**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI**

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

**USULAN TUGAS AKHIR**

# **IDENTITAS PENGUSUL**

Nama : **Rindang Nisa Prasasti**

NRP : **5107 100166**

Dosen Wali : **Dr. Agus Zainal Arifin, S.Kom, M.Kom**

1. **JUDUL TUGAS AKHIR**

***“Implementasi Metode Klasifikasi Tumor Menggunakan Transformasi Wavelet dan Neighbourhood Rough Set”***

1. **LATAR BELAKANG**

Saat ini, analisa ekspresi gen dari data microarray menjadi salah satu tema menarik dalam penelitian. Penemuan microarray telah memberikan peluang untuk menganalisa ekspresi ribuan gen dalam satu waktu. Namun, analisa ekspresi gen memiliki kelemahan yaitu minimnya jumlah sampel dibandingkan dengan dimensi dari ruang vector yang dibentuk oleh ekspresi gen. Dalam satu studi teoritik yang berkaitan dengan jumlah sampel dan dimensional data (d), setidaknya disyaratkan lebih dari 2(d+1). Jika jumlah gen manusia sekitar 30 ribu, maka jumlah sampel yang disyaratkan setidaknya menjadi sekitar 60 ribu. Sementara biaya yang diperlukan untuk membaca ekspresi satu sampel dari sel manusia saja sangat tinggi. Maka dapat dibayangkan betapa mahalnya eksperimen analisa data microarray.

Salah satu alternatif untuk memecahkan masalah tersebut mengembangkan perangkat lunak (metode pattern recognition) yang dapat mengolah data dengan jumlah sampel relatif kecil dibandingkan dimensionalnya. Alternatif inilah yang membuat banyak metode diajukan untuk menganalisis data ekspresi gen, seperti klasifikasi dan clustering.

Seleksi fitur adalah salah satu metode pengolahan awal data (*pre*-*processing*) untuk menentukan subset fitur yang akan diolah pada tahap berikutnya. Seleksi fitur mereduksi jumlah fitur dan menghilangkan data yang tidak relevan, berlebihan, atau noise. Seleksi fitur telah diterapkan secara ekstensif pada klasifikasi tumor yang menggunakan data microarray.

Berdasarkan cara mengevaluasi kualitas dari subset fitur, terdapat dua kategori pada metode seleksi fitur, yaitu pendekatan *filter* dan pendekatan *wrapper*. Pada pendekatan filter, sekumpulan fitur yang baik dipilih sebagai hasil dari *pre-processing* berdasarkan karakteristik data itu sendiri dan tidak bergantung pada algoritma klasifikasi. Sedangkan pada pendekatan *wrapper*, dibutuhkan performa dari sebuah algoritma mining untuk mengevaluasi dan menentukan fitur mana yang dipilih. Akan tetapi, secara komputasional pendekatan *wrapper* jauh lebih mahal dibandingkan pendekatan *filter*. Oleh karena itu pendekatan *filter* lebih banyak digunakan pada seleksi fitur ekspresi gen data microarray.

Transformasi wavelet adalah dekomposisi suatu sinyal dengan keluarga dari basis orthonormal real yang diperoleh melalui translasi dan dilasi sebuah fungsi kernel yang dikenal sebagai *mother wavelet*. Transformasi wavelet multi resolusi dapat memproses data stasioner dan nonstasioner dan memiliki kemampuan multi resolusi yang bagus. Karena kelebihan inilah, wavelet telah digunakan secara efektif pada beberapa aplikasi bioinformatika. X. Zhou dkk [2] mengusulkan metode seleksi gen menggunakan kesamaan informasi dan transformasi wavelet. Mereka menggunakan transformasi untuk merepresentasikan data microarray. Setiap fitur (koefisien wavelet) dari transformasi wavelet berelasi dengan beberapa gen pada original gen data microarray. Kemudian, kesamaan informasi berdasarkan metode seleksi fitur digunakan untuk memilih fitur paling kuat dari sekian koefisien wavelet.

Tranformasi wavelet dirasa lebih sesuai untuk pengukuran struktur biological daripada tranformasi matematika yang lain karena tranformasi wavelet menyediakan representasi matematika yang ekonomis dan informatif dari banyak objek yang menarik. Tranformasi wavelet dianggap mampu menyediakan analisa ragam secara umum, yang dibutuhkan pada kasus analisis ekspresi gen pada data microarray.

Teori *rough set* pertama kali dijelaskan oleh Pawlak[4]. *Rough set* adalah pendekatan formal yang menyediakan algoritma efisien untuk menemukan pola yang tersembunyi dalam data, mereduksi data, mengevaluasi signifikansi data, dan menghasilkan sekumpulan aturan keputusan (*decision rules*) dari data. Kapabilitas r*ough set* dalammereduksi *cost* komputasi dan memperbaiki akurasi klasifikasi dianggap sesuai untuk membangun metode seleksi fitur yang efektif untuk data microarray yang berisi sejumlah kecil sampel dengan sejumlah besar gen pada masing-masing sampel.

Walaupun seleksi fitur pada *rough set* klasik adalah metode reduksi yang efektif, dalam penggunaannya, atribut harus didiskritkan terlebih dahulu sebelum direduksi dan hal in dapat memicu kemungkinan terjadinya kehilangan informasi. Namun, model *neighbourhood rough set* (NRS) mengecualikan prosedur diskritisasi sehingga kehilangan informasi tidak terjadi. Algoritma seleksi fitur berdasarkan *neighbourhood rough set* dapat menghilangkan fitur yang redundan dan tidak relevan.

*Small-*, *round-*, *blue-cell tumour* (SRBCT) adalah jenis neoplasma ganas yang terdiri dari *Ewing family of tumors* (EWS), *Burkitt lymphoma* (BL), *neuroblastoma* (NB)dan *rhabdomyosarcoma* (RMS). Dinamakan demikian karena kemunculannya yang sama pada histology rutin.

Tumor ini lebih banyak muncul pada anak-anak daripada orang dewasa. Tumor tersebut biasanya mewakili sel-sel terdiferensiasi. Dominasi pewarnaan biru terjadi karena sebagian besar sel terdiri dari nukleus dan hanya memiliki sedikit sitoplasma. Hal ini terlihat pada saat pewarnaan [hematoxylin](http://id.wikipedia.org/w/index.php?title=Hematoxylin&action=edit&redlink=1) dan [eosin](http://id.wikipedia.org/wiki/Eosin). Pemberian hematoxylin akan memberi warna biru pada [nukleus](http://id.wikipedia.org/w/index.php?title=Nukelus&action=edit&redlink=1) dan eosin memberi warna merah muda pada [sitoplasma](http://id.wikipedia.org/wiki/Sitoplasma).

1. **RUMUSAN MASALAH**

Rumusan masalah yang diangkat dalam Tugas Akhir ini adalah sebagai berikut:

1. Bagaimana menerapkan transformasi paket wavelet dalam proses ekstraksi fitur.
2. Bagaimana membentuk tabel keputusan dengan konsep *rough set.*
3. Bagaimana mereduksi tabel keputusan dengan algoritma FARNeM.
4. Bagaimana menerapkan tiga classifier, yaitu SVM-RBF, KNN, dan NEC.
5. Bagaimana mengevaluasi hasil klasifikasi dengan *4-fold cross-validation.*
6. **BATASAN MASALAH**

Permasalahan yang dibahas dalam Tugas Akhir ini memiliki beberapa batasan, yaitu sebagai berikut :

1. Data yang digunakan adalah *Small Round Blue Cell Tumor* (SRBCT). Data didapatkan dari http://research.nhgri.nih.gov/microarray/Supplement
2. Basis paket wavelet yang digunakan yaitu Haar, Debauchies 2, Debauchies 4, Debauchies 8, Biorthogonal 2.2, Biorthogonal 3.7, Biorthogonal 6.8.
3. Cara memperoleh ketujuh basis paket wavelet yang digunakan pada proses pengujian tidak dibahas dalam tugas akhir ini.
4. *Classifier*  yang digunakan adalah SVM-RBF, KNN, dan NEC.
5. Sistem dibangun dengan menggunakan perangkat lunak MATLAB.
6. **TUJUAN TUGAS AKHIR**

Tugas akhir ini bertujuan untuk mengimplementasikan penggunaan transformasi wavelet dan *neighborhood rough set* sebagai metode untuk mengklasifikasi tumor.

1. **MANFAAT TUGAS AKHIR**

Pengerjaan tugas akhir ini diharapkan dapat memberikan manfaat dalam penyelesaian masalah di bidang medis dan biologi dengan sudut pandang ilmu komputer.

1. **RINGKASAN TUGAS AKHIR**
   * + - 1. **Data Set**

Dataset SRBCT secara keseluruhan berisi 83 sampel yang terbagi dalam 4 kelas, yaitu *Ewing family of tumors* (EWS), *Burkitt lymphoma* (BL), *neuroblastoma* (NB)dan *rhabdomyosarcoma* (RMS). Setiap sampel pada dataset ini berisi 2308 nilai ekspresi gen. Dari 83 sample, terdapat 29 sampel untuk kelas EWS, 11 sampel untuk kelas BL, 18 sampel untuk kelas NB, dan 25 sampel untuk kelas RMS.[5]

* + - * 1. **Metode**

START

END

83 Sampel Tumor

7 fungsi wavelet

Evaluasi hasil klasifikasi menggunakan 4-*fold cross-validation*

Evaluasi hasil klasifikasi menggunakan 4-*fold cross-validation*

Proses ekstraksi fitur dengan *wavelet packet transform*

Membentuk 4 *decision table* sebagai data training

Klasifikasi menggunakan SVM-RBF

Proses perangkingan gen menggunakan *Kruskal-Wallis rank sum test*

Proses reduksi atribut menggunakan algoritma FARNeM

Proses reduksi atribut menggunakan algoritma FARNeM

Klasifikasi menggunakan SVM-RBF, K-NN, dan NEC

Proses ekstraksi fitur dengan *wavelet packet transform*

Gen terpilih

Fungsi wavelet teroptimal

*Decision* *table* teroptimal

*Classifier* teroptimal

Mengambil subset fitur optimal dari hasil reduksi atribut sebagai data testing

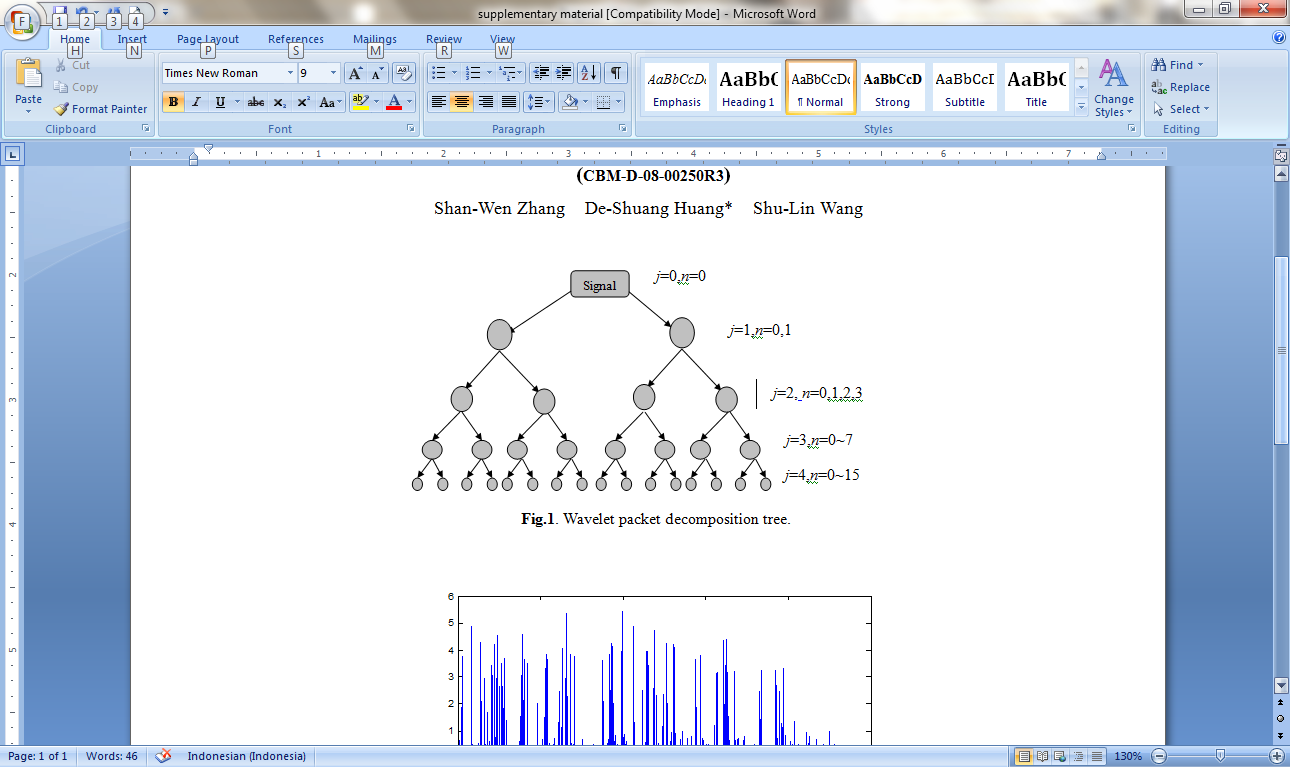
Fungsi wavelet, *decision table*, dan *classifier* yang menghasilkan akurasi tertinggi

83 Sampel Tumor

**Gambar 1. Flowchart klasifikasi tumor**

Penjelasan dari *flowchart* diatas sebagai berikut :

1. Input yang digunakan adalah 83 sampel data microarray
2. Proses ekstraksi fitur dijelaskan sebagai berikut :
3. Menghitung *4-layer full wavelet packet tree decomposition*.



**Gambar 2. *4-layer full wavelet packet tree decomposition***

1. Menghitung *wavelet packet decomposition coefficient* tiap node pada masing-masing layer
2. Menghitung *Mean*, *Energy*, dan *Entropy* tiap node pada masing-masing layer

*Mean :*

*Energy :*

*Entropy :*

1. Membentuk empat *decision table* () berdasarkan metode *neighborhood rough set* yang akan digunakan sebagai data training.

Karena 4-layer dekomposisi graf trasnformasi paket wavelet memiliki 30 *crunodes*, seperti yang ditunjukkan Gambar 1, tabel keputusan , , , dan masing-masing memiliki 30, 30, 60, dan 90 atribut.

1. Reduksi atribut menggunakan algoritma FARNeM. Algoritma FARNeM merupakan metode seleksi fitur baru yang didesain oleh Hu dkk [3] untuk memilih reduksi minimal, yang dijelaskan sebagai berikut :

|  |
| --- |
| **Input : dan**  // *: neighborhood rough decision table*  // *: plant sampel set*  // : *attribute condition*  // : *decision attribute*  // : *threshold* untuk mengontrol ukuran *neighborhood*  **Step 1 :**  // initialisasi diatur mengosongkan set dan merupakan *pool* berisikan atribut yang informatif  **Step 2 : For each**  **Computing**  // dimana didefinisikan , *dependency of subtype table*  *//* set C terhadap himpunan kosong ditetapkan menjadi nol.  **Step 3 : Selecting the informative gene satisfying**    **Step 4 : if**  *//* menambahkan optimal gen ke *reduction set*    **go to step 2**  **return ;**  **Step 5 : Algoritm end;**  **Output :**  // adalah reduksi dari set atribut kondisi |

Reduksi atribut dilakukan menggunakan setting parameter neighborhood , . Untuk satu konstanta , diperoleh satu subset atribut dari penggunaan algoritma FARNeM, jadi 100 subset dapat diperoleh saat bervariasi dari 0 sampai 1 dengan pergerakan 0.01.

1. Melakukan klasifikasi menggunakan tiga classifier, yaitu SVM-RBF, K-NN, dan NEC.
2. Evaluasi hasil klasifikasi menggunakan 4-*fold cross-validation*
3. Menganalisa fungsi wavelet, *decision table*, dan *classifier* yang menghasilkan akurasi tertinggi
4. Proses perangkingan gen pada 83 sampel data SRBCT dengan *Kruskal-Wallis rank sum test*. Hal ini dilakukan dikarenakan hasil eksperimen terkadang tidak stabil walaupun *wavelet packet transform* sangat fungsional untuk mereduksi noise. Dari proses ini diambil sejumlah gen dengan -*value* yang lebih besar.
5. Mengulang kembali langkah (2) hingga (5) dengan input sejumlah gen yang telah terpilih pada langkah (7) serta fungsi wavelet, *decision table*, dan *classifier* yang menghasilkan akurasi tertinggi
6. Menganalisa rentang serta parameter *classifier* yang menghasilkan akurasi tertinggi.

SVM dirasa sesuai untuk analisa ekspresi gen karena kemampuannya mengolah data berdimensi tinggi. Pada umumnya masalah dalam domain dunia nyata (*real world problem*) jarang yang bersifat *linear separable*. Kebanyakan bersifat non linear. Untuk menyelesaikan problem non linear, SVM dimodifikasi dengan memasukkan fungsi *Kernel*. Dalam non linear SVM, pertama-tama data dipetakan ke ruang vektor yang berdimensi lebih tinggi. Pada ruang vektor yang baru ini, hyperplane yang memisahkan *class* dapat dikonstruksikan.

1. **METODOLOGI**

Metodologi yang akan dilakukan dalam Tugas Akhir ini memiliki beberapa tahapan, diantaranya sebagai berikut:

1. Penyusunan Proposal Tugas Akhir

Tahap awal untuk memulai pengerjaan Tugas Akhir adalah penyusunan Proposal Tugas Akhir. Pada proposal ini, penulis mengajukan gagasan pembuatan sistem untuk mengklasifikasi tumor berdasarkan transformasi paket wavelet dan *neighbourhood rough set*.

1. Studi Literatur

Pada tahap ini dilakukan pencarian, pengumpulan, penyaringan, pembelajaran dan pemahaman literatur yang berhubungan dengan proses pengolahan citra dokumen terdegradasi, khususnya yang meliputi permasalahan mengenai wavelet dan*neighbourhood rough set*. Literatur yang digunakan dalam pengerjaan Tugas Akhir ini sebagian besar berasal dari internet berupa makalah ilmiah, tesis, artikel, materi kuliah, serta beberapa buku referensi.

1. Implementasi

Implementasi merupakan tahap untuk membangun sistem tersebut.

1. Pengujian dan Evaluasi

Pada tahap ini dilakukan uji coba terhadap sistem yang telah dibuat, mengamati kinerja sistem yang baru dibuat, serta mengidentifikasi kendala yang mungkin timbul.

1. Penyusunan Buku Tugas Akhir

Tahap terakhir merupakan penyusunan laporan yang memuat dokumentasi mengenai pembuatan serta hasil dari implementasi perangkat lunak yang telah dibuat.

1. **JADWAL KEGIATAN TUGAS AKHIR**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. | Kegiatan | Bulan | | | | | | | |
| September | | Oktober | | November | | Desember | |
| 1. | Penyusunan Proposal Tugas Akhir |  |  |  |  |  |  |  |  |
| 2. | Studi Literatur |  |  |  |  |  |  |  |  |
| 3. | Implementasi |  |  |  |  |  |  |  |  |
| 4. | Pengujian dan Evaluasi |  |  |  |  |  |  |  |  |
| 5. | Penyusunan Buku Tugas Akhir |  |  |  |  |  |  |  |  |

1. **DAFTAR PUSTAKA**
2. S.W. Zhang, D.S.Huang , S.L Wang. *A* method *of* tumor classification based on wavelet *packet transforms and neighborhood rough set*. 2010.
3. X. Zhou, X. Wang, E. Doughert. *Nonlinear-probit gene classification using mutual-information and wavelet based feature selection*. 2004.
4. Q. Hu, D. Yu, Z. Xie. *Neighborhood Classifiers.* 2008.
5. Z. Pawlak. *Roughsets, International Journal of Computer and Information Sciences*. 1982.
6. J. Khan , dkk. *Classification and diagnostic prediction of cancers using geneexpression profiling and artificial neural networks*. 2001.

**LEMBAR PENGESAHAN**

###### **Surabaya, 28 September 2011**

Menyetujui,

Dosen Pembimbing I

# **(Prof. Ir. Handayani Tjandrasa, M.Sc, Ph.D)**

# **( NIP. 19490823 197603 2 001)**

Dosen Pembimbing II

# **(Dr. Nanik Suciati, S.Kom, M.Kom )**

**(NIP. 19710428 199412 2 001)**