Implementasi Algoritma Binary Particle Swarm Optimization (BPSO) dan C4.5 Decision Tree untuk Deteksi Kanker Berdasarkan Klasifikasi Microarray Data

Amalya Citra Pradana[1], Adiwijaya[2], Annisa Aditsania[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]amalyacitra@students.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id,
[3]aaditsania@telkomuniversity.ac.id

Abstrak

Kanker merupakan salah satu penyakit yang mematikan di dunia. Upaya pendeteksian kanker dapat dilakukan dengan merepresentasikan kanker ke dalam *microarray data* dengan mengukur perubahan yang terjadi pada level ekspresi gen. Deteksi gejala kanker dapat dilakukan dengan teknik *data mining,* yaitu klasifikasi terhadap *microarray data*. Salah satu penerapan algoritma untuk klasifikasi adalah C4.5 *Decision Tree* dimana algoritma tersebut mudah diinterpretasi dan termasuk paling berpengaruh dalam klasifikasi namun memiliki kekurangan yaitu sensitif terhadap data *noise*. *Microarray data* memiliki jumlah *feature* yang sangat besar (*high dimensional*) dimana tidak semua *feature* tersebut memiliki informasi yang penting (*high noise*) dan jumlah sampel yang sedikit sehingga penerapan proses klasifikasi saja menjadi sulit karena dapat mempengaruhi nilai akurasi. *Binary Particle Swarm Optimization* (BPSO) merupakan salah satu algoritma optimasi pencarian untuk mendapatkan fitur yang optimal. Pemodelan *rule* pada *Decision Tree* menggunakan nilai diskrit sehingga data perlu didiskritkan. Diskritisasi dilakukan menggunakan K-Means. Sistem dibagi menjadi dua skema yaitu skema *Information Gain* (IG) – C4.5 dan skema BPSO – C4.5. Akurasi yang diperoleh berdasarkan skema IG-C4.5 dan BPSO-C4.5 berturut-turut adalah 54% dan 99%. Pengaruh seleksi fitur terhadap klasifikasi berperan penting dalam menghindari data *noise* untuk memodelkan *rule* yang akurat. Dengan penerapan BPSO sebagai seleksi fitur mampu mencari fitur yang paling signifikan.

Kata kunci : *microarray data, binary particle swarm optimization, C4.5 decision tree, classification, feature selection, K-Means*

Abstract

Cancer is one of deadly disease in the world. Cancer can be detected by representing the cancer into microarray data with measuring the changes occured in gene expression level. Cancer detection can be done by doing classification technique for microarray data. One of most algorithm that applied for classification is Decision Tree C4.5. It is a linier method which is easy to interpret and included into the algorithm which has given impact in classification but it is sensitive to noise data. Microarray data has a large features (high dimensional) which is not all features have important information (high noise) and has a small samples and causing the application is difficult and affected the accuracy. Binary Particle Swarm Optimization (BPSO) is one of searching optimization algorithm that could find an optimal feature. Rule in Decision Tree is modelled with discrete value so the data has to be discretized. Discretization is applied using K-Means. System is divided into two schemas such as Information Gain (IG) – C4.5 and BPSO – C4.5. The accuracy based on IG – C4.5 and BPSO – C4.5 schema are 54% and 99%. Feature selection has given impact to classification for avoiding noise data to build the rule accurately. With applying BPSO as feature selection can find the features significantly.

Keywords: microarray data, binary particle swarm optimization, C4.5 decision tree, classification, feature selection, K-Means

1. Pendahuluan

Pada bagian pendahuluan memuat beberapa sub bab. Sub bab tersebut diantaranya latar belakang, perumusan dan batasan masalah, tujuan, dan organisasi penulisan. Latar belakang mengemukakan masalah dan rencana penelitian, perumusan dan batasan masalah mengemukakan permasalahan dan batasan pada penelitian, sub bab tujuan mengemukakan tujuan yang dicapai pada penelitian, dan organisasi penulisan mengemukakan struktur penulisan pada *paper* ini.

Latar Belakang

Bioinformatics merupakan bidang area yang mengaplikasikan molekular biologi ke dalam data yang berisi rangkaian DNA manusia (*genomic data*) dengan teknologi komputer. Teknologi ini membantu peneliti untuk mendiagnosa masalah kesehatan secara akurat dengan mengukur level mRNA akan didapatkan informasi yang terjadi dengan keadaan sel sehingga dapat memahami informasi proses biologikal yang terjadi didalamnya, contohnya adalah penyakit kanker. Kanker merupakan salah satu penyakit yang mematikan di dunia [1]. Salah

satu cara untuk menghindari penyakit ini adalah dengan mendeteksi gejalanya sejak dini. Data kanker dapat dipresentasikan dengan menggunakan teknologi DNA *microarray*.

DNA *microarray* merupakan sebuah *chip* yang terbuat dari silikon/kaca seperti dalam *Affymetrix array* atau *microscopic head* dalam *Illumina array* dimana ribuan untaian molekul atau *oglinucleotides complimentary* DNA (cDNA) ditanamkan yang dinamai dengan *feature*. Dengan membandingkan hasil dari teknologi DNA *microarray* dengan sel yang normal, peneliti dapat mengukur perubahan yang terjadi pada level ekspresi gen sehingga semua ribuan ekspresi gen dapat diinvestigasi secara bersamaan [2].

Untuk mendeteksi gejala kanker, salah satu teknik yang dapat digunakan adalah melakukan *data mining* terhadap *feature* dalam *microarray data. Data mining* merupakan rangkaian proses untuk mendapatkan sebuah informasi penting yang mulanya tidak diketahui, salah satunya dengan teknik klasifikasi yaitu teknik untuk menggolongkan data yang memiliki karakteristik tertentu yang sama. Dengan teknik tersebut akan dapat menentukan apakah penderita memiliki penyakit kanker [3].

Salah satu penerapan algoritma untuk klasifikasi *microarray data* adalah algoritma *decision tree,* contohnya adalah C4.5. *Decision tree* merupakan metode linier yang mudah di interpretasi dan termasuk dalam sepuluh algoritma *data mining* yang paling berpengaruh dalam klasifikasi *microarray data,* namun sensitif terhadap data dengan banyak *noise* seperti *microarray data* [4]. *Microarray data* memiliki jumlah *feature* yang sangat banyak (*high dimensional*) dimana tidak semua *feature* tersebut memiliki informasi yang penting (*high noise*) dan jumlah sampel yang sedikit sehingga penerapan teknik klasifikasi pada *microarray data* menjadi sulit karena dapat mempengaruhi nilai akurasi [1]. Untuk mengatasi masalah penerapan teknik klasifikasi pada *microarray data* perlu dilakukan reduksi dimensi sebelum dilakukannya klasifikasi, yakni *feature selection* sebagai alat *pre-processing* data.

Feature selection terdiri dari kategori *univariate* dan *multivariate*. Salah satu metode pada kategori *multivariate* adalah *embedded method*. *Embedded method* memiliki kelebihan yaitu dapat menguji kekuatan prediksi gen [5]. Dalam *embedded method* akan mencari *feature* yang kontribusinya terbaik yang merupakan *feature* yang paling optimal. Pencarian untuk mendapatkan *feature* optimal tersebut dapat menggunakan pendekatan *meta-heuristic*. Salah satu algoritma dari pendekatan *meta-heuristic* adalah *Particle Swarm Optimization* (PSO). *Feature selection* merupakan masalah yang menangani data diskrit sehingga PSO perlu dimodifikasi menjadi *Binary* PSO (BPSO) [6].

Dari rekomendasi mengenai penelitian selanjutnya dari penelitian sebelumnya [6], BPSO dapat mencapai hasil yang lebih baik jika inisialisasi parameter, nilai iterasi dan penggunaan nilai konstan faktor *weight* dikembangkan lebih lanjut. Pada penelitian ini menerapkan penggunaan nilai konstan faktor *weight*, nilai iterasi generasi yang digunakan sama seperti penelitian sebelumnya dan mengobservasi beberapa nilai inisialisasi parameter untuk BPSO.

Perumusan dan Batasan Masalah

Perumusan masalah dari penelitian ini yaitu bagaimana mengimplementasikan *feature selection* dan klasifikasi terhadap *microarray data*, bagaimana pengaruh diskritisasi terhadap klasifikasi *microarray data*, bagaimana pengaruh *feature selection* terhadap klasifikasi *microarray data*, dan bagaimana performansi yang dihasilkan dari sistem yang dibangun berdasarkan observasi parameter terhadap deteksi kanker.

Batasan dalam penelitian ini yaitu data yang digunakan adalah *ALL-AML Leukimia, Breast Cancer, Colon Tumor*, *Lung Cancer*, dan *Ovarian Cancer* yang diambil dari Kent Ridge Biomedical Dataset Repository. *Decision Tree* akan bekerja jika data berbentuk diskit. Karena data pada *microarray* berbentuk kontinu maka data perlu di diskritisasi. Untuk diskritisasi data pada penelitian ini menggunakan metode *clustering* K-Means dikarenakan merupakan salah satu penerapan diskritisasi data untuk *microarray* yang mengacu pada observasi penelitian sebelumnya [7].

Tujuan

Tujuan dari penelitian ini yaitu mengimplementasikan *feature selection* dan klasifikasi terhadap *microarray data* dengan menggunakan *Binary Particle Swarm Optimization* (BPSO) sebagai *feature selection* dan C4.5 *Decision Tree* sebagai *classifier*, mengetahui pengaruh diskritisasi dan *feature selection* terhadap klasifikasi *microarray data*, dan mengetahui performansi yang dihasilkan dari sistem yang dibangun berdasarkan observasi parameter terhadap deteksi kanker.

Organisasi Penulisan

Setelah bagian Pendahuluan, bagian selanjutnya terdiri dari bagian Studi Literatur, Pembangunan Sistem, Evaluasi dan Kesimpulan. Bagian Studi Literatur memuat teori *microarray*, studi komparatif dan deskripsi metode-metode yang diterapkan pada penelitian ini. Bagian Pembangunan Sistem memuat pembangunan sistem secara umum dan dibagi dalam dua skema secara detail. Bagian Evaluasi memuat hasil pengujian sistem dan analisis hasil pengujian. Bagian Kesimpulan memuat kesimpulan dan saran untuk penelitian selanjutnya.

2. Studi Literatur

Pada bagian studi literatur memuat beberapa sub bab. Sub bab tersebut berkaitan dengan teori dan metode yang diterapkan pada penelitian ini.

2.1 DNA Microarray

Bioinformatics merupakan bidang area yang mengaplikasikan molekular biologi ke dalam data yang berisi rangkaian DNA manusia (*genomic data*) dengan teknologi komputer. Gen dibuat dalam bentuk protein tergantung kondisi dan waktu yang berbeda. Pada pembuatan protein, instruksi ditranskripsi oleh *messenger* RNA (mRNA) dari DNA pada gen yang tersisa dalam nukleus dan pembuatan terjadi dalam ribosom sel. Sebagai hasilnya, keadaan sel akan berkorelasi dengan perubahan dalam level mRNA. Dengan mengukur level mRNA tersebut akan didapatkan apa yang terjadi dengan keadaan sel tersebut sehingga dapat memahami informasi proses biologikal yang terjadi didalamnya. Teknologi ini membantu peneliti untuk mendiagnosa masalah kesehatan secara akurat. Data kesehatan tersebut dapat dipresentasikan dengan menggunakan teknologi DNA *microarray* [2].

DNA *microarray* merupakan sebuah *chip* yang dibuat dari silikon atau kaca dalam *affymetrix array* dimana ribuan molekul cDNA diimplan. Campuran dari mRNA diturunkan dari sel dan diizinkan untuk melakukan hibridisasi ke rangkaian *probe chip*. Level hibridisasi dideteksi menggunakan *fluorescence dyes* dengan aplikasi *imaging* dari pabrik *chip*. Diasumsikan konsentrasi level setiap mRNA proposional terhadap intensitas yang dideteksi. *Microarray* memiliki karakteristik jumlah *feature* yang sangat besar (*high dimensional*) dimana tidak semua *feature* tersebut memiliki informasi yang penting (*high noise*) dan jumlah sampel yang sedikit [3].

Teknologi DNA *microarray* dikembangkan untuk mengumpulkan data ekspresi gen yang berjumlah sangat besar di waktu yang bersamaan. Gen diekspresikan melalui beberapa kondisi dan waktu ke dalam protein. Instruksi ditranskrip oleh mRNA dari DNA dalam gen yang tersisa dalam nukleus. Sebagai hasilnya, keadaan sel berkorelasi dengan perubahan dalam level mRNA sehingga dapat ditentukan apa yang terjadi dalam sel [1].

2.2 Studi Komparatif

Perkembangan pembelajaran mesin dalam dua puluh tahun terakhir digunakan dalam bidang *bioinformatics* untuk menganalisa masalah kesehatan, salah satunya untuk mendeteksi gejala kanker. Beberapa penelitian sebelumnya yang menerapkan algoritma untuk menyelesaikan klasifikasi *microarray data* antara lain:

a. Pada penelitian yang dilakukan oleh Kun-Huang Chen et al. pada tahun 2014 menganalisa penggunaan berbagai teknik klasifikasi untuk sepuluh *microarray data*, diantaranya *Support Vector Machine*, *Self-organizing map, back propagation neural network, CART decision tree*, C4.5 *decision tree*, *Artificial Immune Recognition System*, *Naive Bayes*. Metode yang ditujukan yaitu *Binary Particle Swarm Optimization* (BPSO) sebagai *feature selection* dan C4.5 sebagai *classifier*. Hasil yang didapatkan bahwa BPSO dengan C4.5 rata-rata mendapatkan hasil akurasi yang lebih unggul yaitu 87% dibandingkan algoritma *Support Vector Machine* (83%), *Self-organizing map* (55%), *back propagation neural network* (44%), *CART decision tree* (70%), C4.5 *decision tree* (73%), *Artificial Immune Recognition System* (50%), dan *Naive Bayes* (75%) [8].
b. Pada penelitian yang dilakukan oleh Meng-Chang Tsai, Kun-Huang Chen, Chao-Ton Su, dan Hung-Chun Lin pada tahun 2012 menganalisa penggunaan berbagai teknik klasifikasi untuk lima *microarray data*, diantaranya *Support Vector Machine*, *back propagation neural network*, *logistic regression,* C4.5 *decision tree*. Metode yang ditujukan yaitu *Binary Particle Swarm Optimization* (BPSO) sebagai *feature selection* dan C4.5 sebagai *classifier*. Hasil yang didapatkan bahwa BPSO dengan C4.5 rata-rata mendapatkan hasil akurasi yang lebih unggul sebesar 75% dibandingkan algoritma *Support Vector Machine* (70%), *back propagation neural network* (58%), *logistic regression* (74%) dan C4.5 *decision tree* (71%) [9].
c. Pada penelitian yang dilakukan oleh Chung-Jui Tu, Li-Yeh Chuang, Jun-Yang Chang, dan Cheng-Hong Yang pada tahun 2007 menganalisa penggunaan berbagai teknik klasifikasi untuk lima *microarray data* dengan SFS, PTA, SFFS, SGA, HGA. Metode yang ditujukan yaitu *Binary Particle Swarm Optimization* (BPSO) sebagai *feature selection* dan *Support Vector Machine* sebagai *classifier*. Hasil yang didapatkan bahwa BPSO dengan SVM rata-rata mendapatkan hasil akurasi diatas 70% [10].
d. Pada penelitian yang dilakukan oleh Jian J. Dai, Linh Lieu, dan David Rocke pada tahun 2006 menganalisa penggunaan berbagai teknik klasifikasi untuk dua *microarray data* dengan *Particle Least Square, Sliced Inverse Regression, Principal Component Analysis*. Hasil *error rate* yang didapatkan bahwa PLS lebih baik sebesar 0.025 untuk leukimia dan 0.136 untuk colon dibandingkan

SIR (0.026 untuk leukimia dan 0.141 untuk colon) dan PCA (0.042 untuk leukimia dan 0.162 untuk colon) [11].

Berdasarkan hasil studi komparatif tersebut, didapatkan bahwa penerapan BPSO dengan kelebihannya yang dapat mencari hasil optimal sebagai *feature selection* dengan algoritma *decision tree* seperti C4.5 sebagai *classifier* terhadap *microarray data* menghasilkan akurasi dengan rata-rata diatas 70% dibandingkan algoritma lain.

2.3 Binary Particle Swarm Optimization (BPSO)

Binary Particle Swarm Optimization (BPSO) merupakan modifikasi dari *Particle Swarm Optimization* (PSO) untuk kasus *feature selection*, dimana metode ini termasuk ke pendekatan meta-heuristik berdasarkan pada perilaku sosial sederhana dari kawanan burung atau sekolah ikan. Dengan memperhatikan kawanan burung jika seekor burung menemukan makanan, informasi ini menyebar ke seluruh kawanan dan pada akhirnya, kawanan burung tersebut dapat mengalokasi makanan tersebut. Menyimulasikan populasi sederhana ini berdasarkan strukturnya akan menghasilkan algoritma optimasi yang sangat kuat. Asumsikan burung tersebut sebagai partikel dan burung dengan kawanannya sebagai populasi partikel. Tiap partikel memiliki kemampuan untuk berkomunikasi satu sama lain dan mencoba untuk mencapai partikel yang memiliki nilai *fitness* terbaik lebih dekat, contohnya yaitu partikel yang posisinya lebih dekat ke *origin*. Hasil akhir menunjukkan semua partikel yang bertepatan di tempat yang sama yang akan memiliki nilai *fitness* yang terbaik seperti *origin* nya. Setiap partikel memiliki parameter *velocity* yang membuat partikel-partikel bergerak ke arah partikel terbaik (partikel yang memiliki nilai *fitness* terbaik). *Velocity* didefinisikan sebagai [6]:

$$v_{id}^{new} = w \times v_{id}^{old} + c_1 r_1 \left(pb_{id}^{old} - x_{id}^{old}\right) + c_2 r_2 \left(gb_{id}^{old} - x_{id}^{old}\right) \quad (2.1)$$

dimana

v_{id}^{new} : *velocity* partikel terbaru
v_{id}^{old} : *velocity* partikel sebelumnya
x_{id}^{old} : posisi partikel sebelumnya
r_1 dan r_2 : nilai *random* antara 0 dan 1
c_1 : faktor *cognitive learning*
c_2 : faktor *social learning*
w : *inertia weight*
gb : *global best*
pb : *personal best*

Posisi sebelumnya yang terbaik dari sebuah partikel akan disimpan sebagai *personal best* (*pb*) dan posisi terbaik yang diperoleh oleh populasi sejauh ini disebut *gb*. X_{id} merupakan posisi saat ini dari partikel *i* dimana *d* merupakan dimensi dari ruang pencarian [4]. Pada metode ini diperlukan *input* pada *parameter value* seperti v_{max}, v_{min}, r_1, r_2, c_1, c_2 dan w agar mendapatkan hasil pencarian yang optimal [12].

Pada PSO, posisi partikel dibuat berdasarkan nilai acak. Namun pada kasus ini representasi posisi tidak dapat hanya nilai acak saja karena tidak dapat mempresentasikan penggunaan fitur sehingga perlu dimodifikasi menjadi biner. Pada BPSO, posisi partikel dimodelkan ke dalam bentuk *bit string* untuk membatasi *velocity* dalam interval [0,1]. *Velocity* didefinisikan sebagai probabilitas sebagai *bit* X_{ij} (i^{th} partikel dan j^{th} *bit*) untuk mencapai angka bernilai satu. Untuk membatasi *velocity* dapat menggunakan fungsi transformasi *limiting* dimana menerapkan fungsi *sigmoid* yang diajukan oleh Kennedy Eberhart [12], yang didefinisikan pada persamaan (2.2) berikut [6].

$$x_{id}^{new} = \begin{cases} 1, & sigmoid(v_{id}^{new}) > rand \\ 0, & sigmoid(v_{id}^{new}) < rand \end{cases} \quad (2.2)$$

dimana *rand* merupakan nilai *random* dalam interval [0,1] dan persamaan *sigmoid* pada persamaan (2.3) berikut.

$$sigmoid(v_{id}^{new}) = \frac{1}{1+e^{-v_{id}^{new}}} \quad (2.3)$$

2.4 Decision Tree C4.5

Decision Tree C4.5 merupakan algoritma *top-down* yang membangun model *decision tree* menggunakan proses rekursif (strategi *divides and conquer*). Algoritma ini merupakan improvisasi dari *decision tree* ID3 dalam seleksi gen dan pemangkasan model *tree*. C4.5 telah digunakan dalam masalah dunia nyata khususnya pembuatan keputusan dalam hal medis, dikarenakan algoritma ini dapat menyediakan akurasi klasifikasi dan dapat direpresentasikan secara sederhana. Dengan menggunakan *gain ratio* sebagai kriteria *splitting*, atribut

yang memiliki informasi penting yang dikomputasi pada *training set* yaitu dipilih pertama, kemudian menyeleksi atribut yang memiliki informasi yang paling informatif, dan seterusnya [9].

Sebelum mengekstrak data ke dalam model *tree*, proses yang perlu dilakukan adalah menentukan atribut yang akan menjadi *root* berdasarkan *gain ratio* terbesar, menentukan atribut yang akan menjadi *internal node* untuk setiap *branch* dari node *parent*, dan memangkas keputusan percabangan *node* dengan melakukan seleksi *pre-prunning* yaitu jika perbandingan antara jumlah pilihan percabangan bertimpang jauh sehingga jika dilakukan percabangan lebih lanjut akan menghasilkan keputusan yang tidak berpengaruh signifikan sehingga proses percabangan lebih lanjut tersebut akan berhenti. Persamaan yang digunakan dalam algoritma ini antara lain [9]:

a. *Entropy Value Calculation*
Entropy digunakan untuk mengukur heterogenitas dari sampel data. Persamaan *entropy* tertera pada persamaan (2.4) berikut.

$$Entropy(S) = -\sum_i^c \; p_i \log_2 p_i \tag{2.4}$$

dimana:
C = nilai dalam atribut target (angka kelas)
p_i = proporsi sampel dalam kelas i

b. *Information Gain Calculation*
Persamaan *information gain* tertera pada persamaan (2.5) berikut.

$$Gain(S, A) = Entropy(S) - \sum_v \; \frac{|s_v|}{s} \; entropy(s_v) \tag{2.5}$$

dimana:
V = nilai *probable* untuk atribut A
$|s_v|$ = nilai sampel untuk *value* v
S = nilai dari semua sampel data
Entropy(S) = entropi untuk sampel dengan nilai v

c. *Gain ratio*
Persamaan *gain ratio* tertera pada persamaan (2.6) berikut.

$$gain\ ratio = \frac{Gain(S,A)}{split\ information\ (S,A)} \tag{2.6}$$

dimana *split information* tertera pada persamaan (2.7) berikut.

$$split\ information = -\sum_i^c \frac{s_i}{s} \log_2 \frac{s_i}{s} \tag{2.7}$$

dimana nilai s_i adalah *subset* C dibuat dari *splitting* S menggunakan atribut A dengan varian nilai C.

2.5 K-Fold Cross Validation

K-fold cross validation merupakan salah satu metode statistikal untuk menjamin perbandingan hasil klasifikasi antar data merata dan menghindari hasil yang menghasilkan nilai random. Metode ini membagi data ke dalam dua bagian yaitu bagian pelatihan mengenal model data dan evaluasi model. Kemudian dilakukan *cross validation* pada dataset yang dibagi sejumlah k [8].

Langkah awal pada metode ini menentukan nilai k. Pada setiap iterasi, bagi data dalam data pelatihan dan data evaluasi, lalu menentukan bagian data yang akan dilatih dan yang diuji dengan perbandingan nilai k. Performansi algoritma setiap kasus dilihat dari akurasi yang didapat. Setelah semua kasus telah dilakukan maka akan didapat rata-rata akurasi yang dihasilkan [8].

2.6 K-Means

K-Means merupakan salah satu metode *clustering* yang sering digunakan dalam *data mining*. Algoritma ini mengelompokkan data yang diolah sebanyak jumlah k dengan membuat inisialisasi nilai *centroid* secara random dan label *cluster* sebanyak k, lalu dihitung jarak tiap data terhadap *centroid*. Jika data yang jaraknya paling terdekat terhadap suatu nilai *centroid* maka data tersebut akan diberi label *cluster* dari *centroid* tersebut. Kemudian nilai *centroid* diperbaharui dan dievaluasi untuk mengurangi tingkat *error* hingga mencapai hasil yang signifikan dengan menggunakan *Sum of Squared Error* (SSE) [13].

Persamaan untuk menghitung nilai centroid tertera pada persamaan (2.8) berikut.

$$Centroid = (\max(p_{i..n}) - \min(p_{i..n})) * rand + \min(p_{i..n}) \tag{2.8}$$

dimana:
n = jumlah data
$p_{i..n}$ = data ke-i sampai data ke-n
rand = nilai random antara [0,1]
max = nilai terbesar dari suatu data
min = nilai terkecil dari suatu data

Persamaan untuk menghitung nilai jarak antara data dengan *centroid* menggunakan *euclidean distance* tertera pada persamaan (2.9) berikut.

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{k}(p_i - c_j)^2} \quad (2.9)$$

dimana:
n = jumlah data
k = jumlah *cluster*
p_i = data ke-i
c_j = *centroid* ke-j

Persamaan untuk menghitung nilai SSE tertera pada persamaan (2.10) berikut.

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{k}\left\|p_i - c_j\right\|^2 \quad (2.10)$$

dimana:
n = jumlah data
k = jumlah *cluster*
p_i = data ke-i
c_j = *centroid* ke-j

3. Pembangunan Sistem

Pada bagian pembangunan sistem memuat beberapa sub bab. Sub bab tersebut berkaitan dengan perancangan sistem secara umum dan detail.

3.1 Gambaran Sistem

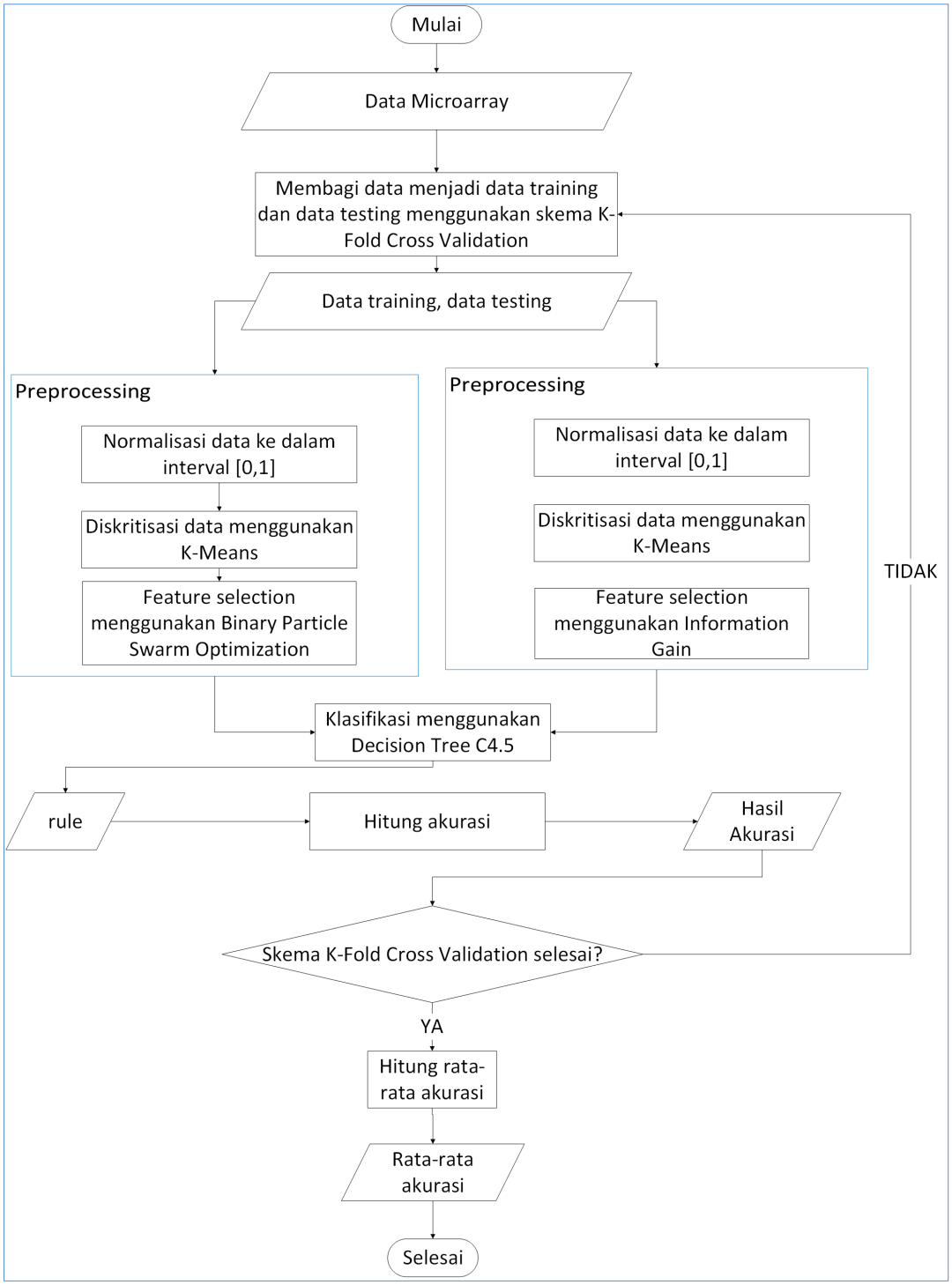
Secara umum, sistem dimulai dengan pembagian data *training* dan data *testing*. Pembagian data dilakukan berdasarkan skema K-Fold Cross Validation. Setelah data terbagi, skema sistem dibagi menjadi dua yaitu skema klasifikasi dengan *Information Gain* sebagai seleksi fitur dan skema klasifikasi dengan BPSO sebagai seleksi fitur dengan C4.5 *Decision Tree* sebagai klasifier.

Pada skema klasifikasi menggunakan *Information Gain*, pada *preprocessing* dilakukan normalisasi dan diskritisasi data menggunakan K-Means. Setelah *preprocessing*, nilai *threshold* ditentukan untuk pemilihan fitur berdasarkan nilai *information gain*. Setelah fitur diseleksi, dilakukan klasifikasi menggunakan *Decision Tree* C4.5 menggunakan data *training* dan menghasilkan sebuah *rule*. Setelah *rule* dibangun, dilakukan evaluasi dengan menghitung akurasi dari data *testing* terhadap *rule*. Setelah akurasi didapat, akurasi disimpan dan skema ini dilakukan pada kasus selanjutnya.

Pada skema klasifikasi menggunakan BPSO, pada *preprocessing* dilakukan normalisasi dan diskritisasi data menggunakan K-Means. Setelah itu dilakukan seleksi fitur menggunakan *Binary Particle Swarm Optimization*. Pada seleksi fitur akan menyeleksi fitur yang digunakan sebelum ke tahap klasifikasi kemudian hasil akurasi dari klasifikasi menjadi nilai *fitness* untuk acuan evaluasi selanjutnya hingga mendapatkan hasil akurasi terbaik secara signifikan. Setelah akurasi didapat, akurasi disimpan dan skema ini dilakukan pada kasus selanjutnya.

Setelah skema dilakukan sebanyak k, dari akurasi yang didapatkan dihitung rata-ratanya per skema. Setelah akurasi rata-rata didapatkan maka sistem ini telah selesai.

Gambaran sistem yang dibangun secara umum dapat dilihat pada Gambar 3-1 sebagai berikut.



Gambar 3-1 Desain Sistem secara Umum

3.2 Kebutuhan Data

Data yang digunakan adalah data dari Kent Ridge yang dapat diunduh pada situs http://leo.ugr.es/elvira/DBCRepository/. Dataset akan dibagi menjadi data *training* dan data *testing*. Dataset terdiri dari *ALL-AML Leukimia*, *Breast Cancer, Colon Tumor, Lung Cancer* dan *Ovarian Cancer*.

Pada *ALL-AML Leukimia* dari dua jenis kanker yaitu *AML* dan *ALL*. Pada *Breast Cancer* terdiri dari dua jenis kanker yaitu *Relapse* dan *Nonrelapse*. Pada *Colon Tumor* terdiri dari dua kelas yaitu kelas *positive* yaitu yang tidak terkena kanker dan kelas *negative* yaitu yang terkena kanker. Pada *Lung Cancer* dari dua jenis kanker yaitu *ADCA* dan *Mesothelioma*. Pada *Ovarian Cancer* terdiri dari dua kelas yaitu kelas *negative* yaitu yang tidak terkena kanker dan kelas *cancer* yaitu yang terkena kanker.

Representasi kelas pada data adalah kelas 1 dan kelas 0. Distribusi data dapat dilihat pada Tabel 3-1 berikut.

Tabel 3-1 Distribusi Data

Dataset	Jumlah *Record*	Jumlah Fitur	Kelas 1	Kelas 0
ALL-AML Leukimia	72	7.129	*AML*	*ALL*
Breast Cancer	97	24.188	*Relapse*	*Nonrelapse*
Colon Tumor	62	2.000	*Positive*	*Negative*
Lung Cancer	181	12.533	*ADCA*	*Mesothelioma*
Ovarian Cancer	253	15.154	*Cancer*	*Normal*

3.3 Pembangunan Sistem

Pada sub bab ini akan memuat bagian dari sistem secara rinci. Sistem terdiri dari bagian normalisasi data, diskritisasi data, seleksi fitur dengan *Information Gain* dan BPSO, proses klasifikasi, perhitungan akurasi dan validasi.

3.3.1 Normalisasi Data

Pada tahap normalisasi, data awal akan dikonversi ke dalam interval [0,1]. Langkah pertama yaitu mencari nilai maksimum dan minimum dari data tersebut. Kemudian langkah kedua melakukan perhitungan data awal dikurangi dengan nilai minimum, lalu dibagi dengan nilai maksimum dikurangi nilai minimum. Langkah-langkah tersebut diulang hingga semua data telah dikonversi.

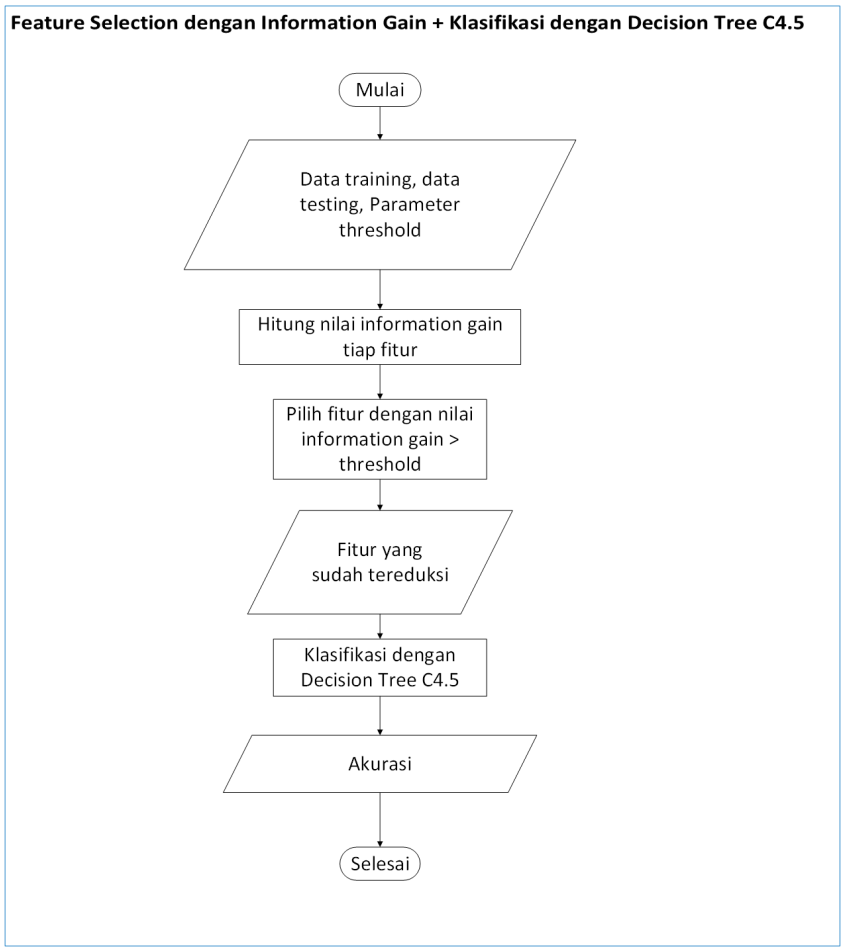
3.3.2 Diskritisasi Data menggunakan K-Means

Pada tahap diskritisasi, data yang telah dinormalisasi akan dikonversi ke dalam bentuk diskrit dengan rentang dari nilai satu hingga nilai k. Langkah pertama yaitu menentukan nilai k, mencari nilai maksimum dan minimum dari data tersebut, menentukan batas iterasi dan menentukan maksimal batas konvergensi turunnya nilai *error*. Pada penelitian ini, nilai k yang digunakan untuk observasi yaitu 3, 5 dan 7, nilai batas iterasi diset 100 dan nilai maksimal batas konvergensi diset 3.

Langkah kedua yaitu inisialisasi nilai *centroid* secara random sebanyak k yang mewakili nilai *cluster*. Kemudian selama iterasi dan batas konstan faktor weight belum mencapai nilai maksimum, dilakukan perhitungan antara data terhadap nilai *centroid*. Jika jarak yang dihitung paling dekat diantara *centroid* maka data tersebut diberi label *cluster* tersebut. Setelah itu nilai *centroid* diperbaharui dan dievaluasi dengan menggunakan SSE hingga hasilnya signifikan. Langkah-langkah tersebut diulang hingga semua data telah dikonversi.

3.3.3 Skema Klasifikasi dengan Information Gain sebagai Seleksi Fitur

Pada skema ini, fitur yang digunakan diseleksi oleh *Information Gain*. Data yang akan dimodelkan adalah data *training*. Pada langkah pertama menginisialisasi nilai *threshold* sebagai acuan untuk memilih atribut yang informatif. Kemudian pada setiap fitur dihitung nilai *information gain* nya. Lalu di cek apakah fitur memiliki nilai *information gain* yang melebihi nilai *threshold,* jika nilainya melebihi maka fitur tersebut digunakan untuk klasifikasi. Setelah dilakukan klasifikasi akan mendapatkan nilai akurasi. Nilai *threshold* yang digunakan untuk observasi yaitu 0.1, 0.2, dan 0.3. Pemilihan nilai *threshold* tersebut didapatkan berdasarkan hasil observasi rata-rata nilai *information gain* yang dihasilkan setiap fitur berada pada rentang batas nilai minimum sebesar 0 dan batas nilai maksimum sebesar 0.5, dengan nilai yang berkumpul pada rentang 0.1 hingga 0.3 lebih besar sehingga nilai 0.1, 0.2 dan 0.3 dipilih menjadi nilai *threshold*. Alur skema ini dapat dilihat pada Gambar 3-2 berikut.



Gambar 3-2 Skema Klasifikasi dengan Information Gain sebagai Seleksi Fitur

3.3.4 Skema Klasifikasi dengan BPSO sebagai Seleksi Fitur

Pada skema ini, fitur yang digunakan diseleksi oleh BPSO. Data yang akan dimodelkan adalah data *training*. Pada langkah pertama menginisialisasi parameter yang diperlukan seperti nilai batas iterasi, nilai batas konstan faktor weight, *cognitive learning* (c_1), *social learning* (c_2), *lower bound velocity* (v_{min}), *upper bound velocity* (v_{max}), *inertia weight* (w), r_1 (nilai random), r_2 (nilai random), jumlah partikel dalam populasi, dan jumlah fitur dalam partikel.

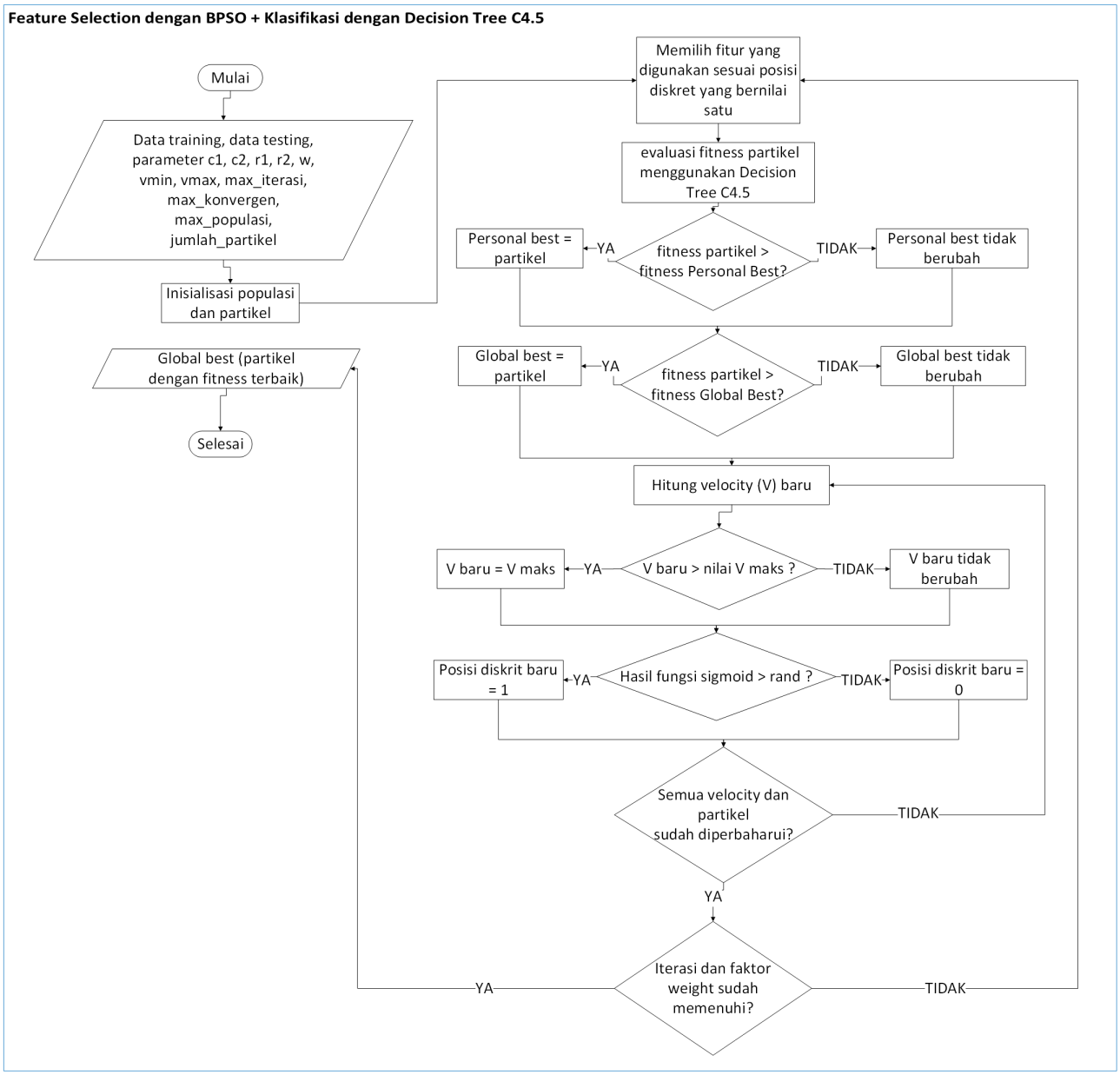
Nilai pada parameter *cognitive learning* (c_1), *social learning* (c_2), *lower bound velocity* (v_{min}), *upper bound velocity* (v_{max}), *inertia weight* (w) menggunakan skema observasi yang dapat dilihat pada Tabel 3-2, sedangkan nilai batas iterasi di set 100, nilai konstan faktor *weight* di set 25, r_1 (nilai random) dan r_2 (nilai random) di set nilai acak dalam interval [0,1], jumlah partikel dalam populasi di set 10, dan jumlah fitur dalam partikel di set sebanyak jumlah fitur kanker.

Tabel 3-2 Observasi Parameter BPSO

Observasi ke-	Parameter				
	cognitive learning **(c_1)**	***social learning*** **(c_2)**	***lower bound velocity*** **(v_{min})**	***upper bound velocity*** **(v_{max})**	***inertia weight*** **(w)**
1	1	1	-1	1	0.1
2	2	2	-2	2	0.2
3	3	3	-3	3	0.3

Pada langkah kedua menginisialisasi populasi yang terdiri dari beberapa partikel. Setiap partikel memiliki nilai *fitness* (akurasi yang dihasilkan dari klasifikasi) dan isi. Bagian isi terdiri dari posisi real (bernilai nilai dalam interval [0,1]), velocity (nilai dalam interval [-4,4]), dan posisi diskrit (bernilai 0 atau 1) sebagai representasi penggunaan fitur. Panjang partikel adalah jumlah seluruh fitur pada data. Kemudian tiap partikel memiliki *personalbest* yaitu partikel terbaik secara individu/lokal. Dan dalam populasi memiliki *globalbest* yaitu partikel terbaik secara kelompok/global.

Pada langkah ketiga, selama iterasi dan konstan faktor weight belum mencapai nilai maksimal dilakukan evaluasi *fitness* per partikel menggunakan data *training* dengan klasifikasi *decision tree* C4.5.Fitur pada data *training* diseleksi berdasarkan inisialisasi posisi diskrit partikel. Kemudian di cek jika nilai *fitness* lebih baik dibandingkan nilai *fitness personalbest* dan *globalbest* maka nilai *fitness personalbest* dan *globalbest* diperbaharui. Kemudian menghitung nilai *velocity* baru tiap partikel, lalu di cek jika *velocity* baru tidak dalam interval nilai v_{max} dan v_{min} maka nilai *velocity* baru diperbaharui. Setelah itu di menghitung fungsi *sigmoid* berdasarkan *velocity*, di cek jika hasil fungsi *sigmoid* lebih besar dibandingkan dengan nilai random dengan interval [0,1] maka posisi diskrit diperbaharui menjadi satu, jika sebaliknya maka posisi diskrit diperbaharui menjadi nol. Langkah-langkah tersebut diulang kembali hingga mencapai iterasi atau konstan faktor *weight* maksimum dan menghasilkan partikel dengan *fitness* terbaik. Alur skema ini dapat dilihat pada Gambar 3-3 berikut.



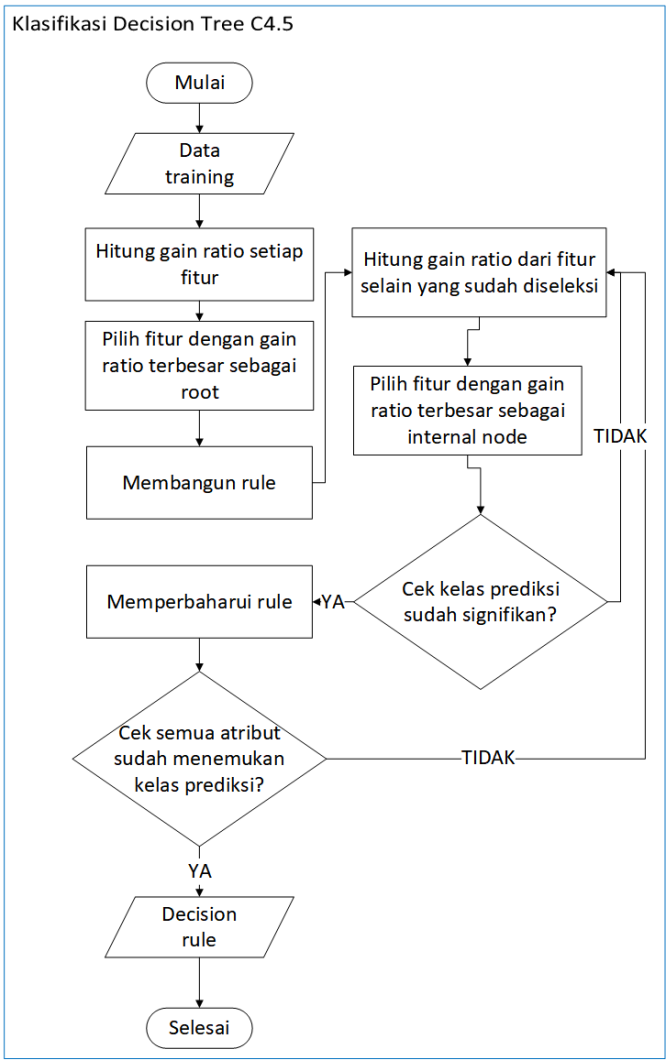
Gambar 3-3 Skema Klasifikasi dengan BPSO sebagai Seleksi Fitur

3.3.5 Klasifikasi menggunakan Decision Tree C4.5

Pada skema ini menggunakan semua fitur pada data *training*. Data yang akan dimodelkan adalah data *training*. Pada langkah pertama, untuk setiap fitur akan dihitung nilai *gain ratio* nya.

Pada langkah kedua, fitur dengan nilai *gain ratio* terbesar akan menjadi *root*. Kemudian dibangun rule berdasarkan atribut pada fitur. Jika tiap atribut pada fitur belum signifikan menemukan

kelas prediksi, maka perlu membangun node selanjutnya. Langkah-langkah tersebut diulang kembali hingga *rule* yang dibangun memenuhi kriteria menemukan kelas prediksi yang signifikan. Alur skema ini dapat dilihat pada Gambar 3-4 berikut.



Gambar 3-4 Klasifikasi Decision Tree C4.5

3.4 Menghitung Akurasi

Perhitungan akurasi digunakan untuk mengukur performansi klasifikasi. Bentuk pengukuran dapat direpresentasikan melalui *Confussion Matrix*, Representasi tersebut dibentuk ke dalam matriks yang dapat dilihat pada Tabel 3-3 berikut [14].

Tabel 3-3 Confussion Matrix

		Prediction Class	
		Class 1	*Class* 0
Actual Class	*Class* 1	*True Class* 1	*False Class* 1
	Class 0	*False Class* 0	*True Class* 0

Perhitungan akurasi menggunakan persamaan yang tertera pada persamaan (3.1) berikut.

$$Accuracy = \frac{True\ Class\ 1 + True\ Class\ 0}{True\ Class\ 1 + False\ Class\ 1 + True\ Class\ 0 + False\ Class\ 0} \quad (3.1)$$

Untuk kasus data *ALL-AML Leukimia* yang memiliki jenis kanker *AML* dan *ALL*, representasi tabel *Confussion Matrix* dapat dilihat pada Tabel 3-4 berikut.

Tabel 3-4 Confussion Matrix untuk ALL-AML Leukimia

		Prediction Class	
		AML (class 1)	*ALL (class 0)*
Actual Class	*AML (class 1)*	*True AML*	*False AML*
	ALL (class 0)	*False ALL*	*True ALL*

Untuk kasus data *Breast Cancer* yang memiliki jenis kanker *Relapse* dan *Nonrelapse*, representasi tabel *Confussion Matrix* dapat dilihat pada Tabel 3-5 berikut.

Tabel 3-5 Confussion Matrix untuk Breast Cancer

		Prediction Class	
		Relapse (class 1)	*Nonrelapse (class 0)*
Actual Class	*Relapse (class 1)*	*True Relapse*	*False Relapse*
	Nonrelapse (class 0)	*False Nonrelapse*	*True Nonrelapse*

Untuk kasus data *Colon Tumor* yang memiliki *Positive* (tidak kanker) dan *Negative* (kanker), representasi tabel *Confussion Matrix* dapat dilihat pada Tabel 3-6 berikut.

Tabel 3-6 Confussion Matrix untuk Colon Tumor

		Prediction Class	
		Positive (class 1)	*Negative (class 0)*
Actual Class	*Positive (class 1)*	*True Positive*	*False Positive*
	Negative (class 0)	*False Negative*	*True Negatives*

Untuk kasus data *Lung Cancer* yang memiliki jenis kanker *ADCA* dan *Mesothelioma*, representasi tabel *Confussion Matrix* dapat dilihat pada Tabel 3-7 berikut.

Tabel 3-7 Confussion Matrix untuk Lung Cancer

		Prediction Class	
		ADCA (class 1)	*Mesothelioma (class 0)*
Actual Class	*ADCA (class 1)*	*True ADCA*	*False ADCA*
	Mesothelioma (class 0)	*False Mesothelioma*	*True Mesothelioma*

Untuk kasus data *Ovarian Cancer* yang memiliki *Normal* (tidak kanker) dan *Cancer* (kanker), representasi tabel *Confussion Matrix* dapat dilihat pada Tabel 3-8 berikut.

Tabel 3-8 Confussion Matrix untuk Ovarian Cancer

		Prediction Class	
		Cancer (class 1)	*Normal (class 0)*
Actual Class	*Cancer (class 1)*	*True Cancer*	*False Cancer*
	Normal (class 0)	*False Normal*	*True Normal*

3.5 Validasi menggunakan K-Fold Cross Validation

Untuk validasi perhitungan akurasi menggunakan *K-Fold Cross Validation*. Pada penelitian ini nilai k di set nilai 5 dan pembagian data *training* dan data *testing* dibagi dengan perbandingan 4:1 yaitu sekitar 80% data dipilih menjadi data *training* dan 20% dipilih menjadi data *testing*. Skema *cross validation* dimulai dengan bagian pertama berperan sebagai data *testing* dan sisa bagian lainnya seperti bagian kedua, ketiga, keempat dan kelima berperan menjadi data *training*, dilanjutkan hingga data kedua, ketiga, keempat dan kelima bergantian menjadi data *testing*.

4. Hasil dan Analisis

Pada bagian hasil dan analisis memuat beberapa sub bab. Sub bab tersebut diantaranya memuat hasil pengujian dan analisa dari hasil pengujian yang telah dilakukan.

4.1 Hasil dan Analisis tanpa Seleksi Fitur terhadap Akurasi

Hasil pengujian klasifikasi tanpa seleksi fitur didapatkan akurasi masing masing untuk data kanker dapat dilihat pada Tabel 4-1 berikut.

Tabel 4-1 Hasil Pengujian tanpa Seleksi Fitur

Dataset	Akurasi rata-rata (%)	Akurasi kelas 1 (%)	Akurasi kelas 0 (%)
AML-ALL Leukimia	47,05	86,75	19,44

Dataset	Akurasi rata-rata (%)	Akurasi kelas 1 (%)	Akurasi kelas 0 (%)
Breast Cancer	50,83	39,28	67,11
Colon Tumor	44,09	83,33	41,66
Lung Cancer	36,27	18,07	100
Ovarian Cancer	60,36	75,68	37,38

Dari hasil pengujian didapatkan rata-rata akurasi untuk keseluruhan data kanker sebesar 47,72 dengan rata-rata akurasi kelas 1 sebesar 60,62 dan rata-rata akurasi kelas 0 sebesar 53,11. Hasil tersebut didapatkan karena terdapat fitur *noise* sehingga *Decision Tree* tidak dapat memodelkan *rule* dengan fitur yang paling optimal yang mengakibatkan rendahnya hasil akurasi.

4.2 Hasil dan Analisis Pengaruh Seleksi Fitur terhadap Akurasi

Hasil pengujian pada penelitian ini dibedakan sesuai nilai observasi. Nilai observasi tersebut adalah nilai k pada diskritisasi dengan nilai 3, 5 dan 7, sedangkan parameter yang digunakan ketika seleksi fitur menggunakan *Information Gain* yaitu nilai *threshold* sebesar 0.1, 0.2, dan 0.3 dan parameter yang digunakan ketika seleksi fitur menggunakan BPSO yaitu menerapkan skenario 1, 2 dan 3 yang dapat dilihat pada Tabel 3-2.

4.2.1 Diskritisasi Data menjadi 3 Cluster

Hasil pengujian diskritisasi menggunakan K-Means dengan nilai K = 3 dapat dilihat pada Tabel 4-2 berikut.

Tabel 4-2 Hasil Pengujian Data Diskritisasi K = 3 Secara Umum

Dataset	Rata-Rata Akurasi Hasil Klasifikasi + Seleksi Fitur (%)					
	Information Gain			BPSO		
	Threshold > 0.1	Threshold > 0.2	Threshold > 0.3	Skenario ke -1	Skenario ke -2	Skenario ke -3
AML-ALL Leukimia	48,05	64,84	67,16	100	100	100
Breast Cancer	48,61	19,74	0	100	100	100
Colon Tumor	51	72,67	46	100	100	97,14
Lung Cancer	83,86	73,33	70,88	99,39	100	99,39
Ovarian Cancer	45,84	51,98	52,16	100	100	100
Rata-rata	**55,47**	**56,51**	**47,24**	**99,87**	**100**	**99,3**

Detail hasil pengujian beserta akurasi per kelas uji pada diskritisasi menggunakan K-*Means* dengan nilai K = 3 dapat dilihat pada Tabel 4-3 berikut.

Tabel 4-3 Hasil Pengujian Data Diskritisasi K = 3 Secara Detail

Dataset		Rata-Rata Akurasi Hasil Klasifikasi + Seleksi Fitur (%)					
		Information Gain			BPSO		
Nama Data	Jenis Kelas	Threshold > 0.1	Threshold > 0.2	Threshold > 0.3	Skenario ke -1	Skenario ke -2	Skenario ke -3
AML-ALL Leukimia	ALL	63,89	78,84	85,6	100	100	100
	AML	37,14	74,41	68,57	100	100	100
Breast Cancer	Relapse	64,7	0	0	100	100	100
	Nonrelapse	37,8	21,05	0	100	100	100
Colon Tumor	Positive	53,33	56,25	33,33	100	100	83,33
	Negative	49,16	76,67	60	100	100	100
	ADCA	74,76	100	81,25	100	100	100

Dataset		Rata-Rata Akurasi Hasil Klasifikasi + Seleksi Fitur (%)					
		Information Gain			BPSO		
Nama Data	Jenis Kelas	Threshold > 0.1	Threshold > 0.2	Threshold > 0.3	Skenario ke -1	Skenario ke -2	Skenario ke -3
Lung Cancer	Mesothelioma	87,5	55,56	64,99	96,42	100	96,42
Ovarian Cancer	Cancer	30,91	38,59	38,55	100	100	100
	Normal	51,72	51,72	53,72	100	100	100

Dari hasil pengujian tersebut didapatkan rata-rata akurasi dari skema klasifikasi dengan *Information Gain* sebagai seleksi fitur adalah 53,07. Sedangkan rata-rata akurasi dari skema klasifikasi dengan BPSO sebagai seleksi fitur adalah 99,72. Rata-rata akurasi total berada dalam rentang rata-rata per kelasnya. Namun beberapa data pada skema tertentu rata-rata akurasi totalnya tidak berada dalam rentang rata-rata per kelasnya.

Pada data *AML-ALL Leukimia* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.2, rata-rata akurasi total sebesar 64,84 namun rata-rata per kelas nya adalah 78,84 untuk kelas 1 dan 74,41 untuk kelas 0. Hal ini dikarenakan tidak imbangnya kemerataan data kelas 1 dan kelas 0 pada saat melakukan skema *k-fold cross validation* yang kelima sehingga mempengaruhi akurasi rata-rata untuk kelas 0.

Pada data *Breast Cancer* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.2, rata-rata akurasi total sebesar 19,74 namun rata-rata per kelas nya adalah nol untuk kelas 1 dan 21,05 untuk kelas 0. Hal ini dikarenakan fitur yang direduksi terlalu sedikit sehingga pembangunan *rule* tidak signifikan dan menyebabkan kegagalan prediksi untuk kelas 1 juga akurasi yang kecil untuk kelas 0.

Pada data *Ovarian Cancer* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.2, rata-rata akurasi total sebesar 51,98 namun rata-rata per kelas nya adalah 38,59 untuk kelas 1 dan 51,72 untuk kelas 0. Hal ini dikarenakan tidak imbangnya kemerataan data kelas 1 dan kelas 0 pada saat melakukan skema *k-fold cross validation* yang kedua sehingga mempengaruhi akurasi rata-rata untuk kelas 0.

Pada data *AML-ALL Leukimia* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.3, rata-rata akurasi total sebesar 67,16 namun rata-rata per kelas nya adalah 85,6 untuk kelas 1 dan 68,57 untuk kelas 0. Hal ini dikarenakan tidak imbangnya kemerataan data kelas 1 dan kelas pada saat melakukan skema *k-fold cross validation* yang ketiga sehingga mempengaruhi akurasi rata-rata untuk kelas 0.

Pada data *Breast Cancer* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.3, rata-rata akurasi total sebesar nol dengan rata-rata per kelas nya masing masing bernilai nol untuk kelas 1 dan kelas 0. Hal ini dikarenakan tidak ada data yang dipilih oleh seleksi fitur dimana tidak ada fitur yang memiliki *threshold* bernilai lebih dari 0.3.

4.2.2 Diskritisasi Data menjadi 5 Cluster

Hasil pengujian diskritisasi menggunakan K-*Means* dengan nilai K = 5 dapat dilihat pada Tabel 4-4 berikut.

Tabel 4-4 Hasil Pengujian Data Diskritisasi K = 5 Secara Umum

Dataset	Rata-Rata Akurasi Hasil Klasifikasi + Seleksi Fitur (%)					
	Information Gain			BPSO		
	Threshold > 0.1	Threshold > 0.2	Threshold > 0.3	Skenario ke -1	Skenario ke -2	Skenario ke -3
AML-ALL Leukimia	34,48	66,35	73,72	100	100	100
Breast Cancer	78,19	58,01	0	100	100	100
Colon Tumor	77,78	71,56	44,28	97,78	100	100
Lung Cancer	42,17	36,36	86,77	100	100	100
Ovarian Cancer	60,46	60,09	60,09	100	100	99,42

Dataset	Rata-Rata Akurasi Hasil Klasifikasi + Seleksi Fitur (%)					
	Information Gain			BPSO		
	Threshold > 0.1	Threshold > 0.2	Threshold > 0.3	Skenario ke -1	Skenario ke -2	Skenario ke -3
Rata-rata	**58,61**	**58,47**	**52,97**	**99,56**	**100**	**99,88**

Detail hasil pengujian beserta akurasi per kelas uji pada diskritisasi menggunakan K-Means dengan nilai K = 5 dapat dilihat pada Tabel 4-5 berikut.

Tabel 4-5 Hasil Pengujian Data Diskritisasi K = 5 Secara Detail

Dataset		Rata-Rata Akurasi Hasil Klasifikasi + Seleksi Fitur (%)					
		Information Gain			BPSO		
Nama Data	Jenis Kelas	Threshold > 0.1	Threshold > 0.2	Threshold > 0.3	Skenario ke -1	Skenario ke -2	Skenario ke -3
AML-ALL Leukimia	ALL	57,5	100	80,5	100	100	100
	AML	0	33,14	78,29	100	100	100
Breast Cancer	Relapse	85,29	48,12	0	100	100	100
	Nonrelapse	58,22	70,95	0	100	100	100
Colon Tumor	Positive	88,75	74,16	66,66	93,75	100	100
	Negative	64,58	63,68	30	100	100	100
Lung Cancer	ADCA	43,2	33,6	62,22	100	100	100
	Mesothelioma	35,5	54,39	95	100	100	100
Ovarian Cancer	Cancer	75,86	75,23	75,23	100	100	100
	Normal	37,38	37,38	37,38	100	100	98,27

Dari hasil pengujian tersebut didapatkan rata-rata akurasi dari skema klasifikasi dengan *Information Gain* sebagai seleksi fitur adalah 56,68. Sedangkan rata-rata akurasi dari skema klasifikasi dengan BPSO sebagai seleksi fitur adalah 99,81. Rata-rata akurasi total berada dalam rentang rata-rata per kelasnya. Namun beberapa data pada skema tertentu rata-rata akurasi totalnya tidak berada dalam rentang rata-rata per kelasnya dan terdapat data pada skema yang mendapatkan rata-rata akurasi sebesar nol.

Pada data *AML-ALL* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.1, rata-rata akurasi total sebesar 34,48 namun rata-rata per kelas nya adalah 57,5 untuk kelas 1 dan nol untuk kelas 0. Hal ini dikarenakan fitur yang direduksi kurang representatif sehingga pembangunan *rule* tidak signifikan dan menyebabkan kegagalan prediksi untuk kelas 0.

Pada data *AML-ALL Leukimia* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.3, rata-rata akurasi total sebesar 73,72 namun rata-rata per kelas nya adalah 80,55 untuk kelas 1 dan 78,29 untuk kelas 0. Hal ini dikarenakan tidak imbangnya kemerataan data kelas 1 dan kelas 0 pada saat melakukan skema *k-fold cross validation* yang kelima sehingga mempengaruhi akurasi rata-rata untuk kelas 0.

Pada data *Breast Cancer* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.3, rata-rata akurasi total sebesar nol dengan rata-rata per kelas nya masing masing bernilai nol untuk kelas 1 dan kelas 0. Hal ini dikarenakan tidak ada data yang dipilih oleh seleksi fitur dimana tidak ada fitur yang memiliki *threshold* bernilai lebih dari 0.3.

4.2.3 Diskritisasi Data menjadi 7 Cluster

Hasil pengujian diskritisasi menggunakan K-*Means* dengan nilai K = 7 dapat dilihat pada Tabel 4-6 berikut.

Tabel 4-6 Hasil Pengujian Data Diskritisasi K = 7 Secara Umum

Dataset	Rata-Rata Akurasi Hasil Klasifikasi + Seleksi Fitur (%)					
	Information Gain			BPSO		
	Threshold > 0.1	Threshold > 0.2	Threshold > 0.3	Skenario ke -1	Skenario ke -2	Skenario ke -3
AML-ALL Leukimia	50,86	54,68	72,87	100	100	100
Breast Cancer	67,25	34,57	10,47	100	100	100
Colon Tumor	49,78	44,16	55,15	96,36	93,8	93,86
Lung Cancer	57,41	34,84	35,97	100	100	100
Ovarian Cancer	86,18	86,6	66,32	100	100	100
Rata-rata	**62,29**	**50,97**	**43,37**	**99,27**	**98,77**	**98,72**

Detail hasil pengujian beserta akurasi per kelas uji pada diskritisasi menggunakan K-Means dengan nilai K = 7 dapat dilihat pada Tabel 4-7 berikut.

Tabel 4-7 Hasil Pengujian Data Diskritisasi K = 7 Secara Detail

Dataset		Rata-Rata Akurasi Hasil Klasifikasi + Seleksi Fitur (%)					
		Information Gain			BPSO		
Nama Data	Jenis Kelas	Threshold > 0.1	Threshold > 0.2	Threshold > 0.3	Skenario ke -1	Skenario ke -2	Skenario ke -3
AML-ALL Leukimia	ALL	70,37	77,78	100	100	100	100
	AML	57,29	58,23	64,24	100	100	100
Breast Cancer	Relapse	57,29	18,18	0	100	100	100
	Nonrelapse	58,75	42,91	11,57	100	100	100
Colon Tumor	Positive	67,5	74,16	62,91	94,4	77,78	95,83
	Negative	40,78	17,5	57,5	96	96	91
Lung Cancer	ADCA	48,3	0	20	100	100	100
	Mesothelioma	100	38,18	37,34	100	100	100
Ovarian Cancer	Cancer	87,06	79,92	66,37	100	100	100
	Normal	83,09	86,75	55,14	100	100	100

Dari hasil pengujian didapatkan rata-rata akurasi dari skema klasifikasi dengan *Information Gain* sebagai seleksi fitur adalah 53,88. Sedangkan rata-rata akurasi dari skema klasifikasi dengan BPSO sebagai seleksi fitur adalah 98,93. Rata-rata akurasi total berada dalam rentang rata-rata per kelasnya. Namun beberapa data pada skema tertentu rata-rata akurasi totalnya tidak berada dalam rentang rata-rata per kelasnya.

Pada data *AML-ALL Leukimia* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.1, rata-rata akurasi total sebesar 50,86 namun rata-rata per kelas nya adalah 70,37 untuk kelas 1 dan 57,29 untuk kelas 0. Hal ini dikarenakan tidak imbangnya kemerataan data kelas 1 dan kelas 0 pada saat melakukan skema *k-fold cross validation* yang ketiga sehingga mempengaruhi akurasi rata-rata untuk kelas 0.

Pada data *Breast Cancer* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.1, rata-rata akurasi total sebesar 67,25 namun rata-rata per kelas nya adalah 57,29 untuk kelas 1 dan 58,75 untuk kelas 0. Hal ini dikarenakan tidak imbangnya kemerataan data kelas 1 dan kelas 0 pada saat melakukan skema *k-fold cross validation* yang kedua sehingga mempengaruhi akurasi rata-rata untuk kelas 0.

Pada data *AML-ALL Leukimia* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.2, rata-rata akurasi total sebesar 54,68 namun rata-rata per kelas nya adalah 72,78 untuk kelas 1 dan 58,23 untuk kelas 0. Hal ini dikarenakan tidak imbangnya kemerataan data kelas 1 dan kelas 0 pada saat melakukan skema *k-fold cross validation* yang kedua sehingga mempengaruhi akurasi rata-rata untuk kelas 0.

Pada data *Lung Cancer* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.2, rata-rata akurasi total sebesar 34,84 namun rata-rata per kelas nya adalah nol untuk kelas 1 dan 38,18 untuk kelas 0. Hal ini dikarenakan tidak imbangnya kemerataan data kelas 1 dan kelas 0 untuk saat melakukan skema *k-fold cross validation* yang kelima dan kelas prediksi pada kelas 1 gagal menebak kelas kanker karena pemodelan *rule* yang dihasilkan kurang mewakili kelas kankernya sehingga mempengaruhi akurasi rata-rata untuk kelas 1.

Pada data *Breast Cancer* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.3, rata-rata akurasi total sebesar 10,47 namun rata-rata per kelas nya adalah nol untuk kelas 1 dan 11,57 untuk kelas 0. Hal ini dikarenakan fitur yang direduksi terlalu sedikit sehingga pembangunan *rule* tidak signifikan dan menyebabkan kegagalan prediksi untuk kelas 1 juga akurasi yang kecil untuk kelas 0.

Pada data *Colon Tumor* dengan skema *information gain* sebagai seleksi fitur yang *threshold* nya bernilai lebih dari 0.3, rata-rata akurasi total sebesar 55,14 namun rata-rata per kelas nya adalah 62,91 untuk kelas 1 dan 57,5 untuk kelas 0. Hal ini dikarenakan tidak imbangnya kemerataan data kelas 1 dan kelas 0 untuk pada melakukan skema *k-fold cross validation* yang kedua sehingga mempengaruhi akurasi rata-rata untuk kelas 0.

4.3 Hasil dan Analisis berdasarkan Jumlah Cluster pada Diskritisasi terhadap Akurasi

Dari hasil pengujian didapatkan rata-rata akurasi tiap kasus k yang dapat dilihat pada Tabel 4-8 berikut.

Tabel 4-8 Hasil Rata-Rata Akurasi Tiap Kasus K Diskritisasi

Skema	Akurasi berdasarkan Diskritisasi Data yang Diterapkan (%)			Rata-Rata Akurasi (%)
	K = 3	K = 5	K = 7	
Information Gain + Decision Tree C4.5	53,07	56,68	53,88	54,54
BPSO + *Decision Tree C4.5*	99,72	99,81	98,93	99,48

Dari tabel tersebut, didapatkan bahwa akurasi skema *Information Gain + Decision Tree C4.5* dan skema BPSO + *Decision Tree C4.5* tidak stabil seiring dengan bertambahnya jumlah k yang digunakan. Dengan data k = 5 memperoleh akurasi yang lebih baik dibandingkan dengan jumlah k lainnya.

Pada data k = 3, pada skema *Information Gain + Decision Tree C4.5* akurasinya bertambah secara signifikan dengan bertambahnya nilai *threshold* dan mendapatkan akurasi terbaik dengan nilai *threshold* = 0.3. Pada skema BPSO + *Decision Tree C4.5* akurasi tidak stabil dengan bertambahnya skenario dan mendapatkan akurasi terbaik dengan nilai skenario ke - 2.

Pada data k = 5, pada skema *Information Gain + Decision Tree C4.5* akurasinya menurun secara signifikan dengan bertambahnya nilai *threshold* dan mendapatkan akurasi terbaik dengan nilai *threshold* = 0.1. Pada skema BPSO + *Decision Tree C4.5* akurasi tidak stabil dengan bertambahnya skenario dan mendapatkan akurasi terbaik dengan nilai skenario ke - 2.

Pada data k = 7, pada skema *Information Gain + Decision Tree C4.5* akurasinya bertambah secara signifikan dengan bertambahnya nilai *threshold* dan mendapatkan akurasi terbaik dengan nilai *threshold* = 0.1. Pada skema BPSO + *Decision Tree C4.5* akurasi tidak stabil dengan bertambahnya skenario dan mendapatkan akurasi terbaik dengan nilai skenario ke - 1.

Dari observasi tersebut didapatkan bahwa nilai k diskritisasi berpengaruh secara signifikan terhadap akurasi skema *Information Gain + Decision Tree C4.5* dan tidak berpengaruh secara signifikan terhadap akurasi skema BPSO + *Decision Tree C4.5*. Skenario ke-2 mendapatkan hasil akurasi yang stabil terhadap skema BPSO + *Decision Tree C4.5*, yaitu dengan parameter *cognitive learning* (c_1) = 2*, social learning* (c_2) = 2, *lower bound velocity* (v_{min}) = -2, *upper bound velocity* (v_{max}) = 2, *inertia weight* (w) = 0.2.

Penerapan seleksi fitur *Information Gain* terbukti mampu menyeleksi data yang informatif dan berpengaruh dalam kontribusi pemodelan *rule* juga meminimalisir data *noise*. Akurasi skema BPSO + *Decision Tree C4.5* stabil pada penggunaan parameter di skenario ke-2 dan meminimalisir data *noise* sehingga mendapatkan hasil akurasi yang lebih baik.

5. Kesimpulan

Dari penelitian yang dilakukan, performansi dari sistem yang dibangun untuk lima data yang digunakan mendapatkan rata-rata akurasi 54% untuk skema klasifikasi dengan *Information Gain* sebagai seleksi fitur dan

rata-rata akurasi 99% untuk skema klasifikasi dengan BPSO sebagai seleksi fitur. Pemilihan nilai k pada diskritisasi didapatkan hasil yang terbaik adalah nilai k = 5. Pemilihan skenario penggunaan parameter pada BPSO didapatkan hasil yang terbaik adalah skenario kedua dimana parameternya terdiri dari *cognitive learning* (c_1) = 2, *social learning* (c_2) = 2, *lower bound velocity* (v_{min}) = -2, *upper bound velocity* (v_{max}) = 2, dan *inertia weight* (w) = 0.2.

Hasil akurasi dari skema klasifikasi dengan BPSO sebagai seleksi fitur lebih baik dibandingkan dengan hasil akurasi dari skema klasifikasi dengan *Information Gain* sebagai seleksi fitur. Penggunaan jumlah fitur yang memiliki banyak *noise* berpengaruh terhadap klasifikasi yang dilakukan. Pengaruh seleksi fitur terhadap klasifikasi berperan penting dalam menghindari data *noise* untuk memodelkan *rule* yang akurat pada prediksi data uji. Dengan penerapan BPSO sebagai seleksi fitur mampu mencari fitur yang paling signifikan.

Faktor lain yang mempengaruhi hasil akurasi didapatkan dari menentukan metode diskritisasi, penentuan nilai k pada skema pengujian *K-Fold Cross Validation*, penentuan batas iterasi dan konstan faktor *weight*, dan observasi inisialisasi parameter yang dilakukan. Menganalisis lebih lanjut mengenai penentuan metode diskritisasi, batas iterasi dan nilai konstan faktor *weight*, inisialisasi parameter dengan nilai observasi lain dapat menentukan performansi sistem yang dibangun untuk meningkatkan hasil akurasi yang dihasilkan.

Daftar Pustaka

[1] Yip, Wai-Ki., Amin, Samir B., and Li, Cheng. 2011. *Chapter 10: A Survey of Classification Techniques for Microarray Data Analysis*, Springer Handbooks of Computational Statistics.

[2] NCBI. Microarrays. 2007. [online] Available at https://www.ncbi.nlm.nih.gov/probe/docs/techmicroarray/ [Accessed 1 May 2018].

[3] Speed, T. (Ed). 2003. *Statistical Analysis of Gene Expression Microarray Data (Chap. 3)*. New York: Chapman & Hall/CRC.

[4] Ulfrenborg, Benjamin., Klinga-Levan, Karin., and Olsson, Bjorn. 2013. *Classification of Tumor Samples from Expression Data Using Decison Trunks*. Libertas Academica.

[5] P.K., Ammu and V., Preeja. 2013. *Review on Feature Selection Techniques of DNA Microarray Data*, International Journal of Computer Applications, Vol.61, No.12.

[6] Pashaei, Elnaz., Ozen, Mustafa., and Aydin, Nizamettin. 2015. *A Novel Gene Selection Algorithm for Cancer Identification based on Random Forest and Particle Swarm Optimization*. IEEE.

[7] Li, Yong., et al. 2010. *Comparative study of discretization methods of microarray data for inferring transcriptional regulatory networks*. BMC Bioinformatics.

[8] Chen, Kun-Huang., et al. 2014. *Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm*. BMC Bioinformatics.

[9] Tsai, Meng-Chang., Chen, Kun-Huang., Su, Chao-Ton., and Lin, Hung-Chun. 2012. *An application of PSO algorithm and decision tree for medical problem*. 2nd ICS Bali, pp. 13-14.

[10] Tu, Chung-Jui., Chuang, Li-Yeh., Chang, Jun-Yang., and Hong, Cheng. 2007. *Feature Selection using PSO-SVM*. IAENG International Journal of Computer Science.

[11] Dai, Jian J., Lieu, Linh., and Rocke, David. 2006. *Dimension Reduction for Classification with Gene Expression Microarray Data*. Statistical Application in Genetics and Molecular Biology, Vol.5, Issue 1, 2006.

[12] Eberhart, Russell., and Kennedy, James. 1995. *A New Optimizer using Particle Swarm Theory*. Sixth International Symposium on Micro Machine and Human Science, IEEE.

[13] Qi, Jianpeng., et al. 2016. *K-Means: An Effective and Efficient K-Means Clustering Algorithm*. International Conferences on Big Data and Cloud Computing, Social Computing and Networking, Sustainable Computing and Communications, IEEE.

[14] Guilet, Fabrice., and Hamilton, Howard J., *Quality Measures in Data Mining*, Springer, 2007.