Analisis Seleksi Fitur *Genetic Algorithm* dan Ekstraksi Fitur *Wavelet* pada Klasifikasi *Microarray* Data Menggunakan *Naïve Bayes*

Analysis of Genetic Algorithm Feature Selection and Wavelet Feature Extraction on Microarray Data Clasification using Naïve Bayes

Milah Sarmilah[1], Adiwijaya[2], Aniq Atiqi R[3]

[1,3]Prodi S1 Ilmu Komputasi, Fakultas Informatika, Universitas Telkom
[2]Prodi S1 Ilmu Komputasi, Fakultas Informatika, Universitas Telkom
[3]Prodi S1 Ilmu Komputasi, Fakultas Informatika, Universitas Telkom
[1]sarmilah61@gmail.com, [2]adiwijaya@telkomuniveristy.co.id, [3]aniqatiqi@telkomuniversity.ac.id

ABSTRACT

Microarray is a modern technique facilitating the simulation analysis of a number of large gene expression data which is needed to solve many complex biological problems. Therefore, there is a scheme which required the process of dimensional reduction and classification process. In this case, the dimension reduction process aims to relieve the computational load the classification, the reduction process used is the feature selection of Genetic Algorithm and feature extraction of Wavelet Haar. The process of classification aims to classify data, whether it is cancer or not, by using the Naïve Bayes classification method. As a result, the accuracy of Genetic Algorithm selection of the data features of detection of Colon Tumor is 76,4706%, 98.0769% of Lung Cancer, and 75% for ovarian. While, Haar Wavelet feature extraction from Colon Tumor data has 80% of accuracy, 94,1176% of Lung Cancer and 100% for ovarian.

Keywords: microarray data, naïve bayes, genetic algorithm, haar wavelet

ABSTRAK

Microarray adalah teknik modern yang memfasilitasi analisis simulasi dari sejumlah data ekspresi gen yang besar yang diperlukan untuk memecahkan masalah biologis yang kompleks. Oleh karena itu, diperlukan skema yang didalamnya terdapat proses reduksi dimensi dan proses klasifikasi. Dalam hal ini, proses reduksi dimensi bertujuan untuk meringankan beban komputasi pada klasifikasi, proses reduksi yang digunakan yaitu seleksi fitur *Genetic Algorithm* dan ekstraksi fitur *Wavelet* Haar. Kemudian, proses klasifikasi bertujuan untuk mengklasifikasikan data kanker atau bukan kanker, dengan menggunakan metode klasifikasi *Naïve Bayes.* Adapun akurasi terbaik dari seleksi fitur *Genetic Algorithm* pada data c*olon tumor* 52,9412%, penyakit *lung cancer* 88,2452% dan *ovarian* 75%. Sedangkan, performansi terbaik dari ekstraksi fitur *wavelet* Haar memberikan hasil untuk penyakit c*olon tumor* sebesar 80%, penyakit *lung cancer* 94,1176% dan *ovarian* 100%.

Kata kunci : *data microarray, naïve bayes, genetic algorithm, wavelet haar*

I. PENDAHULUAN

Kanker adalah penyakit yang disebabkan oleh pertumbuhan sel yang tidak normal. Penyakit ini adalah penyebab utama kematian di berbagai negara dan penyebab ketiga kematian di negara berkembang [1]. Berdasarkan survei yang dilakukan oleh *World Health Organization* (WHO), Kanker adalah penyebab utama kematian kedua di dunia, kanker bertanggung jawab atas 8,8 juta kematian pada tahun 2015 [2]. Oleh karena itu, diperlukan teknologi untuk mendeteksi penyakit kanker sejak dini agar mendapat penanganan lebih awal dengan hasil analisis yang akurat. Namun, mendeteksi kanker bukan suatu hal yang mudah, sehingga untuk mendeteksi dan menganalisis penyakit kanker akan dilakukan dengan teknik *microarray.*

Microarray adalah teknik modern yang memfasilitasi analisis simulasi dari sejumlah data ekspresi gen yang besar yang diperlukan untuk memecahkan masalah biologis yang kompleks. Masalah seperti analisis *microarray* digunakan untuk klasifikasi penyakit. Klasifikasi penyakit adalah masalah pengelompokan sampel yang diberikan ke salah satu sub kelas dari tipe penyakit dimana sub kelas dari suatu penyakit telah ditetapkan.

Proses klasifikasi data *microarray* membutuhkan usaha yang lebih karena dimensi yang besar dan hubungan yang kompleks antara berbagai gen. Aspek penting lainnya dalam pengolahan *microarray* adalah Fenomena *Curse of dimensionality*. Dimana, fenomena ini mempersulit proses pencarian informasi atau bahkan menghalangi informasi penting yang sangat dibutuhkan. Sehingga proses pengolahan data menjadi kurang efektif dan efisien yang akan menyebabkan beban komputasi menjadi tidak stabil [16].

Oleh karena itu, dibutuhkan reduksi dimensi pada data *microarray* sebelum proses klasifikasi, yang bertujuan untuk meringankan beban komputasi pada klasifikasi [17], proses reduksi dimensi yang digunakan seleksi fitur *Genetic Algorithm* dan ekstraksi fitur *Wavelet* Haar. Kemudian, proses klasifikasi bertujuan untuk mengklasifikasikan data kanker

atau bukan kanker, dengan menggunakan metode klasifikasi *Naïve Bayes.* Hal ini dibuktikan pada jurnal Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. "Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages." (2009) [9], mengatakan bahwa "Naïve Bayes Classifier memiliki tingkat akurasi yang lebih baik dibanding model classifier lainnya".

II. TINJAUAN PUSTAKA

Berbagai penelitian telah dilakukan oleh para ahli untuk menangani masalah dimensi tinggi yang dimiliki oleh data microarray, juga metode yang tepat untuk mereduksi dan mengkasifikasikan data tersebut. Berikut beberapa rujukan yang melakukan pengujian dengan menggunakan seleksi fitur Genetic Algorithm, ekstraksi fitur *wavelet,* dan data dengan dimensi tinggi. Sony Sunaryo menggunakan Transformasi Wavelet Disktret dengan akurasi sebesar 96,7% dan 89,4% [12]. Dwi Nugroho dengan menggunakan metode Percentage Split, Cross-validation dengan Hold-Out, PCA, Naive Bayes dan GA dengan akurasi sebesar 100% dan 94,74%[10]. Putri Tsasabila dengan menggunakan metode Functional Link Neural Network (FLNN), Genetic Algorithm (GA) dengan akurasi sebesar 92,3% dan 87,5% [7]. Aniq Atiqi dan Adiwijaya dengan mengguankan metode HPLC, PLS, TWD dengan akurasi sebesar 99,98%[8]. Berdasarkan rujukan tersebut, dapat terlihat bahwa wavelet dan GA merupakan metode yang baik untuk reduksi dimensi data berdimensi tinggi, sehingga diharapkan reduksi dimensi dapat mengurangi beban komputasi pada proses klasifikasi. Dalam hal ini, proses reduksi yang digunakan yaitu seleksi fitur *Genetic Algorithm* dan ekstraksi fitur *Wavelet* Haar. Kemudian, proses klasifikasi bertujuan untuk mengklasifikasikan data kanker atau bukan kanker, dengan menggunakan metode klasifikasi *Naïve Bayes.*

1. Normalisasi

Setiap data kanker yang akan digunakan dalam penelitian ini memiliki perbedaan spesifikasi *range* nilai yang cukup signifikan. Oleh karena itu, diperlukan normalisasi data sehingga skala (*range)* nilai pada setiap data kanker berada pada range 0 sampai 1. Dibawah ini merupakan rumus umum untuk normalisasi data,

$$Normalisasi = \frac{data - \min(data)}{\max(data) - \min(data)} \quad (1)$$

proses normalisasi tersebut dapat membuat range nilai pada dataset memiliki skala nilai yang seragam antara 0 sampai 1, sehingga kompleksitas data pada saat data digunakan sebagai masukan (input) akan berkurang.

2. Seleksi Fitur *Genetic Algorithm*

Adapun untuk proses seleksi fitur *Genetic Algorithm* [23]antara lain:

1. Dalam penerapan *Genetic Algorithm*, setiap kromosom mempresentasikan atribut-atribut yang ada dalam data dengan skema pengkodean *Binary Encoding.* Sehingga setiap individu direpresentasikan ke dalam deretan bilangan biner 0 atau 1. Selanjutnya, inisialisasi populasi berdasarkan bilangan biner 0 dan 1 yang akan dibangkitkan secara acak sebanyak jumlah fitur (atribut) dan ukuran populasi. Setiap individu pada populasi mempresentasikan sebuah solusi kandidat terhadap masalah seleksi fitur. Jika sebuah bit sama dengan 0 artinya fitur tersebut tidak akan terpilih, namun apabila bit tersebut berupa bilangan biner 1, maka fitur tersebut terpilih.
2. Evaluasi *fitness* diperoleh dari hasil perhitungan algoritma *naïve bayes* yang digunakan untuk mengevaluasi individu. *Naïve bayes* akan melakukan klasifikasi terhadap data latih, sehingga diperoleh kelas prediksi yang selanjutnya akan dihitung akurasi klasifikasi menggunakan c*onfusion matrix*. Nilai *fitness* yang tertinggi dalam proses pelatihan data akan dipilih sebagai standar ukuran kualitas evaluasi individu berikutnya.
3. Kemudian setelah melakukan evaluasi *fitness* pada tiap-tiap individu, maka individu yang memiliki nilai *fitness* terbaik akan disimpan proses ini dikenal dengan *elistme. Elitisme* dilakukan agar nilai *fitness* tersebut yang merupakan kandidat solusi terbaik tidak hilang selama operasi genetika yang berlangsung nantinya.
4. Setiap individu akan diseleksi untuk menjadi orangtua dengan metode *Roulette Wheel,* dengan menempatkan setiap kromosom individu pada *Roulette Wheel* sesuai dengan proporsi nilai *fitness* yang dimiliki masing-masing individu. Semakin besar nilai *fitness* suatu kromosom, maka proporsi yang dimilikinya dalam *Roulette Wheel* akan semakin besar pula, sehingga peluang individu tersebut terpilih untuk menjadi orang tua juga semakin besar.
5. Selanjutnya, *crossover* dilakukan terhadap kromosom orang tua yang telah terpilih melalui proses sebelumnya sehingga menghasilkan kromosom *offspring* (anak). Setiap kromosom *offspring* yang telah dibentuk sudah pasti mewarisi gen dari orang tuanya. Dalam penelitian ini dilakukan *single point crossover.* Dengan probabilitas *crossover (Pc)* merujuk dalam jurnal [13] *Pc* sebesar 0,8.
6. Mutasi dilakukan dengan membangkitkan kromosom *offspring* secara acak dalam bilangan biner dengan syarat probabilitas mutasi tertentu. Jika bilangan biner yang telah dibangkitkan secara acak memenuhi kriteria kurang dari probabilitas mutasi *Pm*, maka gen tersebut akan diubah dengan bilangan biner sebaliknya (0

diubah menjadi 1, dan 1 diubah menjadi 0)[3]. Dengan probabilitas Mutasi *(Pm)* merujuk dalam jurnal [13] *Pm* sebesar 0,1.

7. Sesuai dengan konsep GA, yang kuat adalah yang bertahan maka seleksi survivor dilakukan. Metode yang digunakan dalam proses ini adalah *Generational Replacement* yang nantinya kromosom pada suatu generasi diperbaharui atau digantikan sekaligus dengan kromosom baru hasil *crossover* dan mutasi, serta kromosom terbaik yang sudah disimpan dalam *Elitisme.*
8. Kemudian, menghitung data testing menggunakan metode *naïve bayes* dari proses *learning* yang didapat.
9. Selanjutnya kinerja sistem dalam mengklasifikasikan diukur dengan metode *confusion matrix* dengan menghitung akurasi dari proses klasifikasi.

3. Ekstraksi Fitur menggunakan *Wavelet* Haar

Menurut Vidacovic dan Meuller (1991)[4], suatu fungsi $\psi(.)$ bernilai real, disebut *wavelet* jika memenuhi:

$$\int_{-\infty}^{+\infty} \psi_{j,k}^2(t)dt = 1 \tag{2}$$

$$\int_{-\infty}^{+\infty} \psi_{j,k}(t)dt = 0 \tag{3}$$

Menurut Sony dan Notodiputro (1992), dalam Suprapti (2009), *wavelet* Haar merupakan *mother wavelet* othogonal yang sederhana dan banyak dipakai untuk kasus reduksi dimensi data [19].

$$\psi(t) = \begin{cases} 1, & 0 \le t < 1/2 \\ -1, & 1/2 \le t < 1 \\ 0, & selainnya \end{cases} \tag{4}$$

sedangkan untuk *Father wavelet* $\phi_{0,0}(t)$ pada *wavelet* Haar yang biasa ditulis $\phi(t)$, adalah:

$$\phi(t) = \begin{cases} 1, & 0 \le t < 1 \\ 0, & selainnya \end{cases} \tag{5}$$

Dalam transformasi *wavelet* diskret (TWD) suatu vektor pengamatan dinyatakan sebagai kombinasi linear dari fungsi-fungsi basis yang disebut fungsi *wavelet* [4]. Jika ada vektor data berukuran $p = 2^M$, M bilangan bulat positif, maka vektor tersebut dapat dinyatakan dalam fungsi tangga pada interval [0,1) yaitu:

$$f(t) = \sum_{k=0}^{2^M-1} x_k \, I_{\left\{ {}^{k}/_{2^M} \le t < {}^{k+1}/_{2^M} \right\}} \tag{6}$$

dengan transformasi *wavelet* diskret *f(t)* dapat didekomposisikan menjadi:

$$f(t) = c_{0,0}\phi(t) + \sum_{j=0}^{M-1} \sum_{k=0}^{2^j-1} d_{j,k}\, \psi_{j,k}(t) \tag{7}$$

Persamaan (7) disebut dengan persamaan umum TWD, karena nilai *j* hanya diambil pada bilangan bulat bukan negatif saja [18]. Bilangan *j* disebut level resolusi dan *f(t)* dapat diperoleh dengan tepat, jika diambil semua level resolusi untuk dekomposisi, yaitu semua *(M-1)* level resolusi pertama. Koefisien $c_{0,0}$ disebut koefisien pemulus atau bagian pendekatan dari suatu fungsi[24][25][26]. Sedangkan $d_{j,k}$ disebut dengan koefisien *wavelet* atau bagian *detail* suatu fungsi $\psi(t)$ dan $\phi(t)$ [12], maka persamaan (7) dapat dituliskan dengan notasi matriks,

$$\underline{x} = W^T \underline{d} \tag{8}$$

Contoh, matriks *wavelet* Haar jika $p = 4$ maka W^T yaitu:

$$W^T = \begin{bmatrix} 1/2 & 1/2 & 1/\sqrt{2} & 0 \\ 1/2 & 1/2 & -1/\sqrt{2} & 0 \\ 1/2 & -1/2 & 0 & 1/\sqrt{2} \\ 1/2 & -1/2 & 0 & -1/\sqrt{2} \end{bmatrix}$$

untuk menghasilkan W^T [20] yaitu:

$$h = (2^{j-k}) . (2^{j-k+1}) \quad (9)$$

$$h_s, s = (M-1) + (i-1) - 2(j-1) modulo\ (2^{j-k+1}) \quad (10)$$

$$H = h_s, s + (h_s, s = 0) . (2^{j-k+1}) \quad (11)$$

dimana,
k = 1,2,... bilangan bulat,
h = matriks sirkuler,
M = jumlah bilangan bulat positif,
s = pergeseran dari operator modulo.

Karena W orthogonal [21] pada persamaan (8) maka koefisien *wavelet* dapat dihitung dengan,

$$\underline{d} = W\,\underline{x} \quad (12)$$

dimana $\underline{d} = \left(c_{0,0}\ , d_{0,0}, d_{1,1}, d_{1,0}, \dots, d_{n-1,0}\right)^T$.

Matriks transformasi *wavelet* diskret, W, merupakan matriks orthogonal [4] untuk semua jenis *mother wavelet* yang digunakan, sehingga berlaku $W.W^T = I$. W adalah matriks yang elemen-elemen kolomnya adalah nilai dari $\phi(t)$ dan $\psi(t)$ untuk berbagai $t \in [0,1)$, untuk *wavelet* Haar [8] [27]yaitu :

$$\psi(t) = \begin{cases} 1, & 0 \le t < 1/2 \\ -1, & 1/2 \le t < 1 \\ 0, & selainnya \end{cases} \quad (13)$$

$$\phi(t) = \begin{cases} 1, 0 \le t < 1 \\ 0\ , selainnya \end{cases} \quad (14)$$

dari persamaan (12) reduksi dimensi dilakukan dengan cara mengambil $m < p$ dari koefisien-koefisien sehingga diperoleh,

$$D^*_{(nxm)} = X_{(nxp)}\ W^{*T}_{(pxm)} \quad (15)$$

dengan memberi nilai nol pada kolom *(m+1)* sampai dengan *p* dari matriks W^T. Persamaan diatas mereduksi pengamatan *p* variabel menjadi *m* koefisien *wavelet* terpilih, dimana $m < p$ [12].

4. Naive Bayes

Klasifikasi *bayes* mengasumsikan bahwa suatu fitur tidak berpengaruh dengan adanya fitur lain. Hipotesis dalam teorema *Bayes* merupakan label kelas yang menjadi target dalam pemetaan dalam klasifikasi, dan bukti merupakan fitur-fitur yang menjadi masukan dalam model klasifikasi. Jika *B* adalah masukan yang berisi fitur dan *A* adalah label kelas, maka *Naïve Bayes* dituliskan dengan *P(A|B)* [15]. Notasi tersebut berarti probabilitas label kelas *A* didapatkan setelah fitur-fitur *X* diamati. Adapun persamaan *naïve bayes* untuk klasifikasi antara lain sebagai berikut:

$$P(A|B) = \frac{\prod_{i=1}^{q} P(B_i|A)\ P(A)}{P(B)} \quad (16)$$

dimana,
P(A|B) = Probabilitas data dengan fitur *B* pada kelas *A*.
P(A) = Probabilitas awal dengan kelas *A*.
$\prod_{i=1}^{q} P(B_i|A)$ = Probabilitas independen kelas *A* dari semua fitur dalam *X*.

Umumnya *naïve bayes* mudah diimplentasikan untuk fitur bertipe kategorikal. Namun, untuk data numerik pengerjaannya akan sedikit berbeda dengan data kategorikal. Salah satunya adalah dengan mengasumsikan bentuk tertentu dari distribusi menggunakan data training. Distribusi *Gaussian* biasanya dipilih untuk mempresentasikan *conditional probability feature continuous* pada sebuah kelas $P(B_i|A)$. Distribusi *Gaussian* dikarakteristikan dengan dua parameter: mean (μ) dan variansi (σ^2), *x* adalah nilai fitur pada data yang akan diprediksi [22]. Persamaan distribusi *gauss* [15] yaitu:

$$g(x,\mu,\sigma) = P(B_i|A) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \quad (17)$$

dimana,
μ = rata-rata dari data,
σ = standar deviasi dari data,
x_i= data ke i.

III. RACANGAN SISTEM

Sebelum melakukan berbagai tahapan dalam rancangan sistem, maka tahap pertama yang harus dilakukan adalah pengumpulan *dataset microarray* yaitu data kanker, data tersebut akan digunakan untuk serangkaian proses klasifikasi kanker. Data yang digunakan dalam penelitian ini adalah data yang didapat dari Kent-Ridge Bio-medical Data Set Repository[14]. Berikut merupakan spesifikasi *data* kanker yaitu:

Tabel 1 Spesifikasi Data

Data set	Jumlah atribut	Jumlah kelas	Jumlah sampel	
Colon tumor	2000	2	40 Negarif	20 Positif
Lung cancer	12533	2	31 Mesothelioma	150 ADCA
Ovarian	15154	2	91 Normal	162 Cancer

Data *microarray* merupakan data yang memiliki dimensi yang sangat besar. Selain itu, masalah yang terdapat pada data *microarray* yang akan digunakan yaitu skala (*range*) pada setiap fiturnya memiliki perbedaan nilai yang cukup signifikan. Sehingga, diperlukan proses normalisasi data yang membuat *range* nilai pada setiap atribut (fitur) data *microarray* seragam atau berada pada *range* 0 sampai 1. Setelah melakukan proses normalisasi, maka proses selanjutnya yaitu partisi yang akan dibagi menjadi data uji dan data latih. Setelah itu, proses selanjutnya yaitu seleksi fitur dan ekstraksi fitur yang bertujuan untuk menemukan gen informatif dan mengurangi kompleksitas data yang akan digunakan sebagai masukan (input). Proses seleksi dan ekstraksi fitur dilakukan karena jumlah atribut ataupun fitur pada data *microarray* sangat banyak, spesifikasi fitur dapat dilihat pada Tabel 1, yang merupakan tabel spesifikasi data yang akan digunakan. Selanjutnya melakukan proses klasifikasi data *microarray* untuk menentukan kelas kanker atau tidak kanker pada setiap sampel. Selanjutnya kinerja sistem dalam mengklasifikasikan diukur dengan metode *confusion matrix* dengan menghitung akurasi.

Adapun rancangan sistem dari seleksi fitur dan ekstraksi fitur menggunakan Naive Bayes yaitu:

1. Seleksi Fitur Genetic Algorithm

Dalam penelitian ini, maksimum generasi atau maksimum iterasi sebagai kriteria terminasi *Genetic Algorithm*. Sehingga parameter yang dibutuhkan dalam mengimplemen-tasikan *Genetic Algorithm* sebagai seleksi fitur dengan 300 dan 500 individu yang akan dievaluasi antara lain:

Tabel 1 Kriteria Terminasi

Kriteria Terminasi	Nilai
Jumlah Individu	300 dan 500
Ukuran Populasi	60 dan 100
Maksimum Generasi	5
Skema Pengkodean	Binary Encoding
Fungsi Fitness	Akurasi Naive Bayes
Crossover	Single Point Crossover
Peluang Crossover	0,8
Mutasi	Fit Bit Mutation
Peluang Mutasi	0,1
Mekanisme Mutasi	Roulette Wheel
Seleksi Survivor	Generational Replacement

2. Ekstraksi Fitur Wavelet Haar

Adapun langkah-langkah dari ekstraksi fitur *Wavelet* Haar menggunakan transformasi *wavelet* diskret yaitu:

1. Pada metode transformasi *wavelet* diskret data X berukuran nxp, mengharuskan jumlah variabel prediktor (p) dapat dinyatakan dalam 2^M, dengan M adalah bilangan bulat positif. Jika tidak maka perlu ada pemampatan data, sehingga X berukuran nxq, dengan $q = 2^M$.
2. Mentransformasikan variabel prediktor berukuran 2^M ke dalam interval [0,1) dengan persamaan

$$f(t) = \sum_{k=0}^{2^M-1} x_k \, I_{\left\{k/2^M \le t < k+1/2^M\right\}} \tag{18}$$

3. Menghitung matriks transformasi *wavelet* diskret dengan W[4], berukuran qxq, (q adalah jumlah kolom dari data), dimana elemen-elemen kolomnya adalah nilai dari *father* $\phi(t)$ dan $mother\ \psi(t)$ *wavelet* Haar untuk berbagai $t \in$ [0,1) [6][8][27].

$$\psi(t) = \begin{cases} 1, & 0 \le t < 1/2 \\ -1, & 1/2 \le t < 1 \\ 0, & selainnya \end{cases} \tag{19}$$

$$\phi(t) = \begin{cases} 1, 0 \le t < 1 \\ 0\ , selainnya \end{cases} \tag{20}$$

4. Reduksi dimensi dilakukan dengan mengambil $m < q$ dari koefisien-koefisien *wavelet* sehingga $D^*_{(nxq)} = X_{(nxq)}\, W^{*T}_{(qxq)}$ menjadi $D^*_{(nxm)} = X_{(nxq)}\, W^{*T}_{(qxm)}$, yaitu dengan memberi nilai nol pada kolom $(m+1)$ sampai q matriks W^T.
5. Setelah mendapatkan hasil reduksi dari tiap data, maka proses selanjutnya yaitu klasifikasi menggunakan *naïve bayes* [5].
6. Selanjutnya kinerja sistem dalam mengklasifikasikan diukur dengan metode *confusion matrix* dengan menghitung akurasi dari proses klasifikasi

IV. PENGUJIAN DAN ANALISIS

Dalam pengujian ini, untuk reduksi dimensi menggunakan ekstraksi fitur *wavelet* Haar akan memperhatikan jumlah variabel prediktor dan jumlah atribut reduksi pengujian masing-masing data. Sedangkan untuk seleksi fitur *Genetic Algorithm* akan memperhatikan jumlah individu, ukuran populasi serta maksimum generasi pengujian pada masing-masing data.

1. Klasifikasi *Naïve Bayes*

Pada bagian ini akan menampilkan hasil penelitian dari skenario klasifikasi *Naïve Bayes*. Dalam penelitian ini, klasifikasi yang akan dilakukan yaitu data yang sudah dinormalisasi dengan proporsi data latih dan data uji 70:20 dan 80:20.

Tabel 3. Hasil Klasifikasi Naïve Bayes

Data set	Akurasi (%)	
	Proporsi 70:20	Proporsi 80:20
Colon tumor	41,1765	41,1765
Lung cancer	88,2352	88,2352
Ovarian	72,9167	72,9167

Berdasarkan Tabel 3 untuk proporsi pengujian 70:20 dan 80:20 mampu memberikan hasil untuk *colon tumor* 41,1765%, *lung cancer* 88,2352% dan *ovarian* 72, 9167%. Hal ini dapat dilihat bahwa kinerja pengujian 70:20 dan 80:20 memiliki akurasi yang sama untuk pengujian data menggunakan klasifikasi Naïve Bayes tanpa reduksi dimensi.

2. Seleksi Fitur *Genetic Algorithm* dengan Klasifikasi *Naïve Bayes*

Pada bagian ini akan menampilkan hasil penelitian dari skenario seleksi fitur *Genetic Algorithm*. Pada penelitian ini individu yang akan dievaluasi sebanyak 60 dan 100 populasi dengan maksimum generasi 5, sehingga jumlah individu sebanyak 300 dan 500. Selajutnya *Pc* (probabilitas *crossover*) 0,8 dan *Pm* (probabilitas mutasi) 0,1 merujuk pada jurnal [13] dengan proporsi pengujian data latih dan data uji 70:20 yaitu dan 80:20. Adapun hasil seleksi fitur *Genetic Algorithm* dengan klasifikasi *naïve bayes* sebagai berikut:

Tabel 4. Hasil Seleksi Fitur Genetic Algorithm dengan Klasifikasi Naïve Bayes

Data set	Ukuran Populasi	Akurasi (%)	
		Proporsi 70:20	Proporsi 80:20
Colon tumor	60	47,0588	52,9412
	100	47,0588	52,9412
Lung cancer	60	79,4117	85,2411
	100	79,4117	88,2352
Ovarian	60	72,9167	75
	100	72,9167	75

Berdasarkan Tabel 4 untuk proporsi pengujian 70:20 dengan jumlah individu 300 mampu memberikan hasil untuk *colon tumor* 47,0588%, *lung cancer* 79,4117% dan *ovarian* 72,9167%. Sedangkan, proporsi pengujian 70:20 dengan jumlah individu 500 mampu memberikan hasil untuk *colon tumor* 47,0588%, *lung cancer* 79,4117% dan *ovarian* 72,9167%. Kemudian, untuk proporsi pengujian 80:20 dengan jumlah individu 300 mampu memberikan hasil untuk *colon tumor* 52,9412%, *lung cancer* 85,2411% dan *ovarian* 75%. Sedangkan, proporsi pengujian 80:20 dengan jumlah individu 500 mampu memberikan hasil untuk *colon tumor* 52,9412%, *lung cancer* 88,2352% dan *ovarian* 75%.

Sesuai dengan Tabel 4 seleksi fitur *genetic algorithm* mampu menghasilkan performansi terbaik yang didapat dari proporsi 80:20 untuk data *colon tumor, lung cancer* dan *ovarain*. Hal ini disebabkan karena data training yang digunakan lebih besar dibandingkan 70:20, sehingga informasi yang didapatkan akan lebih banyak untuk di uji. Hal ini dapat dilihat dari kenaikan akurasi dari proporsi 70:20 ke proporsi 80:20 dari setiap data set pengujian. Selain itu, penggunaan ukuran populasi lebih dari 100 akan memberikan hasil yang lebih baik karena seleksi akan mencari atribut (kromosom) yang paling berpengaruh untuk di uji, sehingga atribut (kromosom) yang dihasilkan lebih banyak. Dari Tabel 4, dapat disimpulkan bahwa pemilihan jumlah atribut dan proporsi pengujian yang tepat akan memberikan hasil performansi yang baik. Untuk kasus pengujian ini jumlah individu untuk seleksi fitur yang tepat yaitu 500 dengan ukuran populasi 100 dan maksimum generasi 5, dengan pengujian proporsi 80:20.

3. Ekstraksi Fitur *Wavelet* Haar dengan Klasifikasi *Naïve Bayes*

Berikut ini merupakan pemaparan mengenai hasil reduksi dimensi menggunakan *wavelet* Haar dengan klasifikasi *Naïve Bayes.* Proses reduksi ini dilakukan pada data microarray yaitu *colon tumor, lung cancer* dan *ovarian* yang memiliki dimensi tinggi (Tabel 1). Pada metode *wavelet* mensyaratkan jumlah variabel prediktor yang akan direduksi harus memenuhi 2^M. Variabel prediktor pada data *colon tumor* berjumlah 2000 atribut, *lung cancer* 12533 dan *ovarian* 15154 atribut, maka akan diambil *colon tumor* 1024, *Lung cancer* dan *ovarian* 8192 atribut. Adapun hasil ekstraksi fitur *wavelet* Haar dengan klasifikasi *Naïve Bayes* dengan jumlah atribut pengujian yaitu 300 dan 500 dengan proporsi pengujian data latih dan data uji yaitu 70:20 dan 80:20 sebagai berikut:

Tabel 5 Hasil Ekstraksi Fitur Wavelet Haar dengan Klasifikasi Naïve Bayes

Data set	Jumlah atribut	Akurasi (%)	
		Proporsi 70:20	Proporsi 80:20
Colon tumor	300	80	80
	500	80	80
Lung cancer	300	94,1176	94,1176
	500	94,1176	94,1176
Ovarian	300	100	100
	500	100	100

Berdasarkan Tabel 5 untuk proporsi pengujian 70:20 dengan maksimum atribut 300 mampu memberikan hasil untuk *colon tumor* 80%, *lung cancer* 94,1176% dan o*varian* 100%. Sedangkan, proporsi pengujian 70:20 dengan maksimum atribut 500 mampu memberikan hasil untuk c*olon tumor* 80%, *lung cancer* 94,1176% dan *ovarian* 100%. Kemudian, untuk proporsi pengujian 80:20 dengan maksimum atribut 300 mampu memberikan hasil untuk *colon tumor* 80%, *lung cancer* 94,1176% dan o*varian* 100%. Sedangkan, proporsi pengujian 80:20 dengan maksimum atribut 500 mampu memberikan hasil untuk c*olon tumor* 80%, *lung cancer* 94,1176% dan *ovarian* 100%.

Sesuai dengan Tabel 5 ekstraksi fitur wavelet Haar mampu menghasilkan performansi terbaik yang didapat dari proporsi 80:20 dan 70:20 untuk data *colon tumor, lung cancer* dan *ovarain*. Hal ini dapat dilihat dari kinerja yang stabil dari setiap data. Dari Tabel 5, dapat disimpulkan bahwa pemilihan jumlah atribut dan proporsi pengujian yang tepat akan memberikan hasil performansi yang baik. Untuk kasus pengujian ini jumlah atribut reduksi untuk ekstraksi fitur yang tepat yaitu 300 dan 500, sedangkan untuk pengujian proporsi 70:20 dan 80:20.

4. Perbandingan Seleksi Fitur dan Ekstraksi Fitur Dengan Klasifikasi

Berdasarkan kedua pengujian reduksi dimensi yaitu seleksi fitur dan ekstraksi fitur, maka ekstraksi fitur mampu memberikan hasil yang lebih baik dari pada seleksi fitur untuk data *colon tumor, lung cancer* dan *ovarian*. Hal ini disebabkan karena ekstraksi fitur melibatkan semua atribut yang akan digunakan atau yang akan diekstraksi, sedangkan seleksi fitur hanya memilih atribut yang paling berpengaruh untuk diseleksi. Pada dasarnya semua atribut (dimensi) dapat mempengaruhi klasifikasi maka jika diseleksi akan ada informasi yang hilang. Sehingga, seleksi fitur memberikan hasil yang lebih kecil dibandingkan ekstraksi fitur. Untuk kasus pengujian ini dapat disimpulkan bahwa ekstraksi fitur *wavelet* haar dengan klasifikasi *Naïve Bayes* mampu memprediksi penyakit *colon tumor, lung cancer* dan *ovarian* lebih baik dibandingkan seleksi fitur.

Sesuai dengan pengujian yang dilakukan, maka proporsi yang tepat untuk pengujian ini yaitu proporsi 80:20. Hal ini disebabkan karena proporsi 80:20 memiliki data training yang lebih besar dibandingkan 70:20, sehingga data yang akan diproses untuk data training pada standar deviasi dan mean pada proses klasifikasi memiliki informasi yang lebih banyak. Selain itu, pada proses seleksi pemilihan data training yang lebih besar akan memberikan peluang informasi yang penting yang akan terpilih lebih banyak. Sedangkan untuk kasus ini, seleksi fitur yang tepat untuk jumlah individu lebih dari 500 dengan ukuran populasi 100 dan maksimum generasi 5, hal ini disebabkan karena informasi penting yang akan didapatkan lebih banyak. Sedangkan, jumlah atribut ekstraksi fitur yang tepat untuk kasus ini yaitu 300 dan 500.

Berdasarkan analisis tersebut, berikut merupakan tabel perbandingan akurasi reduksi dimensi seleksi fitur *genetic algorithm* dan ekstraksi fitur *wavelet* Haar dengan proporsi pengujian 80:20, jumlah atribut untuk ekstraksi fitur 500 dan jumlah individu untuk seleksi fitur 500 dengan ukuran populasi 100 dan maksimum generasi 5 antara lain:

Tabel 6 Perbandingan Akurasi Genetic Algorithm Dan Wavelet Haar dengan Menggunakan Klasifikasi Naïve Bayes

Data set	Akurasi (%)	
	Genetic Algorithm	Wavelet Haar
Colon tumor	52,9412	80
Lung cancer	88,2452	94,1176
Ovarian	75	100

Sesuai Tabel 6 maka seleksi fitur *Genetic Algorithm* dengan *Naïve Bayes* mampu memberikan hasil performansi untuk penyakit c*olon tumor* 52,9412%, penyakit *lung cancer* 88,2452% dan *ovarian* 75%. Sedangkan, ekstraksi fitur *wavelet* Haar mampu memberikan hasil untuk penyakit c*olon tumor* sebesar 80%, penyakit *lung cancer* 94,1176% dan *ovarian* 100%.

V. KESIMPULAN

Microarray adalah teknik modern yang memfasilitasi analisis simulasi dari sejumlah data ekspresi gen yang besar yang diperlukan untuk memecahkan masalah biologis yang kompleks. Proses klasifikasi data *microarray* membutuhkan usaha yang lebih karena dimensi yang besar dan hubungan yang kompleks antara berbagai gen. Dalam hal ini, proses reduksi dimensi bertujuan untuk meringankan beban komputasi pada klasifikasi, proses reduksi yang digunakan yaitu seleksi fitur *Genetic Algorithm* dan ekstraksi fitur *Wavelet* Haar. Hasil reduksi dimensi dengan metode ekstraksi fitur *wavelet* Haar mampu memprediksi dengan baik untuk data *colon tumor, lung cancer* dan *ovarian.* Seleksi fitur *Genetic Algorithm* dengan *Naïve Bayes* mampu memberikan hasil performansi untuk penyakit c*olon tumor* 52,9412%, penyakit *lung cancer* 88,2452% dan *ovarian* 75%. Sedangkan, estraksi fitur *wavelet* Haar mampu memberikan hasil untuk penyakit c*olon tumor* sebesar 80%, penyakit *lung cancer* 94,1176% dan *ovarian* 100%.

DAFTAR PUSTAKA

[1] C. Hayter. Cancer: “the worst scourge of civilized mankind”. Can Bull Med Hist., 20(2):251–264, 2003

[2] WHO. The World Health Organisation, 2017. Available from: http://www.who.int/mediacentre/factsheets/fs297/en/[Accessed 25 April 2017].

[3] Suyanto, S. M. (2008). Evolutionary Computing. Bandung: Informatika.

[4] Vidacovic B. Dan Meuller P., 1991. *Wavelet for kids. A Tutorial Introduction. AMS Subject Classification*, Duke University

[5] Mubarok, M.S., Adiwijaya and Aldhi, M.D., 2017. Aspect-Based Sentiment Analysis To Review Products Using Naïve Bayes. In AIP Conference Processings (Vol.1867, No. 1, p. 020060). AIP Publishing.

[6] Suprapti, A., 2009. *Pra-pemprosesan Data Luaran GCM CSIRO-Mk3 dengan Metode Transformasi Wavelet Diskret/*Tugas Akhir. Surabaya, Institut Teknologi Sepuluh November.

[7] Ramadhani, P., (2017). *Deteksi Kanker berdasarkan Klasifikasi Data Microarray menggunakan Functional Link Neural Network dengan Seleksi Fitur Genetic Algorithm/*Tugas Akhir. Bandung, Telkom University.

[8] Rohmawati A., Adiwijaya, 2017. *A Daubechies Wavelet Transformation to Optimize Modeling Calibration of Active Compound on Drug Plants. In 5th* International Conference on Information and Communication Technology (ICoICT). Pp.1-4. IEEE

[9] Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. (2009) "Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages".

[10] Nugroho, Dwi. 2016. *Prediksi Penyakit Menggunakan Genetic Algorithm(GA) dan Naive Bayes untuk Data Berdimensi Tinggi*/Tugas Akhir. Bandung: Telkom University.

[11] E. Prasetyo, *Data Mining: Konsep dan Aplikasi menggunakan Matlab*, 1 ed. Yogyakarta: Andi Offset, 2012.

[12] Sunaryo, S., 2005. *Model Kalibrasi dengan Transformasi Wavelet sebagai Metode Pra-pemrosesan/*Disertasi. Bogor: Sekolah Pascasarjana, Institut Pertanian Bogor.

[13] O Babatunde, Leisa Armstrong, Jinsong Leng, and Dean Diepeveen. A genetic algorithm-based feature selection. International Journal of Electronics Communication and Computer Engineering, 5(4):889–905, 2014.

[14] Jinyan Li, Kent-ridge bio-medical data set repository, School of Computer Engineering Nanyang Technological University, Singapore, Downloaded at April 2017 from URL: http://leo.ugr.es/elvira/DBCRepository/.

[15] Suyanto, S. M. (2008). Soft Computing. Bandung: Informatika..

[16] Mukesh Kumar, Sandeep Singh, and Santanu Kumar Rath. Classification of microarray data using functional link neural network. Procedia Computer Science, 57:727–737, 2015.

[17] Nurfalah, A.,Adiwijaya, and Suryani, A.A, 2016. Cancer Detection Based On Microarray Data Classification Using PCA And Modified Bank Propagation. Far East Journal of Electronics and Communications, 16 (2), p. 269.

[18] Kaist, 2005. *The Discrete Wavelet Transform*

[19] Suprapti, A., 2009. *Pra-pemprosesan Data Luaran GCM CSIRO-Mk3 dengan Metode Transformasi Wavelet Diskret/*Tugas Akhir. Surabaya, Institut Teknologi Sepuluh November.

[20] Vidacovic B (1991), *Statistical Modeling by Wavelet.* Pages 115. Duke University.

[21] Percival, 2005. *An Introduction to the wavelet Analysis of Time Series.* Seattle: Washington University.

[22] Shantanam, T. & Padmavathi, M.S., 2015. Aplication of K-Means and Genetic Algorithms for Dimensional Reduction by Intergrating SVM for Diabetes Diagnosis. Science Direct.

[23] Nhita, F., Adiwijaya, 2013. A rainfall forecasting using fuzzy system based on genetic algorithm. In Information and Communication Technology (ICoICT), 2013 International Conference of (pp. 111-115). IEEE.

[24] Wisesty, U.N., Nasri, J., and Adiwijaya. 2016. Modified Backpropagation Algorithm for Polycystic Ovary Syndrome Detection Based on Ultrasound Images. In International Conference on Soft Computing and Data Mining (pp. 141-151). Springer, Cham

[25] Adiwijaya. 2014. Aplikasi Matriks dan Ruang Vektor. Yogyakarta: Graha Ilmu

[26] Adiwijaya, 2016, Matematika Diskrit dan Aplikasinya, Bandung: Alfabeta

[27] Adiwijaya, Maharani, M., Dewi, B.K., Yulianto, F.A. and Purnama, B., 2013. digital image compression using graph coloring quantization based on wavelet-SVD. In Journal of Physics: Conference Series(Vol. 423, No. 1, p. 012019). IOP Publishing.