dilakukan seleksi fitur terlebih dahulu untuk meningkatkan performa klasifikasi K-NN. Salah satu yang sering digunakan untuk seleksi fitur adalah algoritma *Binary Particle Swarm Optimization* (BPSO). Pada penelitian ini akan dibuat 2 model klasifikasi K-NN yaitu model klasifikasi K-NN saja tanpa seleksi fitur dan model klasifikasi K-NN dengan seleksi fitur *Binary Particle Swarm Optimization*, kemudian dilakukan perbandingan hasil akurasi yang didapat. Dataset yang akan digunakan adalah *Breast Cancer Wincosin (Diagnostic)* dari *UCI Machine Learning Repository* yang memiliki 569 data dan 30 fitur. Hasil penelitian menunjukkan model K-NN+BPSO menghasilkan akurasi sebesar 95,32% dan model K-NN menghasilkan akurasi sebesar 94,15%. Berdasarkan akurasi yang didapatkan algoritma K-NN dengan seleksi fitur menghasilkan akurasi yang lebih baik daripada algoritma K-NN tanpa seleksi fitur sebesar 1,17%. Algoritma *Binary PSO* juga berhasil mereduksi fitur dari 30 fitur menjadi 5 fitur dengan tidak mengurangi akurasi dari model klasifikasi. Sehingga dapat disimpulkan bahwa pada kasus klasifikasi kanker payudara dengan dataset *Breast Cancer Wincosin (Diagnostic)* algoritma K-NN dapat dikombinasikan dengan seleksi fitur *Binary PSO* untuk membuat model klasifikasi yang memberikan performa cukup baik.

**Kata kunci:** K-NN, Kanker payudara, Seleksi fitur, Binary PSO

# Implementation of Binary Particle Swarm Optimization Feature Selection on K-NN Algorithm for Breast Cancer Classification

## Abstract

Breast cancer is the most common type of cancer that affects women all over the world. Early detection of breast cancer is critical to patient treatment because the sooner cancer is diagnosed, the sooner treatment can begin. To make an accurate detection of breast cancer patients, this study was conducted with the aim of obtaining a classification model that can provide an accurate classification of breast cancer. The classification algorithm that is often used and is quite well known is K-Nearest Neighbor (K-NN). The K-NN algorithm performs classification using the nearest neighbor concept based on previous data, but the K-NN algorithm is weak on data with a large number of features. Therefore, for datasets with many features, feature selection can be used to improve the performance of K-NN classification. One that is often used for feature selection is the Binary Particle Swarm Optimization (BPSO) algorithm. In this study, 2 models of K-NN classification will be made, namely the K-NN classification model without feature selection and the K-NN classification model with BPSO feature selection. The accuracy results were then compared. The dataset that will be used is Breast Cancer Wincosin (Diagnostic) from the UCI Machine Learning Repository, which has 569 data points and 30 features. The results showed that the K-

NN+BPSO model obtained an accuracy of 95.32% and the K-NN model obtained an accuracy of 94.15%. Based on the accuracy obtained, the K-NN algorithm with feature selection gets better accuracy than the K-NN algorithm without feature selection by 1.17%. The PSO Binary Algorithm also succeeded in reducing features from 30 to 5 without reducing the accuracy of the classification model. So it can be concluded that in the case of breast cancer classification with the Breast Cancer Wincosin (Diagnostic) dataset, the K-NN algorithm can be combined with Binary PSO feature selection to create a classification model that provides quite good performance.

**Keywords:** K-NN, Breast cancer, Feature selection, Binary PSO

## I. PENDAHULUAN

Kanker Payudara adalah jenis kanker paling umum yang sering menyerang kalangan wanita di seluruh dunia. Kanker payudara merupakan salah satu masalah kesehatan yang serius karena banyak menyebabkan kematian wanita di seluruh dunia. Oleh karena itu diagnosa awal yang akurat dalam mendeteksi kanker payudara memainkan peran penting dalam pengobatan pasien [1].

Banyak penelitian telah berusaha menerapkan *machine learning* untuk melakukan klasifikasi kanker payudara sebagai upaya untuk membantu dalam deteksi dini kanker payudara. Salah satu penelitian terdahulu mengenai performa machine learning untuk klasifikasi kanker payudara dengan algoritma K-NN mendapatkan akurasi terbaik sebesar 96,7% dan menggunakan algoritma Decision tree didapat akurasi terbaik sebesar 93,7% [2].

Klasifikasi K-Nearest Neighbor (K-NN) adalah salah satu algoritma klasifikasi yang populer dalam pengenalan pola [3]. Algoritma KNN memiliki beberapa kelebihan antara lain memberikan waktu yang cepat dalam training dan mampu menghasilkan prediksi dengan rasio kesalahan yang kecil [4]. Akan tetapi algoritma KNN lemah terhadap data dengan jumlah fitur yang besar atau data berdimensi tinggi. [5].

Secara teori semakin banyak fitur semakin banyak informasi yang disediakan untuk classifier dan semakin bagus klasifikasi, tetapi dalam praktiknya proses *training* dan hasil klasifikasi kurang bagus dengan dimensi fitur yang besar. Banyaknya fitur membuat *learning speed* lebih lama dan akurasi klasifikasi menurun karena terdapat *redundansi* fitur [6]. Seleksi fitur adalah proses pemilihan set fitur terbaik dari semua fitur yang ada pada *dataset*. Seleksi fitur bukanlah hal mudah dikarenakan ruang pencarian yang luas dan kompleks yang membuat pencarian lengkap atau menyeluruh menjadi kurang efektif, oleh karena itu diperlukan teknik pencarian global yang efisien untuk kasus seleksi fitur [7].

Algoritma *swarm intelligence* banyak dilirik dan diterapkan dalam kasus seleksi fitur karena sederhana dan memiliki kemampuan pencarian global yang baik. Pada penelitian terkait algoritma *swarm intelligence*, *Binary Particle Swarm Optimization* digunakan untuk seleksi fitur dan berhasil meningkatkan akurasi klasifikasi K-NN sebesar 2.5% [8].

Berdasarkan masalah dan penjelasan diatas penelitian ini bertujuan untuk mengimplementasikan algoritma seleksi fitur *Binary PSO* pada algoritma klasifikasi KNN untuk klasifikasi kanker payudara dengan harapan dapat meningkatkan akurasi klasifikasi algoritma KNN dibandingkan tanpa seleksi fitur. Dataset yang akan digunakan pada penelitian ini adalah *Breast Cancer Wincosin (Diagnostic)* dari *UCI Machine Learning Repository* yang memiliki 569 data dan 30 fitur.

## II. METODOLOGI

Pada bab ini memuat langkah - langkah peneliti dalam melakukan penelitian secara lengkap dan padat. Penelitian akan dimulai dari penjelasan *dataset*, pembagian data, seleksi fitur, dan pembuatan model klasifikasi dan evaluasi model. Alur penelitian dapat dilihat pada gambar 1.

### A. Dataset

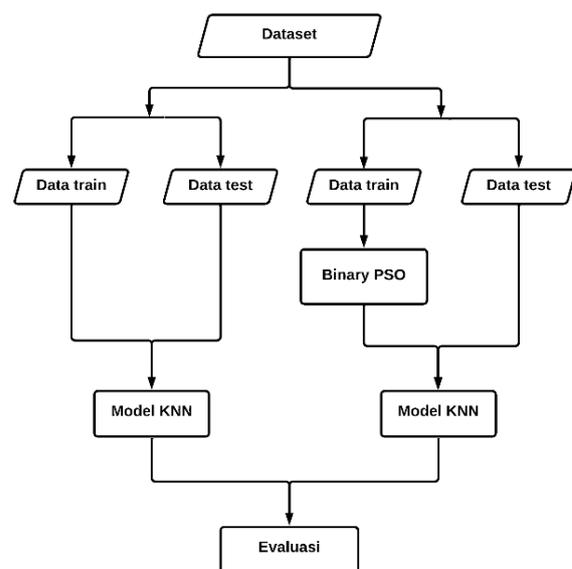
Dataset yang digunakan adalah dataset *Breast Cancer Wincosin (Diagnostic)* yang diperoleh dari web *UCI Machine Learning Repository*. Dataset ini digunakan untuk mengklasifikasikan jenis kanker payudara pada pasien apakah termasuk jenis kanker jinak (*benign*) atau kanker ganas (*malignant*). Dataset ini memiliki 569 *record* data dengan 30 fitur.

### B. Pembagian Data

*Dataset* dibagi menjadi 2 bagian yaitu data train dan data test. Pembagian data bertujuan untuk melakukan validasi model klasifikasi yang telah dibuat. Proporsi pembagian data train dan data testing pada penelitian ini adalah 70% data train dan 30% data test. 70% sebagai set pelatihan dan 30% sebagai set pengujian. Selama prosedur pelatihan, setiap agen mendemonstrasikan satu subset fitur. Setelah prosedur pelatihan, fitur yang dipilih dianalisis pada set pengujian untuk mencapai akurasi klasifikasi pengujian [9].

### C. Binary Particle Swarm Optimization

*Binary PSO* adalah pengembangan dari algoritma PSO yang digunakan untuk mengatasi optimasi dalam ruang pencarian biner. Pada algoritma Binary PSO posisi partikel hanya memiliki dua nilai yaitu 0 dan 1.



Gambar 1. Diagram alur penelitian

Mekanisme biner diperkenalkan untuk memetakan solusi dari ruang kontinyu kedalam ruang diskrit atau biner, sehingga membuat algoritma cocok untuk masalah seleksi fitur. Fungsi *sigmoid* digunakan dalam algoritma *binary PSO* untuk mengubah nilai kontinyu kedalam range [0, 1] dan kemudian diubah menjadi angka biner dengan proses treshold dengan angka random, detail fungsi ini adalah sebagai berikut :

$$x_{sig} = \frac{1}{1+e^{-x^{t+1}}} \tag{1}$$

$$x_{binary} = \begin{cases} 1, & N_{random} \leq x_{sig} \\ 0, & N_{random} > x_{sig} \end{cases} \tag{2}$$

dimana  $x_{binary}$  adalah posisi baru partikel dan  $N_{random}$  adalah angka random [10].

1) *Fitness Function* : Fitness function adalah fungsi objektif yang digunakan untuk mengukur seberapa dekat solusi dengan solusi optimal. Tujuan dari seleksi fitur adalah untuk mengurangi sejumlah fitur dan memaksimalkan akurasi klasifikasi, hal tersebut dapat dianggap sebagai masalah pengoptimalan multi-objective. Mempertimbangkan dua tujuan tersebut, dirancang fitness function untuk kasus seleksi fitur sebagai berikut :

$$Fitness = a \cdot E + b \cdot \frac{R}{C} \tag{3}$$

dimana E adalah error rate klasifikasi, R adalah jumlah fitur terpilih dan C adalah total fitur, kemudian a = 0,8 dan b = 0,2 adalah bobot yang digunakan untuk menyeimbangkan kedua objektif [11].

2) *Inertia* : Parameter *Inertia* ( $\omega$ ) pada penelitian ini menggunakan pendekatan *Linearly decreasing inertia weight* (LDIW) yang bertujuan agar algoritma melakukan pencarian global pada awal perulangan dan seiring waktu mengecil menjadi pencarian lokal di akhir perulangan, rumus inertia yang digunakan adalah sebagai berikut :

$$\omega = \omega_{max} - \left( \frac{\omega_{max} - \omega_{min}}{Iter_{max}} \right) x \text{ Iter} \tag{4}$$

dimana  $\omega_{max} = 0,9$  dan  $\omega_{min} = 0,4$  dan jumlah iterasi maksimal adalah 30 iterasi [12].

3) *C1 dan C2* : Pada penelitian ini C1 = 2,1 dan C2 = 1,9. Konstansa akselerasi dalam persamaan pembaruan kecepatan mengontrol pergerakan partikel menuju posisi *local best* dan *global best*. Nilai yang besar akan membuat divergensi dari solusi optimal dan nilai yang kecil akan membatasi pergerakan partikel. Oleh karena itu nilai harus dipilih memenuhi kondisi  $C1 + C2 \leq 4$  [11].

4) *Jumlah Partikel* : Pada penelitian ini jumlah partikel yang digunakan adalah 5, 10, 15, 20, 25, dan 30. Jumlah partikel ditentukan untuk mempercepat waktu perhitungan, dimana diketahui jika jumlah partikel besar maka perhitungan akan lebih lama. Pada kasus yang kompleks lebih direkomendasikan jumlah partikel yang besar [4].

#### D. Klasifikasi K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* (KNN) adalah metode untuk melakukan klasifikasi terhadap objek yang

berdasarkan dari data pembelajaran yang jaraknya paling dekat dengan objek tersebut. KNN merupakan algoritma supervised learning dimana hasil dari query instance yang baru diklasifikasi berdasarkan mayoritas dari kategori pada algoritma KNN dimana kelas yang paling banyak muncul yang nantinya akan menjadi kelas hasil dari klasifikasi. Kedekatan didefinisikan dalam jarak metrik, seperti jarak *Euclidean* [13].

Algoritma *K-Nearest Neighbor* atau K-NN adalah algoritma yang digunakan untuk mengklasifikasi objek berdasarkan pada perhitungan data dengan jarak terdekat dari objek tersebut [14]. Pada penelitian ini akan dibuat dua model K-NN yaitu model K-NN tanpa seleksi fitur dan model K-NN dengan seleksi fitur *Binary PSO* untuk melihat pengaruh seleksi fitur pada algoritma K-NN. Parameter  $k$  yang digunakan adalah bilangan ganjil dari rentang 1 sampai 21. nilai k dipilih ganjil untuk menghindari seri untuk data yang genap. Jadi dalam kasus klasifikasi biner, perlu untuk mengambil nilai k yang ganjil sehingga instance mayoritas akan muncul[1].

#### E. Evaluasi Model

Merupakan penilaian hasil kinerja dari semua model yang telah dibuat menggunakan dataset yang dipakai. Pada tahap ini akan dibandingkan hasil akurasi klasifikasi model klasifikasi K-NN tanpa seleksi fitur dan model klasifikasi K-NN dengan seleksi fitur *Binary PSO*.

Perhitungan akurasi menggunakan confusion matrix. Confusion matriks adalah alat ukur yang digunakan untuk mengevaluasi performa suatu model klasifikasi. Akurasi adalah ukuran seberapa tepat model dalam mengklasifikasikan dengan benar [15]. Berikut adalah rumus akurasi dengan *confusion matrix* :

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

### III. HASIL DAN PEMBAHASAN

Pada bab ini akan disampaikan hasil seleksi fitur menggunakan *Binary PSO* dan juga hasil akurasi yang didapatkan oleh model K-NN tanpa seleksi fitur dan model K-NN dengan seleksi fitur. Kemudian akan dilakukan perbandingan akurasi terbaik antar model K-NN tanpa seleksi fitur dan model K-NN dengan seleksi fitur.

#### A. Seleksi fitur

Proses seleksi fitur hanya menggunakan *data train* untuk menemukan kombinasi fitur terbaik. Kemudian setelah didapatkan hasil fitur terbaik, fitur lain akan dihapus dari data train maupun data test menyisakan fitur yang sudah diseleksi. Hasil seleksi fitur menggunakan algoritma *Binary PSO* dapat dilihat pada tabel 1.

TABEL I  
HASIL SELEKSI FITUR BINARY PSO

Jumlah Partikel	Fitness Score	Jumlah Fitur	Hasil
5	0,094908	7	[1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0]
10	0,082211	6	[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0]



*Binary PSO*. Berdasarkan hasil penelitian, model K-NN menghasilkan akurasi terbaik sebesar 94,15% pada k=13 dengan menggunakan 30 fitur dan model K-NN+BPSO menghasilkan akurasi terbaik sebesar 95,32% pada k=9 dengan menggunakan 5 fitur yaitu *compactness se, symmetry se, texture worst, perimeter worst, dan symmetry worst*. Hasil yang didapatkan akurasi algoritma K-NN dengan seleksi fitur lebih baik daripada algoritma K-NN tanpa seleksi fitur sebesar 1,17%. Algoritma Binary PSO berhasil mereduksi fitur dari 30 fitur menjadi 5 fitur dengan tidak mengurangi akurasi dari model klasifikasi, selain itu dengan berkurangnya jumlah fitur yang di proses oleh algoritma dapat mengurangi tenaga komputasi atau perhitungan algoritma menjadi lebih ringan. Dengan demikian dapat disimpulkan bahwa pada penelitian ini untuk kasus klasifikasi kanker payudara menggunakan dataset Breast Cancer Wincosin (Diagnostic) dengan algoritma K-NN dapat dikombinasikan dengan seleksi fitur BPSO untuk membuat model klasifikasi yang memberikan performa lebih baik dibandingkan tanpa adanya seleksi fitur.

#### DAFTAR PUSTAKA

- [1] H. Harafani and H. A. Al-Kautsar, "Meningkatkan Kinerja K-NN Untuk Klasifikasi Kanker Payudara Dengan Forward Selection," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 18, no. 1, p. 99, 2021, doi: 10.23887/jptk-undiksha.v18i1.29905.
- [2] O. I. Obaid, M. A. Mohammed, M. K. A. Ghani, S. A. Mostafa, and F. T. Al-Dhief, "Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer," *Article in International Journal of Engineering and Technology*, vol. 7, no. 4, pp. 160–166, 2018, doi: 10.14419/ijet.v7i4.36.23737.
- [3] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat, "Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach," vol. 12, no. 8, pp. 33–39, 2014, [Online]. Available: <http://arxiv.org/abs/1409.0919>
- [4] W. Yunus, "Algoritma K-Nearest Neighbor Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Ginjal Kronik," *Jurnal Teknik Elektro CosPhi*, vol. 2, no. 2, pp. 51–55, 2018.
- [5] N. Kouiroukidis and G. Evangelidis, "The effects of dimensionality curse in high dimensional knn search," *Proceedings - 2011 Panhellenic Conference on Informatics, PCI 2011*, pp. 41–45, 2011, doi: 10.1109/PCI.2011.45.
- [6] X. Zhang, Z. Shi, X. Liu, and X. Li, "A Hybrid Feature Selection Algorithm For Classification Unbalanced Data Processing," *2018 IEEE International Conference on Smart Internet of Things (SmartIoT)*, pp. 269–275, 2018, doi: 10.1109/smartiot.2018.00055.
- [7] B. H. Nguyen, B. Xue, and M. Zhang, "A survey on swarm intelligence approaches to feature selection in data mining," *Swarm and Evolutionary Computation*, vol. 54, no. April 2019, p. 100663, 2020, doi: 10.1016/j.swevo.2020.100663.
- [8] H. Gohzali, S. Megawan, and J. Onggo, "Rekomendasi Buku Menggunakan K-Nearest Neighbor (KNN) dan Binary Particle Swarm Optimization (BPSO)," *Jurnal SIFO Mikroskil (JSM) STMIK Mikroskil*, vol. 20, no. 1, pp. 1–5, 2019, [Online]. Available: <https://www.ranks.nl/stopwords>.
- [9] T. A. Khan, K. Zain-Ul-Abideen, and S. H. Ling, "A modified particle swarm optimization algorithm used for feature selection of UCI biomedical data sets," *60th International Scientific Conference on Information Technology and Management Science of Riga Technical University, ITMS 2019 - Proceedings*, pp. 1–4, 2019, doi: 10.1109/ITMS47855.2019.8940760.
- [10] H. Lubis, P. Sirait, and A. Halim, "KNN Method on Credit Risk Classification With Binary Particle Swarm Optimization Based Feature Selection," *Jurnal INFOKUM*, vol. 9, no. 2, pp. 211–218, 2021.
- [11] B. Ji, X. Lu, G. Sun, W. Zhang, J. Li, and Y. Xiao, "Bio-Inspired Feature Selection: An Improved Binary Particle Swarm Optimization Approach," *IEEE Access*, vol. 8, pp. 85989–86002, 2020, doi: 10.1109/ACCESS.2020.2992752.
- [12] W. Han, P. Yang, H. Ren, and J. Sun, "Comparison study of several kinds of inertia weights for PSO," *Proceedings of the 2010 IEEE International Conference on Progress in Informatics and Computing, PIC 2010*, vol. 1, no. 09, pp. 280–284, 2010, doi: 10.1109/PIC.2010.5687447.
- [13] A. Johar, D. Yanosma, and K. Anggriani, "Implementasi Metode K-Nearest Neighbor (Knn) Dan Simple Additive Weighting (Saw) Dalam Pengambilan Keputusan Seleksi Penerimaan Anggota Paskibraka," *Pseudocode*, vol. 3, no. 2, pp. 98–112, 2017, doi: 10.33369/pseudocode.3.2.98-112.
- [14] S. H. A. Aini, Y. A. Sari, and A. Arwan, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 9, pp. 2546–2554, 2018.
- [15] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.