METODE EKSTRAKSI FITUR PADA PENGKLASIFIKASIAN DATA *MICROARRAY* BERBASIS INFORMASI PASANGAN GEN

Rully Soelaiman[1,2], Sheila Agustianty [1], Yudhi Purwananto[1], dan I. K. Eddy Purnama[2]

[1]Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.

[2]Program Pascasarjana, Jurusan Teknik Elektro, Fakultas Teknologi Industri, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.
rully@is.its.ac.id

Abstrak

Pengenalan teknologi DNA *microarray* membuat perolehan data *microarray* menjadi lebih mudah. Hal ini semakin memicu persoalan tentang bagaimana cara terbaik dalam mengekstraksi dan memilih fitur dari data yang berdimensi besar tersebut. Metode-metode terdahulu mengabaikan adanya hubungan antargen sehingga memungkinkan hilangnya informasi penting yang tersimpan dalam suatu gen pada saat ekstraksi fitur. Meskipun berbagai macam metode telah digunakan, pengembangan metode ekstraksi dan seleksi fitur dari data *microarray* yang lebih *powerful* dan efisien masih diperlukan untuk meningkatkan performa klasifikasi kanker. Dalam penelitian ini diimplementasikan sebuah metode dalam melakukan ekstraksi fitur dari data *microarray* yang memanfaatkan model klasifikasi berbasis informasi pasangan gen, yaitu pasangan gen yang memiliki perbedaan signifikan pada dua jenis *sampel tissue*. Hasil uji coba terhadap dua data *microarray* menunjukkan bahwa fitur hasil ekstraksi menggunakan metode ini dapat meningkatkan performa klasifikasi. Bahkan akurasi 100% dapat diperoleh pada uji coba terhadap data *lymphoma*.

Kata kunci : *algoritma genetika, data microarray, ekstraksi fitur, informasi pasangan gen, klasifikasi kanker.*

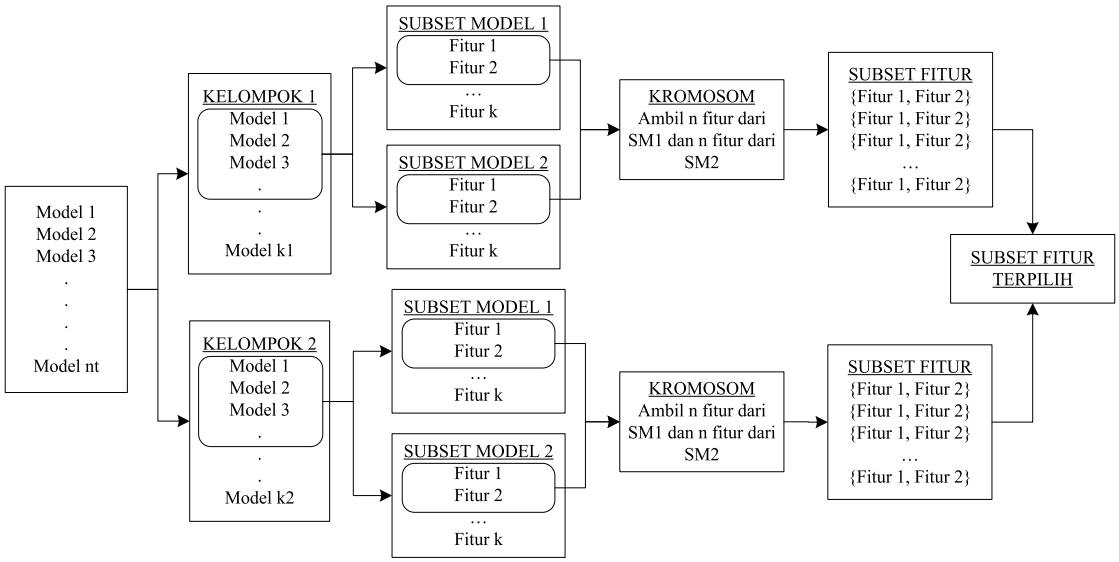
1. Pendahuluan

Pengenalan teknologi DNA *microarray* membuat perolehan data *microarray* semakin mudah. Hal ini memicu persoalan tentang bagaimana cara terbaik dalam melakukan ekstraksi dan seleksi fitur dari data yang berdimensi sangat besar tersebut. Berbagai macam metode telah diusulkan dalam melakukan ekstraksi dan seleksi fitur dari data *microarray*. Akan tetapi, metode-metode terdahulu mengabaikan adanya hubungan antar gen (*interrelation*) sehingga memungkinkan hilangnya informasi penting yang tersimpan pada suatu gen pada saat ekstraksi fitur. Hal ini mengakibatkan metode tersebut masih belum dapat membantu para ilmuwan biologi untuk menemukan informasi penting pada gen, terutama mengenai masalah klasifikasi kanker [1].

Meskipun berbagai macam metode telah digunakan untuk mengekstrak dan memilih fitur dari data *microarray*, pengembangan metode ekstraksi dan seleksi fitur dari data *microarray* yang lebih kuat dan efisien untuk meningkatkan performa klasifikasi kanker masih menjadi persoalan yang perlu diselesaikan [1].

Oleh karena itu dalam penelitian ini diusulkan sebuah implementasi metode ekstraksi fitur yang memperlakukan informasi pasangan gen, yaitu pasangan gen yang memiliki kolerasi tinggi pada satu jenis sampel jaringan (*tissue sampel)* dan memiliki perbedaan yang signifikan pada tipe *tissue sampel* lain, sebagai suatu kesatuan yang digunakan untuk ekstraksi fitur dari model klasifikasi berbasis informasi pasangan gen.

Secara umum informasi pasangan gen didapatkan melalui pembangunan kemungkinan-kemungkinan model klasifikasi yang memiliki tingkat koefisien korelasi yang tinggi pada satu kelas dan rendah di kelas lainnya secara acak. Sebanyak n *top ranked* model klasifikasi dengan tingkat akurasi yang lebih tinggi dibagi menjadi dua kelompok. Kelompok 1 berisi model klasifikasi yang berkorelasi tinggi di kelas 1. Kelompok 2 berisi model klasifikasi yang berkorelasi tinggi di kelas 2. Kemudian *top-ranked* model klasifikasi dari masing-masing kelompok digunakan untuk membuat dua subset model klasifikasi. Subset model 1 berisi fitur-fitur yang memiliki peran besar untuk menjadi gen pertama (*g1*) pada informasi pasangan gen. Subset model 2



Gambar 1. Desain model sistem secara umum diawali dari proses pembangkitan model klasifikasi sebanyak *nt* sampai didapatkan subset fitur terpilih.

berisi fitur-fitur yang memiliki peran besar untuk menjadi gen kedua (*g2*) pada informasi pasangan gen. Proses seleksi subset fitur diterapkan menggunakan algoritma genetika untuk mendapatkan subset paling optimal di masing-masing kelasnya. Subset fitur terbaik akan terpilih di akhir sistem. Dalam subset tersebut terdapat pasangan-pasangan informasi gen yang terlibat beserta tingkat akurasi pengklasifikasiannya. Algoritma sistem ini merupakan usulan dari [1]. Desain model sistem yang diimplementasikan dapat dilihat pada Gambar 1.

2. Informasi Pasangan Gen

Dua macam gen *g1* dan *g2* yang diuji ke dalam dua jenis jaringan sampel (sebagai contoh, jaringan normal dan jaringan kanker) dapat disebut dengan informasi pasangan gen apabila memenuhi karakteristik sebagai berikut:

- Keduanya memiliki tingkat korelasi tinggi ke kelas 1 (atau kelas 2).
- *Expression level* dari *g1* dan atau *g2* memiliki perubahan yang signifikan yang membuat kedua jenis sampel dapat dipisahkan.

Apabila terdapat pasangan gen *g1* dan *g2* yang diuji pada dua tipe *tissue sampel* yang berbeda, relasi di antara keduanya dapat digambarkan menggunakan model regresi linier ketika keduanya memiliki korelasi tinggi pada sebuah tipe *tissue sampel* (tipe *tissue sampel* untuk selanjutnya disebut dengan kelas). Contohnya ketika pasangan gen tersebut memiliki korelasi tinggi pada kelas pertama, artinya bahwa pada kelas pertama nilai ekspresi *g1* dapat secara akurat diprediksi dari nilai ekspresi *g2* menggunakan model linier. Model yang didapat dari kelas pertama tersebut masih dapat digunakan untuk memprediksi nilai ekspresi *g1* dari *g2* pada kelas kedua ketika relasi antara *g1* dan *g2* memiliki perbedaan yang signifikan pada kelas kedua, yang menghasilkan nilai bias yang lebih besar antara nilai hasil prediksi dengan nilai sesungguhnya. Nilai besar atau kecilnya bias inilah yang menunjukkan sampel berasal dari kelas kedua atau pertama, sehingga dua macam kelas dapat dibedakan berdasarkan nilai prediksi bias. Berdasarkan pemikiran tersebut diperkenalkan model klasifikasi berbasis informasi pasangan gen [2].

3 Model Klasifikasi

3.1 Pembuatan Model Klasifikasi

Diasumsikan dua jenis sampel diuji dalam percobaan *microarray*, *k* merupakan jumlah gen, *n1* dan *n2* ($n = n1 + n2$) merupakan jumlah sampel pada kelas 1 dan 2. Data *microarray* dapat direpresentasikan ke dalam bentuk matriks $Y=(y_{ip})_{k\times n1}$, $X=(x_{iq})_{k\times n2}$, di mana $y_{ip}(x_{iq})$ menunjukkan *expression level* dari gen ke-*i* pada sampel ke-*p*(*q*) yang dimiliki oleh kelas ke-1 (atau kelas ke-2).
Jika diberikan gen ke-*i* dan ke-*j* (informasi pasangan gen) yang sangat berkolerasi pada kelas 1, maka untuk sampel ke-*p* dari kelas 1, y_{ip} dapat diprediksi melalui model regresi berikut.

$$\hat{y}_{ijp} = \hat{\beta}_{ij0} + \hat{\beta}_{ij1} y_{jp} \quad 1 \le p \le n1 \tag{1}$$

$\hat{\beta}_{ij0}$ dan $\hat{\beta}_{ij1}$ diestimasikan dari dataset, (y_{i1}, y_{j1}), (y_{i2}, y_{j2}), ..., (y_{in1}, y_{jn1}) menggunakan metode *least square* [3]. Nilai residual didapatkan dengan rumus $e_{ijp} = |y_{ip} - \hat{y}_{ijp}|$ yakni selisih antara nilai yang ditinjau y_{ip}

dengan nilai yang diprediksi $\hat{y}_{ijp}$. Untuk semua sampel pada kelas 1, dapat diperoleh subset E1 seperti pada model (2).

$$E1_{ij} = \{e_{ijp} \mid e_{ijp} = \mid y_{ip} - \hat{\beta}_{ij0} - \hat{\beta}_{ij1} y_{jp} \mid, 1 \le p \le n1\} \quad (2)$$

Untuk sampel ke-*q* pada kelas 2, tetap digunakan model (1) untuk memprediksi nilai x_{iq} seperti yang terlihat pada model (3) berikut:

$$\hat{x}_{ijq} = \hat{\beta}_{ij0} + \hat{\beta}_{ij1} x_{jq} \quad 1 \le q \le n2 \quad (3)$$

dengan nilai residual: $eijq = |xiq - \hat{x}ijq|$. Untuk semua sampel dari kelas 2, dapat diperoleh model (4).

$$E2_{ij} = \{e_{ijq} \mid e_{ijq} = \mid x_{iq} - \hat{\beta}_{ij0} - \hat{\beta}_{ij1} x_{jp} \mid, 1 \le q \le n2\} \quad (4)$$

Aturan klasifikasi yang digunakan model regresi untuk meminimalisasi kesalahan dalam membedakan elemen menjadi dua subset $E1ij$ dan $E2ij$ diawali dengan mendefinisikan fungsi berikut:

$$f_i(e) = count(\{e_{ijp} < e, e_{ijp} \in E1_{ij}, n1 \ge p \ge 1\} \cup \{e_{ijq} < e, e_{ijq} \in E2_{ij}, n2 \ge q \ge 1\}) \quad (5)$$

dengan *e* merupakan bilangan *real*, count(.) menunjukkan jumlah elemen yang terdapat dalam subset. Jika $e = e1$, $fi(e) = \max(fi(e))$, maka nilai *threshold ed* dapat diperoleh melalui rumus:

$$e_d = (\max(\{ e_{ijp} \mid e_{ijp} \le e_1, e_{ijp} \in E1_{ij}) + \min(\{ e_{ijq} \mid e_{ijq} \le e_1, e_{ijq} \in E2_{ij}\}))/2. \quad (6)$$

Oleh karena itu ketika dipilih sampel secara acak dari keseluruhan sampel, maka *expression level* gen ke-*i* dan ke-*j* dalam masing-masing sampel adalah w_i dan w_j, proses klasifikasi dapat dilakukan dengan aturan berikut.

Ditempatkan ke dalam kelas 1 jika:

$$| w_i - \hat{w}_i | \le e_d \text{ , yakni, } | w_i - \hat{\beta}_{ij0} - \hat{\beta}_{ij1} w_j | \le e_d \quad (7)$$

dan ditempatkan ke dalam kelas 2 jika sebaliknya.

3.2 Evaluasi Model Klasifikasi

Terdapat sejumlah besar model klasifikasi yang diusulkan dalam data *microarray* tetapi banyak di antaranya tidak relevan dengan fungsi klasifikasi. Dengan demikian diperlukan adanya penyaringan terhadap model klasifikasi yang tidak baik tersebut dan memilih model dengan perform,a klasifikasi yang lebih baik [1].

Karena untuk mengevaluasi seluruh kemungkinan model klasifikasi akan membutuhkan waktu komputasi yang sangat lama, model klasifikasi yang akan dievaluasi hanyalah model klasifikasi yang memenuhi kriteria yaitu melebihi nilai *threshold* koefisien korelasi ($\hat{\partial}_h$). Sehingga model klasifikasi yang nilai koefisien korelasi [4] di kelas 1 dan kelas 2 berada di bawah nilai *threshold* tidak akan dievaluasi.

Terdapat tiga metode yang sering digunakan dalam mengevaluasi performa model klasifikasi [5]. Apabila keseluruhan sampel digunakan sebagai data *training* sekaligus data *testing* maka tingkat akurasi klasifikasi menunjuk pada penggunaan metode yang disebut *within sampel classification accuracy* (WSCA).

Metode WSCA diterapkan dalam dua proses. Proses pertama adalah digunakan untuk mengevaluasi tiap model klasifikasi hasil pembangkitan sedangkan yang kedua adalah digunakan untuk mengevaluasi fitur subset pada saat masuk ke langkah seleksi fitur. Perhitungan nilai WSCA secara mudah dapat diperoleh melalui hasil bagi jumlah elemen yang benar diklasifikasikan dengan jumlah elemen yang diklasifikasikan.

4. Seleksi Subset Fitur

Seleksi fitur merupakan teknik yang digunakan untuk memilih fitur terbaik dari sekian banyak fitur yang tersedia pada sebuah data. Proses seleksi fitur yang diterapkan pada kasus implementasi metode ekstraksi fitur dalam pengklasifikasian data *microarray* berbasis informasi pasangan gen ini digunakan untuk memilih subset model klasifikasi dalam hal ini di dalamnya terdapat pasangan-pasangan informasi gen pilihan terbaik yang dapat mengklasifikasikan dengan tingkat akurasi yang lebih tinggi.

Metode yang digunakan dalam proses ini adalah algoritma genetika. Algoritma genetika dipilih karena terbukti merupakan metode optimasi evolusioner yang efektif [6,7]. Pada kasus ini algoritma genetika digunakan untuk mencari subset fitur yang optimal dari *top ranked* subset model klasifikasi pada kelas 1 (atau 2) [1].

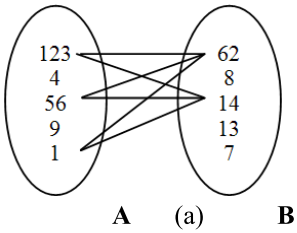
Terdapat beberapa parameter yang perlu dijelaskan dalam mengimplementasikan algoritma genetika, diantaranya adalah:

4.1 Definisi individu

Individu atau yang sering disebut dengan kromosom dinyatakan dalam representasi biner. Satu individu terdiri atas dua gen, gen pertama merepresentasikan *top ranked* subset model yang memiliki peran penting untuk menjadi gen pertama dalam suatu informasi pasangan gen sedangkan gen

kedua merepresentasikan *top ranked* subset model yang memiliki peran penting untuk menjadi gen kedua dalam suatu informasi pasangan gen. Misal subset A dan B pada Gambar 2a di bawah ini merupakan subset model klasifikasi yang berisi fitur, subset A berisi *top ranked* 5 fitur terbaik yang yang berperan di suatu kelas untuk menjadi gen pertama sedangkan subset B berisi *top ranked* 5 fitur terbaik yang berperan di kelas yang sama untuk menjadi gen kedua.

Representasi kromosom biner dari subset di atas pada algoritma genetika dapat dilihat pada Gambar 2b. Tiap bit dari kromosom ini merupakan representasi satu fitur. Bit 1 dan 0 merepresentasikan hadir atau tidaknya fitur tersebut dalam sebuah individu. Dalam satu kromosom terdapat satu titik potong P yang berfungsi membedakan fitur dari subset A (fitur untuk gen pertama) dengan fitur dari subset B (fitur untuk gen kedua).



123	4	56	9	1	62	8	14	13	7
1	0	1	0	1	1	0	1	0	0

P

(b)

Gambar 2. Contoh representasi kromosom. (a) Dua subset model klasifikasi yang akan dikonversi menjadi kromosom. (b) Contoh representasi kromosom biner dari gabungan subset A dan subset B.

4.2 Inisialisasi populasi

Populasi awal dibangkitkan secara random dari *top ranked LC* model klasifikasi yang telah didapat pada langkah sebelumnya dengan asumsi bahwa apabila didekodekan ke dalam desimal tiap individu dalam populasi awal ini memiliki angka yang unik untuk menghindari adanya individu yang kembar dan tidak dibolehkan ada angka 0 untuk menghindari kesalahan. Populasi yang dipakai berukuran 50. Semakin tinggi nilai populasi akan membutuhkan waktu eksekusi yang lebih lama.

4.3 Pemilihan individu untuk dipasangkan

Cara seleksi yang digunakan adalah sebagai berikut: dua kromosom terbaik akan langsung masuk ke generasi selanjutnya (proses elitisme), sedangkan 48 kromosom sisanya diberi bobot sesuai dengan *relative fitness* (14) untuk generasi *parent* (*probabilistically*) atau biasa disebut dengan *Roulette Wheel*

$$f_{ri} = \frac{f_i}{\sum_{k=1}^{48} f_k}, \quad 1 \le i \le 48 \tag{14}$$

dengan f_{ri} adalah nilai *relative fitness* dari koromosom ke-*i*, f_i adalah nilai *fitness* dari kromosom ke-*i*.

4.4 Mengkombinasikan individu

Kombinasi individu ini merupakan hasil pindah silang (*crossover*) dua individu yang terpilih dalam aliran populasi yang bertujuan untuk mencetak individu baru pada generasi berikutnya. *Crossover* yang digunakan pada kasus ini adalah *two point crossover* dengan probabilitas terjadinya *crossover* adalah 0.9. Probabilitas ini termasuk ke dalam *range* parameter *control* yang telah diusulkan oleh [8].

4.5 Mutasi

Mutasi adalah perubahan gen yang bukan berasal dari *parent*. Proses mutasi dilakukan dengan cara penggantian dengan nilai inversinya, bit 0 menjadi 1 atau bit 1 menjadi 0. Proses ini dilakukan secara acak pada posisi tertentu pada individu-individu yang terpilih untuk dimutasikan. Probabilitas terjadinya mutasi pada kasus ini adalah 0.05. Probabilitas ini juga termasuk ke dalam *range* parameter *control* yang telah diusulkan oleh [8].

4.6 Kriteria berhenti (stopping criteria)

Kriteria berhenti yang ditetapkan adalah ketika jumlah generasi mencapai lebih dari 200 dan kenaikan nilai *fitness* (optimal *fitness* value) lebih rendah dari 0.0001 dalam 20 putaran.

4.7 Fungsi fitness

Tujuan dari algoritma genetika ini adalah memilih subset model klasifikasi yang paling optimal untuk mengekstraksi subset fitur yang dapat memberikan performa klasifikasi yang lebih baik dengan menggunakan gen yang lebih sedikit. Oleh karena itu, subset model klasifikasi dievaluasi melalui performa subset fitur yang diekstraksi dari subset model yang bersangkutan.

4.7.1 Ekstraksi fitur dari subset model klasifikasi

Fungsi *fitness* digunakan untuk menghitung performa dari tiap individu. Telah dijelaskan sebelumnya bahwa fungsi *fitness* dihitung dengan menggunakan performa subset fitur yang diekstraksi dari subset model klasifikasi. Oleh karena itu, di sini akan dijelaskan terlebih dahulu mengenai metode ekstraksi fitur dari subset model klasifikasi untuk mendapatkan subset fitur. Setelah individu

didekodekan menjadi subset model klasifikasi, selanjutnya kita harus mendaftarkan semua kemungkinan informasi pasangan gen yang dapat dibuat dari subset model tersebut. Pasangan informasi gen ini yang akan digunakan untuk proses ekstraksi subset fitur.

Setiap informasi pasangan gen (gen ke-*i* dan gen ke-*j*) akan memproyeksikan *expression values* gen ke-*i* pada dua jenis sampel ke dalam dua subset E1*i* dan E2*i*. Untuk *m1* pasangan informasi gen: (i_1,j_1), ..., (i_{m1},j_{m1}) yang memiliki tingkat kolerasi tinggi pada kelas 1, dapat dibuat *m1* model regresi liniernya yang memproyeksikan *expression values* gen *m1* $(i_1, i_2, ..., i_{m1})$ pada dua jenis sampel ke dalam *m1* pasang subset: $(E1_{i1j1},E2_{i1j1})$, $(E1_{i2j2},E2_{i2j2})$, ..., $(E1_{im1jm1},E2_{im1jm1})$.

Untuk sampel ke-*p* dari kelas 1 dan sampel ke-*q* dari kelas 2 diterapkan rumus (8) dan (9):

$$\mu 1_p = \frac{1}{m1}\sum_{l=1}^{m1} e_{i_l j_l p}, \quad e_{i_l j_l p} \in E1_{i_l j_l} \tag{8}$$

$$\mu 2_q = \frac{1}{m1}\sum_{l=1}^{m1} e_{i_l j_l q}, \quad e_{i_l j_l q} \in E2_{i_l j_l} \tag{9}$$

dipilih $\mu 1_p$ sebagai fitur dari sampel ke-*p* dan $\mu 2_q$ sebagai fitur dari sampel ke-*q*. Jadi untuk semua sampel dalam data *microarray* akan menghasilkan subset fitur (10).

$$U = \{\mu 1_1, \mu 1_2, ..., \mu 1_{n1}, \mu 2_1, \mu 2_2, ..., \mu 2_{n2}\} \tag{10}$$

Dengan cara yang sama, untuk *m2* pasangan informasi gen: (i'_1,j'_1), ..., (i'_{m2},j'_{m2}) yang memiliki tingkat kolerasi tinggi pada kelas 2, dapat dibuat *m2* model regresi liniernya yang memproyeksikan *expression values* gen $m2(i'_1, i'_2, ..., i'_{m2})$ pada dua jenis sampel ke dalam *m2* pasang subset: $(E1_{i'1j'1},E2_{i'1j'1}),(E1_{i'2j'2},E2_{i'2j'2}),...,(E1_{i'm1j'm1},E2_{i'm1j'm1})$.

Untuk sampel ke-*p* dari kelas 1 dan sampel ke-*q* dari kelas 2, melalui rumus (11) dan (12):

$$\mu 1'_p = \frac{1}{m2}\sum_{l=1}^{m2} e_{i'_l j'_l p}, \quad e_{i'_l j'_l p} \in E1_{i'_l j'_l} \tag{11}$$

$$\mu 2'_q = \frac{1}{m2}\sum_{l=1}^{m2} e_{i'_l j'_l q}, \quad e_{i'_l j'_l q} \in E2_{i'_l j'_l} \tag{12}$$

dapat dipilih juga $\mu 1'_p$ sebagai fitur dari sampel ke-*p* dan $\mu 2'_q$ sebagai fitur dari sampel ke-*q*. Jadi untuk semua sampel dalam data *microarray* akan menghasilkan subset fitur (13).

$$U' = \{\mu 1'_1, \mu 1'_2, ..., \mu 1'_{n1}, \mu 2'_1, \mu 2'_2, ..., \mu 2'_{n2}\} \tag{13}$$

4.7.2 Rumusan Fungsi Fitness

Proses ekstraksi fitur akan menghasilkan subset fitur untuk tiap subset model klasifikasi. Subset fitur ini akan digunakan untuk menghitung nilai *fitness* karena performa subset fitur merepresentasikan performa dari subset model klasifikasi (atau individu dalam algoritma genetika).

Untuk proses perhitungan *fitness*, ada 3 poin ukuran (*terms*) yang dapat digunakan untuk mengukur performa dari subset fitur, yaitu:

1. Tingkat akurasi dari klasifikasi subset fitur,
2. Batasan (*margin*) dari *classifier* yang dilatih oleh subset fitur,
3. Jumlah gen yang terlibat dalam subset fitur.

Jika subset fitur yg diekstrak dari 2 subset model klasifikasi mempunyai nilai akurasi yang sama, subset fitur dipilih adalah yang dapat melatih *classifier* dengan margin yang lebih besar. Jika nilai akurasi dan margin sama, subset dengan jumlah gen lebih sedikit yang akan dipilih. Untuk mengkombinasikan ketiga ukuran di atas, kita gunakan fungsi *fitness* (15) berikut:

$$fitness = \begin{cases} Acc + 10^{-4}\dfrac{(LC - Fn)}{LC} & if \quad Acc < 1 \\ Acc + 10^{-2}\dfrac{Mg}{MM} + 10^{-4}\dfrac{(LC - Fn)}{LC} & if \quad Acc = 1 \end{cases} \tag{15}$$

dengan Acc = WSCA dari subset fitur, *Fn* = jumlah pasangan gen yang terdapat dalam subset fitur, *LC* = panjang kromosom, Mg/MM = *magnitude* dari margin *classifier*.

WSCA dihitung menggunakan aturan klasifikasi yang sama pada diskriminasi subset E1 dan E2. Sebagai contoh, WSCA subset fitur $U = \{\mu 1_1, \mu 1_2, \ldots, \mu 1_{n1}, \mu 2_1, \mu 2_2, \ldots, \mu 2_{n2}\}$, dihitung menggunakan aturan berikut:

- Pilih sampel secara random dari total sampel (nilai fitur valuenya μ_i).
- Tempatkan sampel pada kelas 1 jika $|\mu_i| \le \mu_d$, kelas 2 jika sebaliknya, di mana μ_d merupakan nilai optimal *threshold* yang dapat meminimalisir error dalam membedakan elemen ke dalam 2 subset: $\{\mu 1_1, \mu 1_2, \ldots, \mu 1_{n1}\}$, $\{\mu 2_1, \mu 2_2, \ldots, \mu 2_{n2}\}$.

Jika Acc = 100%, maka Mg = $\min(\mu 2_1, \mu 2_2, \ldots, \mu 2_{n2}) - \max(\mu 1_1, \mu 1_2, \ldots, \mu 1_{n1})$, dan MM berdasarkan model (10) dapat dihitung menggunakan rumus berikut:

$$MM = abs\left(\frac{1}{n1}\sum_{i=1}^{n1}\mu 1_i - \frac{1}{n2}\sum_{i=1}^{n2}\mu 2_i\right). \tag{16}$$

Range nilai akurasi antara 0.5 sampai dengan 1, Mg/MM antara 0 sampai dengan 1, sedangkan *term* ketiga berkisar antara 0 sampai dengan 0.0001.

5. Uji Coba dan Analisis

Dataset *microarray* yang digunakan sebagai uji coba sistem ini adalah data *colon cancer* [9] dan *diffuse large B cell lymphoma* (DLBCL) [10]. *Colon cancer* memiliki 2000 fitur dan 62 sampel (22 merupakan jaringan normal dan 40 merupakan jaringan kanker), sedangkan DLBCL memiliki 4026 fitur dan 42 sampel (21 merupakan jaringan *germinal center B-like* DLBCL dan 21 merupakan jaringan *actived B-like* DLBCL).

Terdapat tiga data masukan yang dibutuhkan dalam sistem, yaitu dataset, *threshold* koefisien korelasi (∂_h), panjang subset fitur (*LC*). Data keluaran dari sistem ini adalah subset fitur hasil ekstraksi beserta informasi pasangan gen yang terdapat di dalamnya dan nilai *within sampel classification accuracy* (WSCA) dari subset fitur tersebut.

Dua skenario akan diterapkan dalam uji coba sistem Skenario pertama adalah penambahan nilai threshold ∂_h yang digunakan sebagai *filtering* koefisien korelasi sedangkan skenario kedua adalah penambahan nilai *LC* yang merupakan panjang subset fitur yang ingin dihasilkan. Masing-masing skenario akan diterapkan pada dua data uji coba yang telah dijelaskan sebelumnya, yaitu data *Colon Cancer* dan data DLBCL.

5.1. Skenario 1

Pada skenario pertama ini dilakukan penambahan nilai *threshold* (∂_h) kemudian dilakukan analisis pengaruhnya terhadap tingkat akurasi yang dapat dilakukan oleh subset fitur yang terpilih beserta jumlah pasangan gen yang terlibat dalam subset fitur. Selain itu dilakukan juga analisis terhadap tingkat akurasi yang dapat dilakukan oleh masing-masing informasi pasangan gen tertinggi dalam subset fitur yang terpilih.

5.1.1. Data *Colon Cancer*

Uji coba pertama, kedua, ketiga, dan keempat secara berturut-turut ∂_h bernilai 0.6, 0.7, 0.8, 0.9 dengan panjang subset fitur *LC* yang sama yaitu 10. Hasil keluaran berupa *optimal fitness value* (OFV), nilai *within sampel classification accuracy* (WSCA), beserta banyaknya pasangan gen yang terlibat dalam subset yang terpilih (*Fn*) terlihat pada Tabel 1.

Tabel 1. Perbandingan Terhadap Penambahan Nilai Threshold Pada Data Colon Cancer

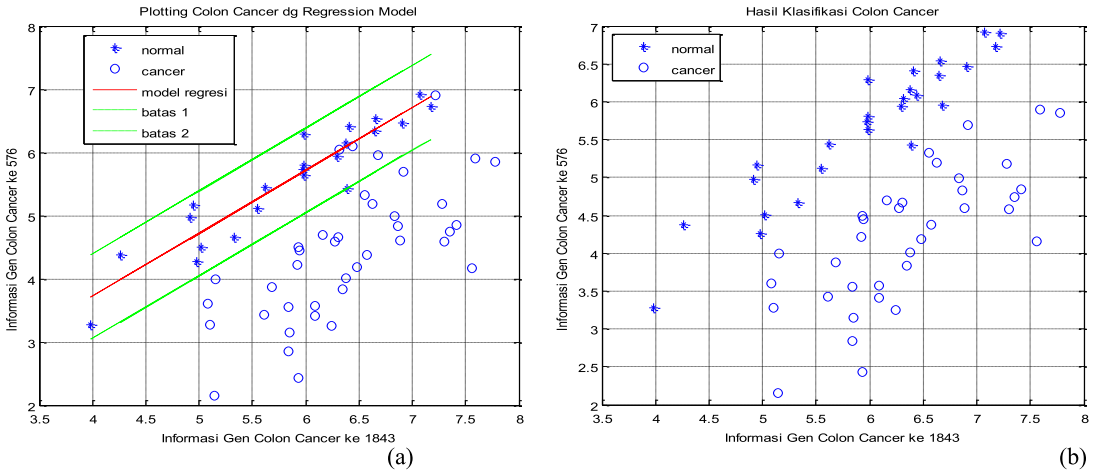
∂_h	OFV	WSCA (%)	Fn
0.6	0.93555	93.548	3
0.7	0.95168	95.161	3
0.8	0.95167	95.161	4
0.9	0.93557	93.548	1

Tabel 2. Daftar Informasi Pasangan Gen yang Terlibat dalam Subset Fitur Data Colon Cancer

∂_h	Gen1	Gen2	WSCA(%)	Koefisien korelasi
0.6	897	656	83.871	0.64246
	1042	656	82.258	0.65043
	1887	656	70.968	0.67675
0.7	1635	576	91.935	0.78924
	183	1771	80.645	0.74948
	1042	1771	64.516	0.79216
0.8	1843	576	93.548	0.94064
	1042	1106	87.097	0.8162
	1843	1590	79.032	0.8735
	1843	1106	72.581	0.81655
0.9	1843	576	93.548	0.94064

Empat uji coba di atas dilakukan dengan menggunakan hasil pembangkitan model klasifikasi secara random yang sama yaitu pada saat ∂_h 0.6. Hal ini dilakukan untuk dapat menganalisis hasil subset fitur yang didapat ketika terdapat kenaikan nilai ∂_h saat proses seleksi. Dari hasil yang terlihat pada Tabel 1 tersebut dapat dianalisis bahwa penambahan nilai ∂_h berpengaruh terhadap nilai WSCA subset fitur yang terpilih beserta jumlah pasangan gen yang terlibat di dalamnya (*Fn*).

Semakin tinggi nilai ∂_h yang diinputkan maka semakin tinggi pula nilai WSCA subset fitur yang didapat, selain itu jumlah pasangan informasi gen yang terlibat dalam subset fitur juga semakin banyak. Namun terjadi penurunan ketika nilai ∂_h yang diinputkan terlalu tinggi. Hal ini dikarenakan proses random model klasifikasi di awal dilakukan dengan menggunakan ∂_h 0.6 sehingga pasangan-pasangan informasi gen yang dapat melebihi nilai ∂_h 0.9 pada saat proses seleksi menjadi semakin sedikit.



(a) (b)

Gambar 3. *Plotting* nilai ekspresi data *colon cancer* menggunakan informasi pasangan gen ke 1843 dan ke 576. (a) Data sebelum diklasifikasikan. Proses klasifikasi dilakukan dengan aturan yaitu data yang berada dalam dua garis batas berwarna hijau akan diklasifikasikan ke dalam kelas normal, dan diklasifikasikan ke dalam kelas *cancer* jika sebaliknya. (b) Hasil setelah diklasifikasikan.

Pencapaian nilai WSCA subset fitur tertinggi adalah ketika ∂_h bernilai antara 0.7 sampai 0.8. Tingkat akurasi pengklasifikasian sebesar 95,16% didapat dengan melibatkan sebanyak 3 sampai 4 informasi pasangan gen.

Untuk mengetahui lebih detail mengenai subset fitur yang terpilih, berikut akan ditampilkan daftar informasi pasangan gen yang terlibat pada masing-masing subset fitur yang terpilih pada keempat uji coba.

Daftar informasi pasangan gen pada uji coba pertama sampai keempat dapat dilihat pada Tabel 2 yaitu berupa pasangan gen (gen 1 dan gen 2), nilai *within sampel classification accuracy* (WSCA), beserta nilai koefisien korelasi pasangan gen yang bersangkutan. Dari tabel tersebut terlihat bahwa pasangan informasi gen yang paling dominan dan memiliki tingkat akurasi pengklasifikasian paling tinggi adalah pasangan 1843 dan 576 yaitu dengan tingkat akurasi 93.55% dan koefisien korelasi 0.94. Visualisasi gambar proses klasifikasi menggunakan informasi pasangan gen ini dapat dilihat pada Gambar 3.

Tingginya koefisien korelasi pada satu kelas terlihat pada berkumpulnya sampel-sampel yang berada dalam kelas normal sedemikian sehingga sampel-sampel tersebut memiliki sebuah jarak yang dapat digunakan untuk membedakannya dengan kelas kanker.

5.1.2. Data DLBCL

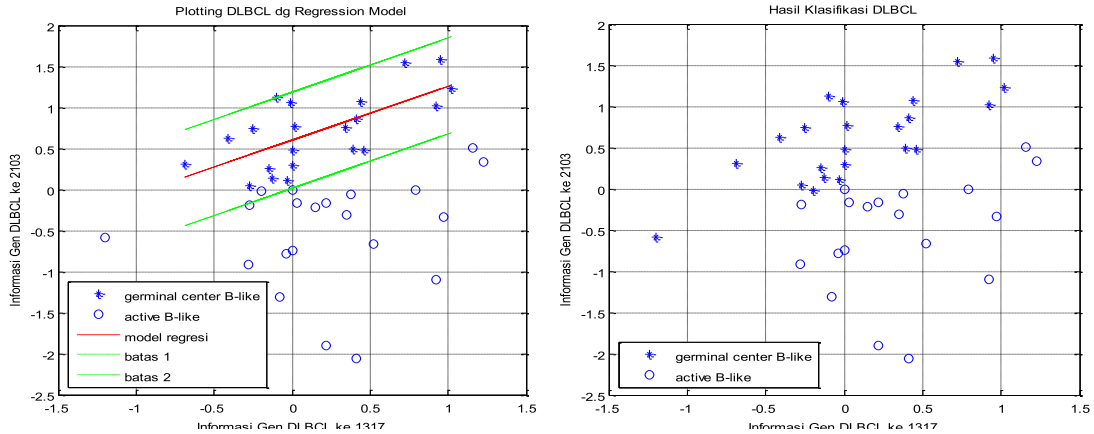
Uji coba pertama, kedua, ketiga, dan keempat secara berturut-turut ∂_h bernilai 0.5, 0.6, 0.7, 0.8 dengan panjang subset fitur *LC* yang sama yaitu 10. Hasil keluaran berupa *optimal fitness value* (OFV), nilai *within sampel classification accuracy* (WSCA), nilai *magnitude classifier* (Mg/MM), beserta banyaknya pasangan gen yang terlibat dalam subset yang terpilih (*Fn*) terlihat pada Tabel 3.

Empat uji coba di atas dilakukan dengan menggunakan hasil pembangkitan model klasifikasi secara random yang sama yaitu pada saat ∂_h 0.5, hal ini dilakukan untuk dapat menganalisis hasil subset fitur yang didapat ketika terdapat kenaikan nilai ∂_h saat proses seleksi. Hasil yang terlihat pada Tabel 3 tersebut dapat dianalisis bahwa penambahan nilai ∂_h berpengaruh terhadap nilai WSCA subset fitur yang terpilih beserta jumlah pasangan gen yang terlibat di dalamnya (*Fn*).

Semakin tinggi nilai ∂_h yang digunakan maka semakin tinggi pula nilai WSCA subset fitur yang didapat., Selain itu jumlah pasangan informasi gen yang terlibat dalam subset fitur juga semakin banyak. Namun terjadi penurunan ketika nilai ∂_h yang diinputkan terlalu tinggi. Hal ini dikarenakan proses random model klasifikasi di awal dilakukan dengan menggunakan ∂_h 0.5 sehingga pasangan-pasangan informasi gen yang dapat melebihi nilai ∂_h 0.8 pada saat proses seleksi menjadi semakin sedikit.

Untuk mengetahui lebih detail mengenai subset fitur yang terpilih, berikut akan ditampilkan daftar informasi pasangan gen yang terlibat pada masing-masing subset fitur yang terpilih pada keempat uji coba.

Daftar informasi pasangan gen pada uji coba pertama dapat dilihat pada Tabel 4 yaitu berupa pasangan gen (gen 1 dan gen 2), nilai *within sampel classification accuracy* (WSCA), beserta nilai koefisien korelasi pasangan gen yang bersangkutan. Dari Tabel 4 terlihat bahwa pasangan informasi gen yang paling dominan dan memiliki tingkat akurasi pengklasifikasian paling tinggi adalah pasangan 1317 dan 2103 yaitu dengan tingkat akurasi 95.24% dan koefisien korelasi 0.68. Visualisasi gambar



(a) (b)

Gambar 4. Plotting nilai ekspresi data DLBCL menggunakan informasi pasangan gen ke 1317 dan ke 2103. (a) Data sebelum diklasifikasikan. Proses klasifikasi dilakukan dengan aturan yaitu data yang berada dalam dua garis batas berwarna hijau akan diklasifikasikan ke dalam kelas *germinal center B-liked*, dan diklasifikasikan ke dalam kelas *active B-like* jika sebaliknya. (b) Hasil setelah diklasifikasikan.

proses klasifikasi menggunakan informasi pasangan gen ini dapat dilihat pada Gambar 4.

Tingginya koefisien korelasi pada satu kelas terlihat pada berkumpulnya sampel-sampel yang berada dalam kelas *germinal center B-like*, sedemikian sehingga sampel-sampel tersebut memiliki sebuah jarak yang dapat digunakan untuk membedakannya dengan kelas *active B-like*.

Tabel 3. Perbandingan Terhadap Penambahan Nilai *Threshold* pada Data DLBCL

∂_h	OFV	WSCA (%)	Mg/MM	Fn
0.5	1.0006	100	0.048912	3
0.6	1.0013	100	0.12287	3
0.7	0.92865	92.857	-	2
0.8	0.85723	85.714	-	1

Tabel 4. Daftar Informasi Pasangan Gen yang Terlibat dalam Subset Fitur Data DLBCL

∂_h	Gen1	Gen2	WSCA(%)	Koefisien korelasi
0.5	1276	1166	90.476	0.70846
	1312	2932	85.714	0.62865
	2136	1411	85.714	0.51807
0.6	1317	2103	95.238	0.6804
	3132	958	85.714	0.72082
	1642	989	83.333	0.86304
0.7	1642	989	85.714	0.86304
	3132	958	85.714	0.72082
0.8	1642	989	83.333	0.86304

5.2. Skenario 2

Pada skenario kedua dilakukan penambahan panjang subset fitur (*LC*) kemudian dilakukan analisis pengaruhnya terhadap tingkat akurasi yang dapat dilakukan oleh subset fitur yang terpilih beserta jumlah pasangan gen yang terlibat dalam subset fitur. Selain itu dilakukan juga analisis terhadap tingkat akurasi yang dapat dilakukan oleh masing-masing informasi pasangan gen tertinggi dalam subset fitur yang terpilih. Hasil uji coba yang dilakukan pada dua dataset akan dijelaskan pada sub bab berikut ini.

5.2.1. Data *Colon Cancer*

Uji coba pertama, kedua, ketiga, sampai ketujuh secara berturut-turut *LC* bernilai 10, 20, 30, 40, 50, 100, 150 dengan nilai *threshold* ∂_h yang sama yaitu 0.8. Hasil keluaran berupa *optimal fitness value* (OFV), nilai *within sampel classification accuracy* (WSCA), beserta banyaknya pasangan gen yang terlibat dalam subset yang terpilih (*Fn*) terlihat pada Tabel 5.

Tabel 5. Perbandingan Terhadap Penambahan Panjang Subset Fitur pada Data Colon Cancer

LC	OFV	WSCA (%)	Fn
10	0.95168	95.161	4
20	0.9678	96.774	8
30	0.98393	98.387	11
40	0.98394	98.387	14
50	0.96781	96.774	17
100	0.9678	96.774	158
150	0.95159	95.161	187

Tujuh uji coba di atas dilakukan dengan menggunakan hasil pembangkitan model klasifikasi secara random yang sama yaitu pada saat ∂_h 0.8. Hal ini dilakukan untuk dapat menganalisis hasil subset fitur yang didapat ketika terdapat kenaikan panjang subset fitur (*LC*). Dari hasil yang terlihat pada Tabel 5 tersebut dapat dianalisis bahwa penambahan nilai *LC* berpengaruh terhadap nilai WSCA subset fitur yang terpilih beserta jumlah pasangan gen yang terlibat di dalamnya (*Fn*).

Semakin tinggi nilai *LC* yang diinputkan maka semakin tinggi pula nilai WSCA subset fitur yang didapat, selain itu jumlah pasangan informasi gen yang terlibat dalam subset fitur juga semakin banyak. Namun terjadi penurunan nilai WSCA ketika nilai *LC* yang diinputkan terlalu besar. Hal ini dikarenakan semakin besar nilai *LC*, maka semakin banyak pula kemungkinan pasangan gen (model klasifikasi) yang dibentuk pada saat proses seleksi fitur menggunakan algoritma genetika. Terlalu banyaknya model klasifikasi yang dibentuk tersebut itulah yang memicu penurunan nilai WSCA yang didapat. Nilai *Fn* berbanding lurus dengan besarnya nilai *LC* yang diinputkan, seperti yang dijelaskan sebelumnya bahwa semakin besar nilai *LC* maka semakin banyak pula pasangan gen yang terlibat dalam suatu subset fitur.

Pencapaian nilai WSCA subset fitur tertinggi adalah ketika *LC* bernilai antara 30 sampai 40. Tingkat akurasi pengklasifikasian sebesar 98,4% didapat dengan melibatkan sebanyak 11 sampai 14 informasi pasangan gen.

5.2.2. Data DLBCL

Uji coba pertama, kedua, ketiga, sampai ketujuh secara berturut-turut *LC* bernilai 10, 20, 30, 40, 50, 100, 150 dengan nilai *threshold* ∂_h yang sama yaitu 0.8. Hasil keluaran berupa *optimal fitness value* (OFV), nilai *within sampel classification accuracy* (WSCA), nilai *magnitude classifier* (Mg/MM), beserta banyaknya pasangan gen yang terlibat dalam subset yang terpilih (*Fn*) terlihat pada Tabel 6.

Tabel 6.
Perbandingan Terhadap Penambahan Panjang Subset Fitur pada Data DLBCL

LC	OFV	WSCA (%)	Mg/MM	Fn
10	1.0002	100	0.015547	2
20	1.0017	100	0.1661	4
30	1.0031	100	0.30246	5
40	1.0034	100	0.33624	5
50	1.005	100	0.49269	9
100	1.0058	100	0.57021	17
150	1.0056	100	0.54951	26

Tujuh uji coba di atas dilakukan dengan menggunakan hasil pembangkitan model klasifikasi secara random yang sama yaitu pada saat ∂_h 0.8. Hal ini dilakukan untuk dapat menganalisis hasil subset fitur yang didapat ketika terdapat kenaikan panjang subset fitur (*LC*). Dari hasil yang terlihat pada Tabel 6 tersebut dapat dianalisis bahwa penambahan nilai *LC* berpengaruh terhadap nilai WSCA subset fitur yang terpilih beserta jumlah pasangan gen yang terlibat di dalamnya (*Fn*). Karena dari ketujuh uji coba di atas menghasilkan nilai WSCA subset fitur yang sama, analisis akan dilakukan terhadap nilai *magnitude* (Mg/MM) dan banyaknya informasi pasangan gen yang terlibat di dalamnya (*Fn*).

Semakin tinggi nilai *LC* yang diinputkan maka semakin tinggi pula nilai magnitude yang didapat, selain itu jumlah pasangan informasi gen yang terlibat dalam subset fitur juga semakin banyak. Namun nilai *magnitude* mengalami penurunan ketika nilai *LC* yang diinputkan terlalu tinggi. Hal ini dikarenakan semakin tinggi nilai *LC*, semakin banyak pula kemungkinan pasangan gen (model klasifikasi) yang dibentuk pada saat proses seleksi fitur menggunakan algoritma genetika. Terlalu banyaknya model klasifikasi yang dibentuk tersebut itulah yang memicu penurunan nilai magnitude yang didapat. Nilai *Fn* berbanding lurus dengan besarnya nilai *LC* yang diinputkan, seperti yang dijelaskan sebelumnya bahwa semakin besar nilai *LC* maka semakin banyak pula pasangan gen yang terlibat dalam suatu subset fitur.

6. Kesimpulan

Setelah dilakukan uji coba dan analisis hasil terhadap aplikasi yang telah dibuat maka dapat diambil kesimpulan sebagai berikut:

a. Suatu informasi pasangan gen yang memiliki karakteristik nilai koefisien korelasi yang sangat tinggi pada suatu kelas, dan nilai dari kedua gen tersebut memiliki perbedaan yang signifikan sehingga dapat digunakan untuk membedakan antara kelas satu dengan lainnya terbukti dapat dijadikan sebagai model untuk proses pengklasifikasian.
b. Ekstraksi fitur yang memanfaatkan model klasifikasi berbasis informasi pasangan gen dapat menghasilkan deretan fitur (subset fitur) yang mampu meningkatkan akurasi proses klasifikasi.
c. Gabungan informasi pasangan gen yang terdapat dalam subset fitur dapat membentuk gabungan model klasifikasi. Gabungan dari beberapa model klasifikasi yang digunakan untuk mengklasifikasikan ini dapat memberikan tingkat akurasi yang lebih tinggi daripada hanya menggunakan satu model klasifikasi saja.
d. Semakin tinggi nilai parameter *threshold* koefisien korelasi dan panjang subset fitur yang digunakan tidak menjamin menghasilkan subset fitur yang baik. Berdasarkan hasil uji coba dapat disimpulkan bahwa subset fitur yang optimal dapat dihasilkan dengan menggunakan parameter *threshold* untuk colon cancer adalah 0.7, sedangkan untuk DLBCL adalah 0.6, dan parameter panjang subset fitur untuk *colon cancer* adalah 30, sedangkan untuk DLBCL adalah 10.

REFERENSI

[1] Li, J., Tang, X., Liu, J,. Huang, J., dan Wang, Y, "A novel approach to feature extraction from classification model based on information gene pairs", *Pattern Recognition,* 41 : 6. Juni, 2008.

[2] Li, J., Tang, X, "A new classification model with simple decision rule for discovering optimal feature gene pairs", *Computers in Biology and Medicine* 37, 2007.

[3] Theodoridis, S., Koutroumbas, K, *Pattern Recognition Third Edition***,** China: Machine Press. Pp 495, 2003.

[4] Walpole, R.E., Myers, R.H., Myers, S.L., Ye, K, *Probability & Statistics for Engineers & Scientist Seventh Edition.* Prentice Hall, pp 356, 2002.

[5] Xiong, M., Fang, X., Zhao, J, "Biomarker identification by feature wrapper", *Genome Res* 11, 2001.

[6] Gen, M., Cheng, R, *Genetic Algorithm and Enginering Design*, Japan: A wiley-Interscience Publication, John Wiley & Sons, Inc, 1997.

[7] Goldberg, D.E, *Genetic Algorithm in Search, Optimization, and Machine Learning,* USA: Addition Wesley Publishing Company, Inc, 1989.

[8] Srinivas, M., Patnaik, L.M, "Genetic algorithm: a survey", IEEE Comput**.** 27, 1994.

[9] Alon, U. Barkai, N., Notterman, Gish, K., Ybarra, S., Mack, D., and Leviner, J, Data pertaining to the article 'Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays', 1999, http://microarray. princeton.edu/oncology/affydata /index.html.

[10] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., et al, The Web Supplement to Distinct Types of Diffuse Large B-Cell Lymphoma Identified By Gene Expression Profiling, 2000, http://llmpp.nih.gov/lymphoma/data/rawdata.