IJCCS, Vol.9, No.1, January 2015, pp. 33~42
ISSN: 1978-1520

Klasifikasi Data Microarray Menggunakan Discrete Wavelet Transform dan Extreme Learning Machine

Khadijah*[1], Sri Hartati[2]
[1] Prodi S2/S3 Ilmu Komputer, FMIPA UGM, Yogyakarta
[2] Jurusan Ilmu Komputer dan Elektronika, FMIPA, UGM, Yogyakarta
e-mail: ***[1]khadijah0303@gmail.com,** [2]shartati@ugm.ac.id

Abstrak

Data microarray digunakan sebagai alternatif untuk diagnosa penyakit kanker karena kesulitan dalam dignosa kanker berdasarkan bentuk morfologis, yaitu perbedaan morfologis yang tipis antar jenis kanker yang berbeda. Penelitian ini bertujuan untuk membangun pengklasifikasi data microarray. Proses klasifikasi diawali dengan reduksi dimensi data microarray menggunakan DWT, dengan cara mendekomposisi sampel hingga level tertentu, kemudian mengambil nilai koefisien aproksimasi pada level tersebut sebagai fitur sampel. Fitur tersebut selanjutnya menjadi masukan untuk klasifikasi. Metode klasifikasi yang digunakan adalah ELM yang diterapkan pada RBFN. Dataset yang digunakan adalah data microarray multikelas, yaitu dataset GCM (16.063 gen, 14 kelas) dan Subtypes-Leukemia (12.600 gen, 7 kelas).

Pengujian dilakukan dengan cara membagi data latih dan data uji secara random sepuluh kali dengan proporsi data yang sama. Classifier yang dihasilkan dari penelitian ini untuk dataset GCM belum memiliki performa yang cukup baik, ditunjukkan dengan nilai akurasi sekitar 75% ± 6,25% dan nilai minimum sensitivity yang masih rendah, yaitu 15% ± 19,95% menunjukkan bahwa sensitivity untuk tiap kelas belum merata, terdapat beberapa kelas yang sensitivity-nya masih rendah. Namun, classifier untuk dataset Subtypes-Leukemia yang memiliki jumlah kelas lebih sedikit dari dataset GCM memiliki performa yang cukup baik, ditunjukkan dengan nilai akurasi 87,68% ± 2,88% dan minimum sensitivity 51,90% ± 20,29%.

Kata kunci— *microarray, ekspresi gen, DWT, ELM, RBFN*

Abstract

Microarray data is used as an alternative in cancer diagnosis because of the difficulties cancer diagnosis based on morphologis structures. Different classes of cancer usually have poor distintion of morphologis structures. The aim of this reserach is to bulid microarray data classfier. The classification process is started by reducing dimension of microarray data. The method used to reduce the microarray data dimension is DWT by decomposing the samples until certain decomposition level and then use approximation coefficients at those level as feature to classifier. Classifier used in this reserach is ELM implemeted on RBFN. Dataset used are GCM (16.063 genes, 14 classes) and Subtypes-Leukemia (12.600 genes, 7 classes).

Testing process is done by randomly dividing the training and testing data ten times with same proprotion of training and testing data. The perfomance of classifier built in this research is not so good for GCM dataset, shown by accuracy 75% ± 6,25% and mean of minimum sensitivity 15% ± 19,95%. The low minimum sensitivity indicate that there are few classes that have low sensitivity. But the classifier for Subtypes-Leukemia dataset give better result, that is accuracy 87,68% ± 2,88% and mean of minimum sensitivity 51,90% ± 20,29%.

Keywords— *microarray, gene expression, DWT, ELM, RBFN*

Received March 4th,2014; Revised January 8th, 2015; Accepted January 20th, 2015

1. PENDAHULUAN

Kanker merupakan salah satu jenis penyakit yang berbahaya, namun diagnosa kanker bukanlah sesuatu yang mudah. Diagnosa kanker berdasarkan struktur morfologisnya memiliki kesulitan karena perbedaan morfologis yang sangat tipis untuk jenis kanker yang berbeda. Kesulitan dalam diagnosa kanker juga dapat disebabkan oleh ketidaklengkapan informasi klinis yang tersedia dari seorang pasien serta kemungkinan adanya faktor subjektivitas pada interpretasi data tersebut. Sejumlah kesulitan tersebut telah mendorong beberapa penelitian untuk menentukan jenis kanker berdasarkan tingkat ekspresi gen [1]. Tingkat ekspresi gen menunjukkan keaktifan sebuah gen di dalam sel tubuh yang diukur berdasarkan jumlah mRNA di dalam sel tubuh tersebut [2]. Tingkat ekspresi gen dalam tubuh manusia dapat diukur melalui eksperimen microarray [3]. Pola ekspresi gen dari sejumlah pasien yang telah diketahui jenis kankernya, dapat digunakan untuk memprediksikan jenis kanker pada pasien yang baru. Permasalahan tersebut disebut sebagai klasifikasi [1, 3, 4].

Permasalahan penting dalam klasifikasi data *microarray* adalah jumlah gen yang sangat besar, dapat mencapai puluhan ribu (*high dimensional*), sedangkan jumlah sampel yang terbatas, hanya berkisar puluhan atau ratusan. Hal tersebut menyebabkan perlunya reduksi dimensi pada data *microarray* sebelum proses klasifikasi yang bertujuan untuk menghemat waktu komputasi [3] serta menghindari *overvitting* pada *classifier* [5]. Salah satu cara untuk mereduksi data *microarray* adalah ekstraksi fitur menggunakan *Discrete wavelet transform* (DWT) [6]. DWT dapat mendekomposisi sinyal untuk memisahkan bagian sinyal yang memiliki frekuensi rendah atau *large scale* dengan sinyal yang memiliki frekuensi tinggi atau *small scale* [7]. Bagian sinyal yang berisi frekuensi rendah menggambarkan identitas atau karakteristik sinyal [8] sehingga bagian tersebut dapat mencirikan sinyal dan berpotensi digunakan sebagai fitur sinyal. Beberapa penelitian mengenai ekstraksi fitur pada data *microarray* menggunakan DWT telah dilakukan [9, 10, 11] dan menunjukkan bahwa penggunaaan DWT dapat membantu *classifier* untuk menghasilkan akurasi lebih dari 90%.

Tahap berikutnya setelah reduksi dimensi adalah klasifikasi. Berbagai penelitian mengenai klasifikasi data *microarray* yang terdiri atas dua kelas (klasifikasi biner) telah banyak dilakukan dan memberikan hasil yang baik [9, 10, 11, 12, 13]. Namun, hasil penelitian pada klasifikasi data *microarray* yang terdiri lebih dari dua kelas (klasifikasi multikelas) tidak sebaik pada klasifikasi biner. Hal tersebut dikarenakan semakin banyak jumlah kelas maka semakin tinggi tingkat kesulitan dalam permasalahan klasifikasi [14]. Metode SVM (*support vector machine*) menghasilkan akurasi yang cukup tinggi dalam permasalahan klasifikasi biner [9, 10, 11, 13, 15], namun saat jumlah kelas bertambah, SVM tidak dapat digunakan secara langsung dan harus dimodifikasi, misalnya SVM-OVA (*one-versus-all*) atau SVM-OVO (*one-versus-one*). Hal tersebut meningkatkan kompleksitas *classifier* sehingga membutuhkan *resource* komputasi yang lebih besar [16].

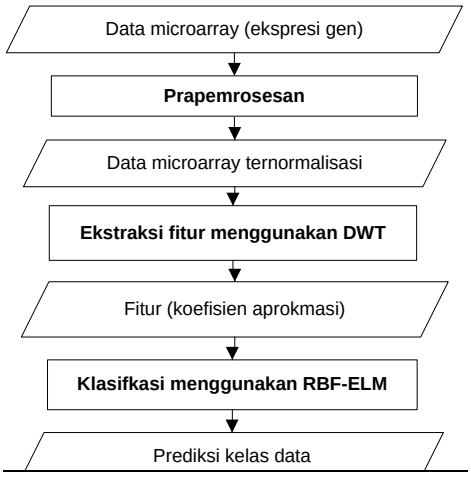
Extreme learning machine (ELM) dikembangkan oleh [17], yaitu algoritma pelatihan untuk *single-hidden-layer feedforward neural network* yang dapat mengatasi kelemahan pada *gradient descent learning*. ELM mampu melakukan pelatihan dalam waktu cepat karena tidak memerlukan iterasi, serta tetap menghasilkan kemampuan generalisasi yang baik [17]. Penggunaan ELM pada *feedforward neural network* dengan fungsi aktivasi sigmoid untuk klasifikasi data *microarray* multikelas telah dilakukan pada penelitian [16]. Hasilnya menunjukkan bahwa ELM mampu menghasilkan akurasi yang lebih baik daripada SVM saat jumlah kelas semakin banyak. [18] membandingkan penerapan ELM pada *feedforward neural network* yang menggunakan fungsi aktivasi sigmoid dengan ELM pada RBFN (*radial basis function network*) untuk klasifikasi data *microarray* dan menunjukkan bahwa ELM yang diterapkan pada RBFN dapat memberikan performa yang lebih baik. [19] mengembangkan ELM untuk *generalized single-hidden-layer feedforward network* (SLFN) yang memiliki performa lebih baik dibanding ELM semula dalam permasalahan [19].

Penelitian ini melakukan klasifikasi data *microarray* multikelas menggunakan DWT untuk ekstraksi fitur dan ELM untuk klasifikasi sebab DWT memiliki kemampuan yang baik

untuk ekstraksi fitur dan ELM memiliki performa yang bagus untuk klasifikasi multikelas seperti yang telah dijelaskan sebelumnya. ELM yang digunakan adalah ELM yang dikembangkan oleh [19] dan diterapkan pada RBFN karena penggunaan ELM pada RBFN memberikan hasil yang lebih baik daripada *feedforward neural network* dengan *additive sigmoid node* [18]. *Dataset microarray* yang digunakan dalam penelitian ini adalah *dataset* GCM (14 kelas, 16.063 gen) dan Subtypes-Leukemia (7 kelas, 12.600 gen).

2. METODE PENELITIAN

Data microarray direpresentasikan dalam bentuk matriks ekspresi gen berukuran $m \times n$ dengan m adalah banyaknya gen dan n adalah banyaknya sampel. Setiap baris menunjukkan *gene-expression vector*, yaitu ekspresi sebuah gen tertentu pada setiap sampel dan setiap kolomnya menunjukkan *sample-expression vector*, yaitu ekspresi setiap gen pada sebuah sampel tertentu. Terdapat tiga tahapan dalam penelitian ini yang digunakan untuk melakukan klasifikasi data microarray seperti pada Gambar 1, yaitu prapemrosesan, ekstraksi fitur dan klasifikasi itu sendiri.



Gambar 1 Tahapan klasifikasi data microarray

2.1 Prapemrosesan

Prapemrosesan bertujuan untuk meminimalkan variabilitas teknis pada saat eksperimen.

2.1.1 Prapemrosesan dataset GCM

Langkah-langkah untuk prapemrosesan pada *dataset* GCM adalah:

1. Koreksi nilai ekspresi gen
   Batas atas dan bawah nilai ekspresi gen diatur ke nilai tertentu seperti yang digunakan pada penelitian aslinya, [1]. Nilai ekspresi gen pada *dataset* yang kurang dari 20 diubah ke 20 karena nilai yang sangat kecil dapat memuat *noise* dan tidak *reproducible*, sedangkan nilai yang melebihi 16.000 dibulatkan ke 16.000 karena nilai yang sangat besar disebabkan oleh *saturation effect.*
2. *Filtering* gen
   Filtering gen dilakukan untuk menghilangkan gen yang tidak mengalami perubahan nilai siginifikan di antara sampel yang ada. Cara yang digunakan untuk melakukan *filtering* gen sama dengan cara yang digunakan pada penelitian aslinya [1]. Gen-gen yang lolos dari *filtering* gen untuk digunakan pada tahap berikutnya adalah gen-gen yang memenuhi kedua syarat berikut:
   a. Perbandingan nilai ekspresi maksimum dan minimum (max/min) untuk sebuah gen di antara semua sampel lebih dari atau sama dengan 5.
   b. Selisih nilai mutlak ekspresi maksimum dan minimum (|max-min|) untuk sebuah gen di antara semua sampel lebih dari atau sama dengan 500.

3. *Normalisasi antar array*
   Normalisasi antar array dilakukan dengan cara centering, yaitu mengurangi setiap nilai intensitas setiap *array* dengan nilai rata-rata intensitas lalu membaginya dengan nilai standar deviasi pada *array* tersebut, sehingga setiap *array* memiliki rata-rata nilai intensitas 0 dan standar deviasi 1. Metode ini paling sering digunakan untuk normalisasi data antar *array*.

2.1.1 Prapemrosesan dataset Subtypes-Leukemia

Langkah-langkah untuk melakukan prapemrosesan pada *dataset* Subtypes-Leukemia adalah sebagai berikut:

1. Koreksi nilai ekspresi gen
   Batas atas dan bawah nilai ekspresi gen diatur ke nilai tertentu seperti yang digunakan pada penelitian aslinya [20]. Nilai ekspresi gen kurang dari 100 dan *absent call* serta nilai ekspresi gen negatif diubah ke nilai 1, sedangkan nilai yang lebih dari 45.000 dibulatkan ke 45.000.
2. *Filtering* gen
   Cara yang digunakan untuk melakukan *filtering* gen sama dengan cara yang digunakan pada penelitian aslinya [20]. Gen-gen diloloskan dari proses *filtering* gen jika memenuhi kedua syarat berikut:
   a. Persentase *present call* (P) untuk sebuah gen di antara semua lebih dari atau sama dengan 1%.
   b. Selisih nilai mutlak intensitas maksimum dan minimum (|max-min|) untuk sebuah gen di antara semua sampel lebih dari atau sama dengan 100.
3. *Normalisasi antar array*
   Normalisasi antar *array* dilakukan dengan cara yang sama seperti prapemrosesan *dataset* GCM.

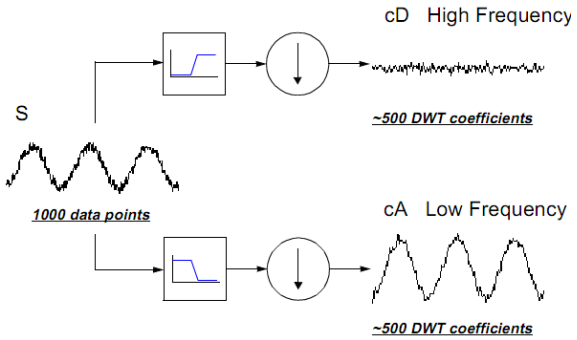
2.2 Ekstraksi Fitur

Ekstraksi fitur pada penelitian ini menggunakan *discrete wavelet transform* (DWT). DWT merupakan salah satu jenis transformasi wavelet yang melakukan pemampatan pada sinyal secara *dyadi*c (bilangan interger pangkat dua), yaitu 2, 4, 8, 16 dan seterusnya. DWT digunakan untuk mendekomposisi sinyal ke dalam beberapa frekuensi band. Contoh ilustrasi dekomposisi pada sebuah sinyal ditunjukkan pada Gambar 2 [8]. Dekomposisi adalah pembagian sinyal menjadi dua bagian menggunakan highpass decomposition filter (H) dan lowpass decomposition filter (L). Sinyal S dikonvolusi menggunakan filter H, kemudian dilakukan *downsampling* sehingga dihasilkan koefisien detail (cD1). Sinyal S juga didekomposisi menggunakan filter L, kemudian dilakukan *downsampling* sehingga dihasilkan koefisien aproksimasi (cA1) [7]. Konvolusi sinyal S dengan filter H meloloskan daerah frekuensi tinggi dari sinyal, sehingga cD1 berisi bagian sinyal yang berfrekuensi tinggi. Konvolusi sinyal S dengan filter L meloloskan daerah frekuensi rendah dari sinyal, sehingga cA1 memuat daerah sinyal berfrekuensi rendah yang menggambarkan identitas sinyal. *Downsampling* pada hasil konvolusi menyebabkan panjang cA1 maupun cD1 menjadi kurang lebih setengah dari panjang sinyal S [8]. Daerah frekuensi rendah dari sinyal memuat informasi yang penting karena daerah tersebut menggambarkan identitas sinyal [8]. Contoh dekomposisi DWT 3 level ditunjukkan pada Gambar 3.

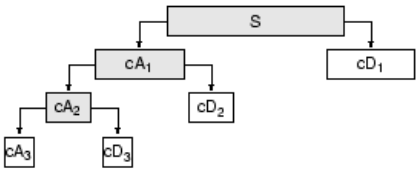
Setiap sampel hasil prapemrosesan didekomposisi menggunakan *wavelet filter* hingga level tertentu. Fitur yang digunakan selanjutnya adalah koefisien aproksimasi yang didapat pada level dekomposisi tersebut sebab koefisien aproksimasi dapat menggambarkan identitas sinyal [8]. *Wavelet filter* yang digunakan adalah Daubachies karena wavelet filter tersebut dapat digunakan untuk DWT dan wavelet tersebut digunakan dalam penelitian [9, 10, 11, 13, 15].

Beberapa parameter yang diujicoba dalam proses ekstraksi fitur untuk mendapatkan hasil klasifikasi terbaik, yaitu tipe wavelet filter Daubechies yang digunakan dan level dekomposisi. Tipe Daubachies yang diuji coba antara lain db2, db4, db8 dan db10. Sejumlah kemungkinan level dekomposisi yang diujicoba yaitu dari 3, 4, 5, 6 dan 7. Pada dekomposisi

level 2 ukuran data hanya menjadi sekitar 1/4 dari ukuran semula (ukuran data semula > 10000) sehingga percobaan dimulai dari level 3, sedangkan dekomposisi level 8 pengurangan jumlah data terlalu banyak sehingga dapat menghilangkan informasi yang penting dari data.



Gambar 2 Dekomposisi sinyal menggunakan DWT [8]

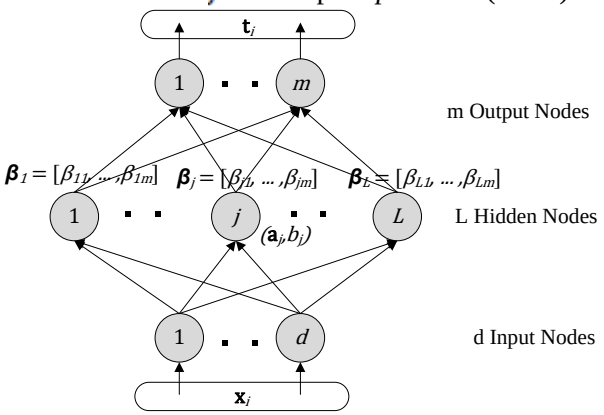


Gambar 3 Dekomposisi sinyal menggunakan DWT 3 level [8]

2.3 Klasifikasi

Metode klasifikasi yang digunakan adalah *Extreme Learning Machine* (ELM) yang diterapkan pada RBFN. Kelebihan ELM antara lain parameter *hidden node* dapat diinisialisasi secara random dan nilainya tidak perlu diubah-ubah selama proses pelatihan, sedangkan bobot output jaringan dapat dihitung melalui sebuah persamaan matematika, sehingga waktu pelatihan sangat singkat. ELM yang akan digunakan adalah ELM untuk *generalized* SLFN yang dikembangkan oleh [19]. ELM tersebut merupakan perbahruan dari ELM sebelumnya [17] dan memiliki performa yang lebih baik untuk klasifikasi.

Arsitektur jaringan yang akan digunakan untuk klasifikasi data *microarray* adalah RBFN seperti yang ditunjukkan pada Gambar 4. Jumlah *input node* adalah d dengan d adalah jumlah fitur atau ukuran sampel yang didapat dari hasil ekstraksi fitur. Jumlah *hidden node* adalah $L = 1000$ sesuai dengan jumlah yang digunakan dalam penelitian [19]. Jumlah *output node* yang digunakan adalah m dengan m adalah banyaknya kelas, pada *dataset* GCM $m = 14$, sedangkan pada *dataset* Subtypes-Leukemia $m = 7$. Algoritma pelatihan yang digunakan adalah ELM yang dikembangkan oleh [19]. Vektor $\mathbf{a}_j$ adalah *center* dan b_j adalah parameter *scale* atau *width* dari fungsi basis radial pada *hidden node* ke-j, sedangkan $\boldsymbol{\beta}_j$ adalah bobot koneksi antara sebuah hidden node ke-j ke setiap *output node* ($1..m$).



Gambar 4 Arsitektur RBFN untuk klasifikasi data microarray

Data ekspresi gen setelah proses ekstraksi fitur terdiri atas pasangan *input-output* $(\mathbf{x}_i, y_i)$, $i = 1, \dots, N$ dengan N adalah jumlah sampel pada data latih. Vektor $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{id}]^T \in \mathbf{R}^d$ adalah *feature set* dari sampel ke-i yang didapat dari proses ekstraksi fitur dan y_i nomor kelas (jenis kanker) untuk sampel ke-i. Nilai input data $\mathbf{x}_i$ yang digunakan untuk pelatihan diskalakan terlebih dahulu ke rentang [-1,1] sesuai yang disarankan oleh [19]. Jumlah *node* yang digunakan adalah sebanyak m kelas, sehingga untuk melakukan pelatihan, *output* y_i dikonversi terlebih dahulu ke dalam vektor $\mathbf{t}_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T \in \mathbf{R}^m$. Nilai pada vektor $\mathbf{t}_i$ yang indeksnya merupakan nomor kelas dari sampel ke-i (y_i) diberi nilai 1, sedangkan nilai pada indeks lainnya adalah -1, sebagai contoh sebuah sampel ke-i pada *dataset* Subtypes-Leukemia ($m = 7$) yang termasuk dalam nomor kelas 4 ($y_i = 4$) akan memiliki vektor output $\mathbf{t}_i = [-1, -1, -1, 1, -1, -1, -1]^T$. Dengan demikian, pasangan data latih yang di-*inputkan* ke RBFN adalah $(\mathbf{x}_i, \mathbf{t}_i)$, $i = 1, \dots, N$. Algoritma pelatihan ELM adalah sebagai berikut:

1) Men-*generate* parameter pada *hidden node* $(\mathbf{a}_j, b_j)_{j=1}^{L}$ secara *random*.
2) Menghitung *hidden layer output matrix* $\mathbf{H}$ menggunakan fungsi aktivasi *hidden node* $G(\mathbf{a}_i, b_i, \mathbf{x})$ Gaussian seperti persamaan (1) atau multiquadric seperti persamaan (2)

$$G(\mathbf{a}_j, b_j, \mathbf{x}) = \exp(-b_j \|\mathbf{x} - \mathbf{a}_j\|^2) \quad (1)$$

$$G(\mathbf{a}_j, b_j, \mathbf{x}) = (\|\mathbf{x} - \mathbf{a}_j\|^2 + {b_j}^2)^{1/2} \quad (2)$$

3) Bobot *output* $\boldsymbol{\beta}$ dihitung menggunakan persamaan (3).

$$\boldsymbol{\beta} = \mathbf{H}^{\mathbf{T}} \left(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^{\mathbf{T}} \right)^{-1} \mathbf{T} \quad (3)$$

$$\mathbf{f}(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\beta} = \mathbf{h}(\mathbf{x})\mathbf{H}^{\mathbf{T}} \left(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^{\mathbf{T}} \right)^{-1} \mathbf{T} \quad (4)$$

Selanjutnya penentuan output hasil klasifikasi dihitung menggunakan persamaan (4).

2.4 Ukuran Hasil Klasifikasi

Ukuran yang dapat digunakan untuk mengevaluasi performa *classifier* pada klasifikasi *multiclass* adalah akurasi atau *correct classification rate* (AC) dan *minimum sensitivity* (MS). Akurasi menunjukkan kemampuan *classifier* untuk mengklasifikasikan dengan benar di antara semua data masukan, sedangkan *minimum sensitivity* menunjukkan kemampuan minimal *classifier* untuk mengklasifikan dengan benar di antara data yang seharusnya berada pada kelas tertentu. Langkah-langkah untuk mengukur AC dan MS adalah [18] :

1) Matriks $\mathbf{P} = \{p_{ij}; \sum_{i,j=1}^{m} p_{ij} = N\}$ disusun dengan dimensi $m \times m$ (m adalah banyaknya *class-label*). Nilai p_{ij} menunujukkan jumlah data yang seharusnya berada pada *class* i, namun diklasifikan ke *class* j oleh *classifier* dan $p_i = \sum_{j=1}^{m} p_{ij}$ menunjukkan jumlah data yang seharusnya berada pada *class* i.
2) *Sensitivity* untuk *class* i dihitung dengan $S_i = {}^{p_{ii}}/_{p_i}$ menunjukkan perbandingan antara data yang diklasifikasikan dengan benar oleh *classifier* ke *class* i dengan jumlah data yang seharusnya berada di *class* i. *Minimum sensitivity* dihitung dengan $MS = \min\{S_i; i = 1, \dots, m\}$.
3) Akurasi atau *correct classification rate* dihitung dengan $AC = {}^{1}/_{N} \sum_{j=1}^{m} p_{jj}$.
4) *Classifier* yang terbaik ($\hat{g}$) di antara sejumlah *classifier* (g_i) adalah *classifier* yang mampu ditunjukkan seperti persamaan (5). Nilai $\overline{AC_{g_i}}$ menunjukkan rata-rata-rata nilai AC dan MS_{g_i} adalah rata-rata nilai MS yang didapat dari beberapa kali percobaan (*cross-validation*) menggunakan *classifier* g_i [18].

$$\hat{g} = \arg\max_{g_i} \frac{\overline{AC_{g_i}} + \overline{MS_{g_i}}}{2} \quad (5)$$

2.5 Dataset

Dataset microarray yang digunakan dalam penelitian ini adalah:

1. Dataset GCM (14 kelas, 16.063 gen), terdiri atas 144 data latih dan 46 data uji dengan distribusi data latih:data uji untuk tiap kelas adalah Breast (8:3), Prostate (8:2), Lung (8:3), Colorectal (8:3), Lymphoma (16:6), Bladder (8:3), Melanoma (8:2), Uterus_Adeno (8:2), Leukemia (24:6), Renal (8:3), Pancreas (8:3), Ovary (8:3), Mesothelioma (8:3), CNS (16:4) [1].
2. Dataset Subtypes-Leukemia (7 kelas, 12.600 gen), terdiri atas 144 data latih dan 46 data uji dengan distribusi data latih:data uji untuk tiap kelas adalah BCR-ABL (9:6), E2A-PBX1 (18:9), Hyperdiploid>50 (42:22), MLL (14:6), T-ALL (28:15), TEL-AML1 (52:27), Others (52:27) [20].

2.6 Rancangan Pelatihan dan Pengujian

Terdapat beberapa parameter yang harus ditentukan dalam algoritma klasifikasi, yaitu fungsi aktivasi dan nilai bilangan C untuk perhitungan bobot output jaringan. Fungsi aktivasi yang dicoba adalah Gaussian dan multiquadric, sedangkan nilai C yang dicoba adalah nilai-nilai dalam rentang $\{2^{-5}, 2^{-4}, \ldots, 2^{24}, 2^{25}\}$. Penentuan fungsi aktivasi dan niai C yang digunakan untuk mendapatkan hasil klasifikasi yang terbaik dilakukan melakukan k*-fold cross-validation* ($k = 5$) pada data latih. Nilai parameter yang memberikan hasil terbaik selanjutnya diterapkan untuk melakukan pelatihan pada semua data latih, selanjutnya model *classifier* yang didapat dari hasil pelatihan tersebut diterapkan untuk memprediksikan kelas pada data uji. Setelah itu, data latih dan data uji tersebut digabungkan menjadi satu, lalu dibagi secara random ke dalam jumlah data latih dan data uji dengan proporsi jumlah data per kelas yang sama. Pembagian tersebut diulang sebanyak 9 kali, kemudian proses pelatihan dan pengujian dilakukan pada masing-masing data latih dan data uji dari hasil pembagian secara random tersebut, sehingga totalnya ada 10 kali proses pengujian yang dilakukan. Hasil akhir yang dilaporkan dari penelitian ini adalah rata-rata akurasi dan *minimum sensitivity* yang didapat dari 10 kali pengujian tersebut.

3. HASIL DAN PEMBAHASAN

Proses 5-*fold cross-validation* dilakukan dengan mengulang tahap pelatihan dan validasi RBF-ELM sebanyak 3 kali, sehingga dalam satu *cross-validation* terdapat 15 kali percobaan yang dilakukan. Nilai akurasi (AC) dan *minimum sensitivity* (MS) dihitung darii rata-rata selama 15 kali percobaan. Setiap kali penggunaan wavelet filter, level dekomposisi dan fungsi aktivasi tertentu, dilakukan 31 kali *cross-validation* dengan nilai C yang berbeda, kemudian dipilih nilai C yang memberikan hasil klasfikasi terbaik. Proses *5-cross-validation* pertama kali pada data latih dilakukan untuk melihat pengaruh penggunaan fungsi aktivasi dalam proses klasifikasi. Hasilnya untuk dataset GCM maupun Subtypes-Leukemia menunjukkan bahwa penggunaan fungsi aktivasi multiquadric memberikan hasil (nilai AC dan MS) yang lebih baik dibanding fungsi aktivasi Gaussian. Selanjutnya, dilakukan cross-validation menggunakan wavelet filter dan level dekomposisi yang berbeda untuk mendapatkan nilai C terbaik pada saat penggunaan parameter tersebut. Fungsi aktivasi dan nilai C yang memberikan hasil terbaik pada 5 cross-validation selanjutnya digunakan untuk melatih seluruh data latih, hasil pelatihan kemudian diujikan pada data uji.

3.1 Dataset GCM

Hasil yang didapat untuk pengujian dataset GCM dengan 10 kali pengacakan data latih dan uji ditunjukkan pada Tabel 1. Hasil terbaik pada proses pengujian didapat saat menggunakan wavelet filter db8 dan level dekomposisi 5, yaitu rata-rata akurasi 75% ± 6,25% dan rata-rata *minimum sensitivity* 15% ± 19,95%. Hasil tersebut sedikit lebih tinggi dari penelitian [5], yaitu 67,39%, namun masih lebih rendah dari penelitian [16], yaitu 76,5% ±

5,3%. *Minimum sensitivity* yang dihasilkan pada sebagian besar percobaan memiliki nilai standar deviasi yang lebih besar dari nilai rata-ratanya menunjukkan bahwa pada percobaan yang telah dilakukan sering terdapat kelas yang tidak pernah diprediksi oleh *classifier*, sehingga nilai *minimum sensitivity* dari percobaan tersebut adalah 0.

Tabel 1 Hasil pengujian dataset GCM

Parameter				Pengujian				
Filter	**Level**	**Jumlah Fitur**	**C**	**Mean AC**	**SD AC**	**Mean MS**	**SD MS**	**(Mean AC + Mean MS) / 2**
db2	3	1370	4	0,7370	0,0452	0,0667	0,1405	0,4018
db2	4	687	4	0,7152	0,0652	0,1000	0,1610	0,4076
db2	5	345	0	0,7391	0,0410	0,1333	0,1721	0,4362
db4	3	1375	3	0,7435	0,0596	0,0667	0,1405	0,4051
db4	4	691	4	0,7304	0,0733	0,0667	0,1405	0,3986
db4	5	348	15	0,6370	0,0481	0,0000	0,0000	0,3185
db8	3	1382	4	0,7326	0,0384	0,0667	0,1405	0,3996
db8	4	698	4	0,7174	0,0703	0,0667	0,1405	0,3920
db8	**5**	**356**	**1**	**0,7500**	**0,0625**	**0,1500**	**0,1995**	**0,4500**
db10	3	1385	3	0,7478	0,0461	0,1000	0,1610	0,4239
db10	4	702	22	0,6652	0,0704	0,0667	0,1405	0,3659
db10	5	360	0	0,7196	0,0610	0,0833	0,1800	0,4014

Tabel 2 Hasil pengujian dataset Subtypes-Leukemia

Parameter				Pengujian				
Filter	**Level**	**Jumlah Fitur**	**C**	**Mean AC**	**SD AC**	**Mean MS**	**SD MS**	**(Mean AC + Mean MS) / 2**
db2	**3**	1353	18	**0,8929**	**0,0168**	**0,5106**	**0,2333**	**0,7017**
db2	4	678	4	0,8848	0,0254	0,4500	0,1581	0,6674
db2	5	340	6	0,8554	0,0303	0,3833	0,1125	0,6193
db4	3	1357	15	0,8955	0,0246	0,4444	0,2400	0,6700
db4	**4**	**682**	20	**0,8768**	**0,0288**	**0,5190**	**0,2029**	**0,6979**
db4	5	344	17	0,7795	0,0355	0,3741	0,1801	0,5768
db7	3	1362	22	0,8938	0,0247	0,3500	0,2284	0,6219
db7	4	687	20	0,8696	0,0295	0,3852	0,2271	0,6274
db7	5	350	5	0,8589	0,0234	0,3000	0,1315	0,5795
db10	3	1367	3	0,9036	0,0118	0,3000	0,2049	0,6018
db10	4	693	21	0,8500	0,0151	0,3167	0,1834	0,5833
db10	5	356	2	0,8902	0,0158	0,3833	0,1372	0,6368

3.2 Dataset Subtypes-Leukemia

Hasil yang didapat untuk pengujian dataset Subtypes-Leukemia dengan 10 kali pengacakan data latih dan uji ditunjukkan pada Tabel 2. Hasil tertinggi pada proses pengujian didapat saat menggunakan wavelet filter db2 dan level dekomposisi 3, yaitu 0,7017. Jika diamati hasil tersebut tidak jauh berbeda dengan hasil yang didapat saat menggunakan wavelet filter db4 dan level dekomposisi 4, yaitu 0,6979, di samping itu penggunaan level dekomposisi 4 menghasilkan jumlah fitur yang lebih kecil, kurang lebih setengah kali jumlah fitur dari level

dekomposisi 3, oleh karena itu disimpulkan bahwa hasil terbaik didapat saat penggunaan wavelet filter db4 dan level dekomposisi 4.

Hasil yang didapat dari penelitian ini, yaitu akurasi 87,68% ± 2,88% lebih rendah dari penelitian [5], yaitu 91,07%. Klasifikasi *dataset* Subtypes-Leukemia juga dilakukan oleh [17], tetapi hanya menggunakan 6 kelas, kelas terakhir yaitu kelas Others tidak digunakan. Untuk melihat perbandingan dengan penelitian [15], maka penelitian ini juga melakukan klasifikasi pada 6 kelas *dataset* Subtypes-Leukemia. Akurasi yang dihasilkan dari penelitian ini untuk klasifikasi *dataset* Subtypes-Leukemia 6 kelas adalah 97,56% ± 2,66% dan *minimum sensitivity* 83,59% ± 16,67%. Hasil tersebut sedikit lebih tinggi dibanding penelitian [15], yaitu akurasi 93,68%.

4. KESIMPULAN

Dari penelitian ini telah dibangun pengklasifikasi data *microarray* untuk *dataset* GCM dan Subtypes-Leukemia. Ekstraksi fitur dilakukan menggunakan DWT dengan wavelet filter hingga level dekomposisi tertentu, lalu menggunakan nilai koefisien aproksimasi level terakhir sebagai masukan untuk tahap klasifikasi. Klasifikasi dilakukan menggunakan RBFN dengan algoritma ELM. Hasil yang didapat untuk setiap *dataset* tersebut adalah sebagai berikut:

1) Hasil terbaik untuk *dataset* GCM, yaitu rata-rata akurasi 75% ± 6,25% dan rata-rata *minimum sensitivity* 15% ± 19,95% dicapai saat menggunakan wavelet filter db8 dan level dekomposisi 5 untuk ekstraksi fitur serta fungsi aktivasi multiquadric dan nilai $C = 2^1$ untuk algoritma pelatihan ELM. Hasil tersebut sedikit lebih tinggi dari penelitian [5], yaitu 67,39%, namun masih lebih rendah dari penelitian [16], yaitu 76,5% ± 5,3%.
2) Hasil terbaik untuk *dataset* Subtypes-Leukemia, yaitu rata-rata akurasi 87,68% ± 2,88% dan rata-rata *minimum sensitivity* 51,90% ± 20,29% dicapai saat menggunakan wavelet filter db4 dan level dekomposisi 4 untuk ekstraksi fitur serta fungsi aktivasi multiquadric dan nilai $C = 2^{20}$ untuk algoritma pelatihan ELM. Hasil tersebut lebih rendah dari penelitian [5], yaitu 91,07%. Jika dibandingkan dengan penelitian [15] menggunakan 5 *cross-validation* pada semua *dataset* dengan 6 kelas, hasil dari penelitian ini adalah 97,56% ± 2,66%, lebih tinggi dibanding penelitian [15], yaitu akurasi 93,68%.

5. SARAN

Beberapa hal yang dapat dilakukan untuk perbaikan penelitian ini antara lain:

1) Penelitian ini dapat diujicobakan pada *dataset microarray* lainnya, sehingga dapat diamati hasil yang didapat pada berbagai *dataset microarray* yang lain.
2) Perbaikan dapat dilakukan pada proses ekstraksi fitur dengan mencoba wavelet filter lainnya atau melakukan pemrosesan lebih lanjut pada nilai koefisien hasil dekomposisi sebelum dijadikan fitur masukan *classifier*, agar didapat hasil klasifikasi yang lebih baik.

DAFTAR PUSTAKA

[1] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S. and Golub, T.R., 2001, Multiclass Cancer Diagnosis Using Tumor Gene-expression Signatures, Proceedings of National Academy Sciences (PNAS), USA, 98, 26, 15149-15154.

[2] Ghanem, M., 2004, Course 341 Introduction to Bioinformatics Microarrays1: Microarray Technology, *Lecture Notes*, Department of Computing, Imperial College, London.

[3] Stekel, D., 2003, *Microarray Bioinformatics*, Cambridge University Press, New York.

[4] Quackenbush, J., 2006, Microarray Analysis and Tumor Classification, *New England J. Medicine*, 354, 23, 2463–2472.

[5] Wang, H., Zhang, H., Dai, Z., Chen, M. dan Yuan, Z., 2013, TSG: A New Algorithm for Binary and Multi-class Cancer Classification and Informative Genes Selection, *BMC Medical Genomics 2013*, 6(Suppl 1):S3.
[6] Lio, P., 2003, Wavelets in Bioinformatics and Computational Biology: State of Art and Perspectives, *Bioinformatics Review*, 19, 1, 2–9.
[7] Fugal, D. L., 2009, *Conceptual Wavelets in Digital Signal Processing,* Space and Sinyal Technical Publishing, Sandiego, California.
[8] Misiti, M., Misiti, Y., Oppenheim, G. dan Poggi, J-M., 2012, *Wavelet Tollbox™ User's Guide R2012b*, The MathWork Inc., Natick.
[9] Li, S., Liao, C. and Kwok, J.T., 2006, Wavelet-Based Feature Extraction for Microarray Data Classification, *Int. Joint Conference on Neural Networks*, Vancouver, BC, Canada, 16-21 Juli, 5028–5033.
[10] Liu, Y., 2008, Detect Key Gene Information in Classification of Microarray Data, EURASIP *J. Advances in Signal Processing*, 2008, 1-10.
[11] Rashid, S. and Maruf, G.M, 2011, An Adaptive Feature Reduction Algorithm for Cancer Classification Using Wavelet Decomposition of Serum Proteomic and DNA Microarray Data, *IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBM11)*, Atlanta, Georgia, USA, 12-15 November, 305-312.
[12] Huynh, H.T., Kim, J.J. dan Won, Y., 2007, DNA Microarray Classification with Compact Single Hidden-Layer FeedForward Neural Networks, *Proceedings of Frontiers in the Convergence of Bioscience and Information Technologies*, Cheju Island, Korea, 11-13 Oktober, 193-198.
[13] Zhang, S.W., Huang, D.S. and Wang, S.L., 2010, A Method of Tumor Classification Based on Wavelet Packet Transforms and Neighborhood Rough Set, *Computers in Biology and Medicine*, 40, 430–437.
[14] Stanikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D. dan Levy, S., 2005, A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis, *Bioinformatics*, 21, 5, 631-643.
[15] Liu, Y., 2009, Wavelet Feature Extraction for High-Dimensional Microarray Data, *Neurocomputing*, 72, 985–990.
[16] Zhang, R., Huang, G.B., Sundararajan, N. and Saratchandran, P., 2007, Multicategory Classification Using an Extreme Learning Machine for Microarray *Gene-expression* Cancer Diagnosis, *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, 4, 3, 485-495.
[17] Huang, G.B., Zhu, Q.Y. and Siew, C.K., 2004, Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks, *Proceedings of International Joint Conference on Neural Networks (IJCNN2004)*, Budapest, Hungary, 25–29 Juli.
[18] Monedero, J.S., Ramirez, M.C., Navarro, F.F., Fernandez, J.C., Gutierrez, P.A. and Martinez, C.H., 2010, On the Suitability of Extreme Learning Machine for Gene Classification Using Feature Selection, *Proceedings of International Conference on Intelligent Systems Design and Applications (ISDA)*, Cairo, 20 Nov – 1 Des, 507-512.
[19] Huang, G.B, Zhou, H., Ding, X. and Zhang, R., 2012, Extreme Learning Machine for Regression and Multiclass Classification, *IEEE Transaction on System, Man, and Cybernetics – Part B: Cybernetics*, 42, 2, 513–529.
[20] Yeoh, E.J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.H., Evans, W.E., Naeve, C., Wong, L. and Downing, J.R., 2002, Classification, Subtype Discovery, and Prediction of Outcome in Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling, *Cancer Cell*, 1, 133–143.