IMPLEMENTASI METODE *SUPPORT VECTOR MACHINE* UNTUK MELAKUKAN KLASIFIKASI PADA DATA BIOINFORMATIKA

(Studi kasus: Klasifikasi *Gene Expression* yang Terjangkit *Medullloblastoma* di Amerika)

TUGAS AKHIR

Lalu Bayu Dwi Cahyo

13611187

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2018

IMPLEMENTASI METODE *SUPPORT VECTOR MACHINE* UNTUK MELAKUKAN KLASIFIKASI PADA DATA BIOINFORMATIKA

(Studi kasus: Klasifikasi Gene Expression
yang Terjangkit Medullloblastoma di Amerika)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana
Jurusan Statistika

Lalu Bayu Dwi Cahyo

13611187

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2018

HALAMAN PERSETUJUAN PEMBIMBING
TUGAS AKHIR

Judul	:	Implementasi Metode *Support Vector Machine* Untuk Melakukan Klasifikasi Pada Data Bioinformatika (Studi Kasus: Klasifikasi *Gene Expression* yang Terjangkit *Medulloblastoma* di Amerika)
Nama Mahasiswa	:	Lalu Bayu Dwi Cahyo
Nomor Mahasiswa	:	13611187

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta,

Pembimbing, 3 Januari 2018

Dr.techn.Rohmatul Fajriyah, S.Si., M.Si

HALAMAN PENGESAHAN

TUGAS AKHIR

IMPLEMENTASI METODE *SUPPORT VECTOR MACHINE* UNTUK MELAKUKAN KLASIFIKASI PADA DATA BIOINFORMATIKA

(Studi Kasus: Klasifikasi *Gene Expression* yang Terjangkit *Medulloblastoma* di Amerika)

Nama Mahasiswa : Lalu Bayu Dwi Cahyo
Nomor Mahasiswa : 13611187

TUGAS AKHIR INI TELAH DIUJIKAN
PADA TANGGAL 13 FEBRUARI 2018

Nama Penguji	Tanda Tangan
1. Fitria Dyah Ayu Suryanegara, M.Sc., Apt.	.....
2. Ayundyah Kesumawati, S.Si, M.Si.	.....
3. Dr.techn.Rohmatul Fajriyah, S.Si., M.Si.	.....

Mengetahui,
Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Drs. Allwar, M.Sc. Ph.D.

KATA PENGANTAR

Assalamu'alaikum Wr.Wb.

Alhamdulillahirabbil'alamiin, puji syukur kehadirat Allah SWT yang telah memberikan hidayah, kesempatan, dan kemudahan kepada kita semua dalam menjalankan amanah yang menjadi tanggung jawab kita. Shalawat serta salam tak henti-hentinya kita panjatkan kepada junjungan kita Nabi Besar Muhammad SAW beserta seluruh keluarga dan sahabatnya, karena dengan syafa'atnya kita dapat hijrah dari zaman jahiliyah menuju zaman yang terang benderang.

Tugas akhir yang berjudul "Implementasi Metode *Support Vector Machine* Untuk Melakukan Klasifikasi pada Data Bioinformatika" ini sebagai salah satu syarat untuk memperoleh gelar sarjana Jurusan Statistika di Universitas Islam Indonesia. Dalam penyusunan skripsi ini penulis banyak mengalami hambatan, namun berkat bantuan, bimbingan, dan kerjasama yang ikhlas dari berbagai pihak, akhirnya skripsi ini dapat terselesaikan dengan baik.

Pada kesempatan ini penulis mengucapkan terima kasih dengan tulus kepada :

1. Bapak Nandang Sutrisno, SH., M.Hum., LLM., Ph.D, selaku Rektor Universitas Islam Indonesia.
2. Bapak Drs. Allwar, M.Sc., Ph.D. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia, Yogyakarta beserta seluruh jajarannya.
3. Bapak Dr. RB. Fajriya Hakim, S.Si, M.Si, selaku Ketua Jurusan Statistika beserta seluruh jajarannya.
4. Ibu Dr.techn.Rohmatul Fajriyah, S.Si., M.Si, selaku dosen pembimbing. Terimakasih atas waktu, tenaga, motivasi, ilmu, nasehat serta bimbingannya sehingga penulis dapat menyelesaikan tugas akhir ini dengan baik.
5. Dosen-dosen Statistika Universitas Islam Indonesia yang telah mendidik dan menginspirasi.

6. Lalu Sukardi, Partinah, Lalu Imam Patra Anjalo, dan Baiq Alivia Safira selaku keluarga, terima kasih atas segalanya dengan penuh kesabaran, ketulusan doa dan rizki dari setiap tetes keringat yang mengalir, dukungan dan kasih sayang.
7. Sahabat-sahabatku semua yang selalu ada untuk penulis selama masa kuliah, atas kebersamaan dari awal kuliah hingga sekarang terimakasih atas segala hal yang pernah dilalui bersama.
8. Teman – teman seperjuanganku di Jurusan Statistika 2013, terimakasih atas kebersamaannya selama ini, menuntut ilmu bersama kalian adalah pengalaman yang tak akan pernah terlupakan.
9. Semua pihak yang telah membantu dalam penyusunan skripsi ini yang tidak dapat disebukan satu persatu

Semoga segala bantuan, bimbingan dan pengajaran yang telah diberikan kepada penyusun mendapatkan imbalan dari Allah SWT. Tidak lupa penulis memohon maaf apabila selama dalam proses penyusunan tugas akhir ini terdapat kekhilafan dan kesalahan. Penulis menyadari sepenuhnya akan keterbatasan kemampuan yang dimiliki. Oleh karena itu, penulis mengharapkan adanya kritik dan saran yang membangun demi kesempurnaan penyusunan dan penulisan tugas akhir ini. Semoga tugas akhir ini dapat bermanfaat bagi semua yang membaca dan membutuhkannya. Akhir kata, semoga Allah SWT selalu melimpahkan rahmat serta hidayah-Nya kepada kita semua, Amin amin ya robbal ‘alamiin.

Wassalamu’alaikum, Wr.Wb.

Yogyakarta, 3 Januari 2018

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan disuatu perguruan tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 3 Februari 2018

Lalu Bayu Dwi Cahyo

IMPLEMENTASI METODE *SUPPORT VECTOR MACHINE* UNTUK MELAKUKAN KLASIFIKASI PADA DATA BIOINFORMATIKA

(Studi kasus: Klasifikasi *Gene Expression* yang Terjangkit *Medullloblastoma* di Amerika)

Lalu Bayu Dwi Cahyo

Program Studi Statistika Fakultas MIPA

Universitas Islam Indonesia

INTISARI

Meduloblastoma merupakan kelompok heterogen tumor ganas dari system saraf pusat (Central Nerveous System(CNS)). Penderita medulloblastoma antara 18%-20% dari semua tumor otak pada anak-anak serta 70% dari penderita medulloblastoma terdeteksi pada usia 10 tahun ke bawah. Fingkelstein melakukan pendataan pada 76 pasien usia muda yang terjangkit medulloblastoma di St Jude Children's Research Hospital disimpan dalam microarray dengan kode akses GSE37418. Penderita medulloblastoma memiliki 4 subrup yaitu WNT(wingless), SHH(sonic hedgehog), Subgroup 3, dan subgroup 4. Berdasarkan data finkelstein akan dilakukan klasifikasi menggunakan metode support vevtor machine (SVM) menggunakan program R.3.4.2 dengan tambahan packages dari bioconductor. Terdapat 4 tahap dalam pengolahannya yaitu Input data, Preprocessing yang memproses data microarray agar datanya representatif, Filtering yang memilih data dan mengkualifikasinya, dan analisis SVM untuk klasifikasi. Berdasarkan hasil analisis SVM mampu memprediksi kelas penderita dengan akurasi 95% dengan nilai AUC 98%. *Terdapat probe yang memiliki bobot cukup tinggi yang merupakan gen yang aktif pada berbagai tumor seperti insulinomas, kanker kerongkongan, dan kanker usus besar.*

Kata kunci: *Medulloblastoma*, Bionformatika, Klasifikasi, *Support Vector Machine*, *Microarray*

IMPLEMENTATION OF SUPPORT VECTOR MACHINE METHOD FOR CLASSIFICATION ON BIOINFORMATIC DATA

(Case study: Classification of Gene Expression Infected by Medullloblastoma in America)

Lalu Bayu Dwi Cahyo

Departement of Statistics, Faculty of Mathematic and Natural Science

Islamic University of Indonesia

ABSTRACT

Meduloblastoma is a heterogeneous group of malignant tumors of the central nerve system (CNS). Medulloblastoma is relatively rare, 18%-20% of all cancerous pediatric brain tumors, 70% of all pediatric medulloblastomas are diagnosed in children under age 10. According Fingkelstein research collected 76 data infected pediatric medulloblastoma at St. Jude Children Research Hospital stored in a microarray with access code GSE37418. Medulloblastoma has 4 subroup namely WNT(wingless), SHH(sonic hedgehog), Subgroup 3, dan subgroup 4. The Data has collected by Finkelstein will be classified using the support vevtor machine method on R.3.4.2 with additional packages of bioconductor. There is 4 step in processing that is input the data, preprocesiing to get representative data, Filtering for qualified data, and implement the SVM method for classification. Base on the result, SVM can predict sample with accuration 95% and AUC 98%. Some of the probe that has a high enough weight, the probe is a gene that is active in various tumors such as insulinomas, esophageal cancer, and colon cancer

Key word: Medulloblastoma, Bionformatic, Classification, Support Vector Machine, Microarray

BAB I
PENDAHULUAN

1.1.Latar Belakang

Perkembangan teknologi pada era globalisasi abad 21 diikuti dengan ledakan data dalam jumlah besar atau yang sering dikenal dengan sebutan *big data*. Big Data adalah sebuah sistem teknologi yang diperkenalkan untuk menanggulangi 'ledakan informasi' seiring dengan semakin bertumbuhnya ekosistem pengguna perangkat *mobile* dan data *internet*. *Big data* memiliki tiga istilah yaitu *volume* , *variety* , dan *velocity*. *Volume* berkaitan dengan ukuran media penyimpanan data yang sangat besar atau mungkin tak terbatas, *variety* berarti tipe atau jenis data yang dapat diakomodasi, dan *velocity* dapat diartikan sebagai kecepatan proses.(IBM,2017)

Perkembangan data terjadi pada hampir semua bidang ilmu tidak terkecuali bidang ilmu biologi molekuler yang dikenal dengan bioinformatika. Bioinformatika merupakan suatu terapan dari ilmu yang mempelajari penerapan ilmu di bidang komputer untuk mengelola dan menganalisis informasi biologis. Bioinformatika mencakup penerapan metode-metode statistika, dan informatika untuk memecahkan masalah-masalah biologis.(Attwood,1999)

Pembahasan tentang bioinformatika akan meliputi *sequencing* dan *microarray* (*microchip)*. Dalam situs resmi Pusat Informasi Bioteknologi National Amerika (NCBI) mendefinisikan *microarray* adalah hibridisasi dari sampel asam nukleat (target) untuk satu set *probe* oligonukleotida yang besar, yang melekat dengan solid, untuk menentukan urutan atau untuk mendeteksi variasi dalam urutan gen atau ekspresi atau untuk pemetaan gen.(NCBI,2017)

Keberadaan *microarray* yang merupakan *sequencing* dari suatu data organisme tentunya memiliki jumlah data dan dimensi yang banyak sesuai dengan penjelasan sebelumnya. *Microarray* tersebut berisikan tentang informasi dari kumpulan *probe* dari suatu gen.

Implementasi dari hal tersebut salah satunya ialah tersimpannya data gen pasien yang terjangkit kanker otak pada anak yaitu *medulloblastoma*.

Meduloblastoma merupakan kelompok heterogen tumor ganas dari system saraf pusat *(Central Nerveous System*(CNS)) yang menyerang laki-laki usia muda. Penyakit ini tergolong langka dan agresif, ditandai dengan kecendrungan yang metastatis.(ABTA,2015)

Berdasarkan *e-book* berjudul *medulloblastoma* yang dirilis ABTA(American Brain Tumor Association) tahun 2015, *medullobastoma* relatif jarang terjadi, terhitung kurang dari 2% dari semua tumor otak primer (tumor yang dimulai di otak atau di permukaannya), antara 18%-20% dari semua tumor otak yang menyerang anak-anak. *Medulloblastoma* adalah tumor otak ganas yang paling sering pada anak usia empat tahun kebawah dan kedua terbanyak menyerang anak usia 5-14 tahun serta 70% dari penderita *medulloblastoma* terdeteksi pada usia 10 tahun ke bawah.

Perkembangan terbaru dalam biologi molekuler dari *medulloblastoma* menunjukkan bahwa klasifikasi tumor embrio yang semata-mata berdasarkan kriteria histologi dan klinis tidak cukup memadai. Pemahaman yang lebih baik tentang mekanisme pengendalian pertumbuhan dalam pengembangan dan kemajuan *medulloblastoma* akan memungkinkan klasifikasi yang lebih baik, yang mengarah pada perbaikan terapi yang ada, serta pengembangan pendekatan terapeutik baru. (Rosi, 2008). Terdapat 4 subgroup dalam penyakit *medulloblastoma* yakni WNT (*wingless*) , SHH (*sonic hedge hog*), Group 3 dan Group 4.

Penelitian Finkelstein yang dipublis tahun 2012 diunnggah pada halaman NCBI dengan kode GSE37418. Fingkelstein melakukan pendataan pada 76 pasien usia muda yang terjangkit *medulloblastoma* dengan judul *"Novel mutations target distinct subgroups of medulloblastoma"*. Penelitian Finkelstein mengklasifikasikan penderita ke dalam 4 kelas WNT, SHH, *group* 3 dan *group* 4.

Perbedaan kelas pada setiap penderita *medulloblastoma* memiliki jenis perlakuan pada tahap pengobatan baik terapi dan target biologis, oleh karena itu perlu dilakukan klasifikasi. Penelitian Finkelstein diangkat dalam

penelitian untuk diklasifikasikan dengan menggunakan metode *Support Vector Machine*.

1.2. Rumusan Masalah

Berdasarkan latar belakang dan fokus masalah, rumusan masalah yang akan dikaji dalam penelitian ini ialah:

1. Bagaimana mengolah data *bioinformatics*?
2. Bagaimana hasil klasifikasi gen yang terbentuk dengan menggunakan SVM?
3. Bagaimana model SVM yang tepat untuk klasifikasi pasien yang ada?

1.3.Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah:

1. Bahasa pemrograman dan *software* yang digunakan dalam penelitian ini adalah *software* R 3.4.2 dengan beberapa *package bioinformatics*.
2. Melakukan klasifikasi dengan menggunakan *hyperplane* yang terbentuk berdasarkan dimensi data yang terangkum dalam algoritma SVM *linear*.
3. Menjelaskan hasil analisis yang berkaitan dengan data gen *microarray*.

1.4.Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Mengetahui data *bioinformatics* dan cara mengolah data tersebut.
2. Mengklasifikasikan data gen *microarray* menggunakan SVM linear.
3. Mengetahui kelas dari data yang baru berdasarkan model SVM yang terbentuk.

1.5.Manfaat Penelitian

Adapun manfaat dilakukan penelitian ini adalah:

1. Secara teoritis

   Hasil penelitian ini diharapkan dapat digunakan oleh berbagai kalangan terutama untuk disiplin ilmu bioinformatika yang didalamnya termasuk ilmu-ilmu hayati (kimia, biologi, farmasi, dan kedokteran), fisika, statistika, dan teknologi informasi.

2. Secara prkatek

   Hasil penelitian ini diharapkan mampu memberikan informasi kepada berbagai kalangan terkait pengelompokkan dan pola yang dibentuk dan dapat memprediksi kelompok dari data yang baru.

BAB II
TINJAUAN PUSTAKA

2.1.Penelitian dengan *Support Vector Machine*

Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 di *Annual Workshop on Computational Learning Theory*. Konsep dasar SVM sebenarnya merupakan kombinasi harmonis dari teori-teori komputasi yang telah ada puluhan tahun sebelumnya, seperti margin *hyperplane* (Duda & Hart (1973), Cover (1965), Vapnik (1964), dan sebagai nya.), kernel (Aronszajn, 1950) dan konsep-konsep pendukung yang lain. Belum pernah ada upaya merangkaikan komponen-komponen tersebut hingga tahun 1992.

Nugroho (2003) menulis makalah dengan judul Suport Vector Machines : Teori Aplikasinya dalam Bioinformatika. Berdasarkan penelitian tersebut aplikasi SVM pada bioinformatika, khususnya analisa ekspresi gen yang diperoleh dari eksperimen *microarray* terhadap pasien penderita penyakit kanker. Penelitian tersebut diangkat dari eksperimen yang dilakukan oleh *group* Terrence S. Furey, dengan tujuan memakai SVM untuk mengklasifikan apakah suatu pasien terkena penyakit kanker atau tidak, berdasarkan hasil analisa *microarray* terhadap sel pasien tersebut.

SVM menunjukkan hasil yang lebih baik daripada *perceptron*. Walaupun demikian, dikarenakan jumlah sampel yang relatif sedikit, hasil eksperimen itu belum dapat memberikan kesimpulan final bahwa SVM *superior* terhadap *perceptron*.

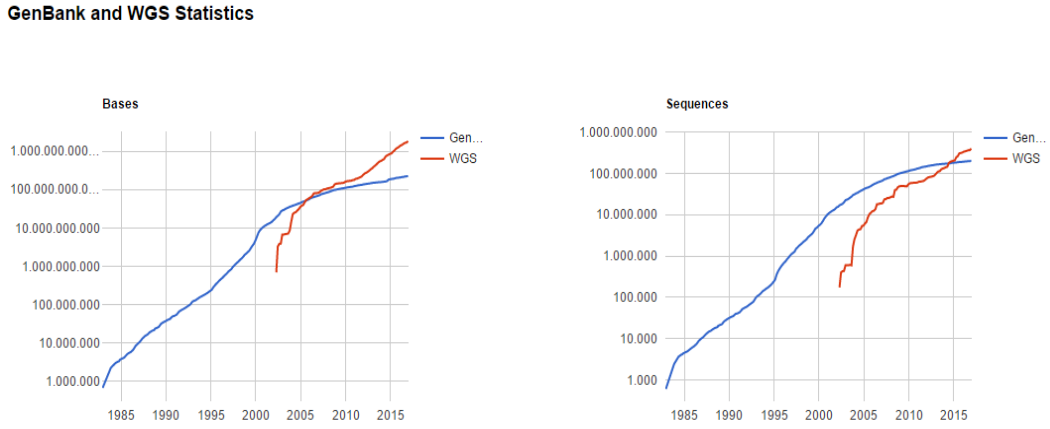
Brown (1999) juga menulis jurnal yang diterima pada 15 November 1999 di *Stanford University School of Medicine, Stanford, CA*, menuliskan penerapan SVM untuk data bioinformatika yang berjudul "*Knowledge-based analysis of microarray gene expression data by using support vector machines*". Penelitian tersebut menggunakan metode SVM dan analisisi jalur. Makalah tersebut menyimpulkan bahwa SVM dapat secara akurat mengklasifikasikan gen-gen ke dalam beberapa kategori-kategori fungsional

dan membuat prediksi untuk mengidentifikasi fungsi dari gen-gen *unannotated yeast*.

Terrence (2000) melalukan penelitian dengan judul "*Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data*". Penelitian Terrence digunakan metode SVM dan perceptron. Berdasarkan penelitian tersebut didapatkan hasil SVM bisa mengklasifikasikan tipe-tipe jaringan dan sel, seperti halnya algoritma *perceptron*. Selain itu SVM juga dapat digunakan untuk mengidentifikasi data yang *mis-labeled*. Dalam penelitian ini diaplikasikan diaplikasikan *simple kernel*.

2.2.Penelitian Terkait *Bioinformatics*

Bioinformatics mengalami perkembangan yang signifikan. Hal ini dibuktikan berdasarkan data *bionformatics* yang tekumpul dari tahun ke tahun yang dicatat oleh Kantor Pusat Informasi Bioteknologi Nasional (NCBI) di Amerika Serikat yaitu:



Sumber: https://www.ncbi.nlm.nih.gov/genbank/statistics

Gambar 2. 1 Perkembangan Data *Bioinformatics*

Aprijani (2004) dalam buku berjudul "Bioinformatika: Perkembangan, Disiplin Ilmu dan Penerapannya di Indonesia" menjelaskan bioinformatika merupakan perpaduan dari disiplin ilmu biologi molekul,

matematika dan teknik informasi. Disimpulkan bioinformatika merupakan aplikasi dari komputasi dan analisis data pada data–data biologi molekul. Bioinformatika mempunyai potensi untuk berkembang karena sifatnya multi disipliner, akan tetapi bioinformatika di Indonesia belum memasyarakat.

Lesk (2011) dalam atikel berjudul "*Bioinformatics*" menjelasakan bioinformatika ialah suatu persilangan ilmu pengetahuan yang merujuk kepada data biologi dan teknik pengumpulan informasi, distribusi, dan analisis untuk mendukung gabungan penelitian ilmiah, termasuk biomedicine.

Pengembangan algoritma yang efisien untuk mengukur kemiripan *sequence* merupakan tujuan dari bioinformatika. Algorima Needleman-Wunsch yang berdasar pada pemrograman dinamis menjamin untuk menemukan keselarasan pasangan *sequence* yang optimal. Algoritma tersebut pada dasarnya membagi masalah besar (*full sequence*) menjadi serangkaian masalah yang lebih kecil (*short sequence segment*) dan menggunakan solusi masalah yang lebih kecil untuk membangun solusi masalah besar. Kesamaan dalam sequence dinilai dalam suatu matriks dan algorithmanya memungkinkan untuk mendeteksi kesenjangan dalam keselarasan *sequence*.

Penelitian Alshamlan (2013) dengan judul " *A Study of Cancer Microarray Gene Expression Profile:Objectives and Approaches*". Berdasarkan hasil penelitian tersebut disimpulkan *Profiling* data *gene expression microarray* merupakan teknik yang efisien pada pengklasifikasian penyakit kanker. Secara umum tujuan dan metode yang digunakan adalah *gene finding, class discovery,*dan *class prediction*

2.3. Penelitain teerkait Medulloblastoma

Northcott (2014) melakukan penelitian dengan judul *Medulloblastoma comprises four distinct molecular variants*. Berdasarkan pnelitaiannya tumor otak merupakan penyebab utama kematian terkait kanker pada anak-anak dan medulloblastoma merupakan penyakit tumor otak pada anak-anak yang paling ganas. Berdasarkan *gene expression* dan

penyimpangan DNA 103 penderita medulloblastoma, dengan mengaplikasikan metode statistika (Anova, NMF, PCA, SubMap, analisis jalur dan *hierarchical clustering*) teridentifikasi 4 group *non overlapping molecular variant*, yaitu WNT, SHH, *group* C, dan *group* D.

Upadhyay (2014) menyebutkan dalam tesis dengan judul "*Pediatric Medulloblastoma: Molecular biology, correlation with histopathological and clinical outcome*" anak-anak dengan medulloblastoma bila di treatment dengan eksisi bedah radikal dan terapi adjuvant secara jangka panjang hasilnya lebih signifikan.

Hasil terbaik dan terburuk dari *treatment* masing-masing ada pada kelompok *wnt pathway* dan *nonwnt/nonshh*, anak-anak pada kelompok *shh* hasilnya akan berada diantara keduanya.

Studi di India berdasarkan *molecular subtyping* membuktikan bahwa *molecular subtyping* tersebut: layak, hemat biaya dan nilai tambah pada *prognosticating outcome* untuk anak-anak penderita medulloblastoma.

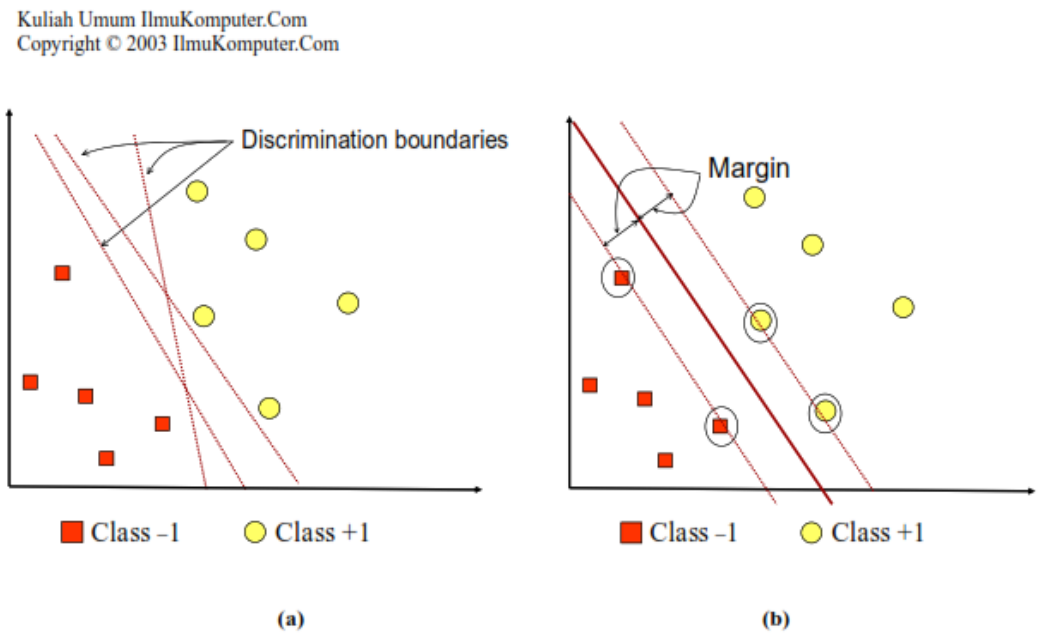
BAB III
LANDASAN TEORI

3.1.*Support vector machine*

SVM(*support vector machine*) dalam *mechine learning* dikenal juga dengan *support vector network* yang merupakan metode *supervised* terkait dengan *learning algorithm* untuk analisa pola data yang digunakan untuk klasifikasi dan regresi. (Mohammed, 2017)

3.1.1. *Support Vector Machine linear* (biner)

Metode SVM menggunakan fungsi dot produk. SVM merupakan usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah *class* pada *input space.*(Nugroho,2003).



Gambar 3.1 Penentuan *Hyperplane* Terbaik

Gambar 3.1a memperlihatkan beberapa *pattern* yang merupakan anggota dari dua buah *class* : +1 dan –1. *Pattern* yang tergabung pada *class* –1 disimbolkan dengan warna merah (kotak), sedangkan *pattern* pada *class* +1, disimbolkan dengan warna kuning(lingkaran). Permasalahan klasifikasi dapat diterjemahkan dengan usaha menemukan garis (*hyperplane*) yang

memisahkan antara kedua kelompok tersebut. Berbagai alternatif garis pemisah (*discrimination boundaries*) ditunjukkan pada gambar 3.1a.

Hyperplane pemisah terbaik antara kedua *class* dapat ditemukan dengan mengukur margin *hyperplane* dan mencari titik optimum *hyperplane* tersebut. Margin adalah jarak antara *hyperplane* dengan *pattern* terdekat dari masing-masing *class*. *Pattern* yang paling dekat disebut sebagai *support vector*. Garis solid pada gambar 3.1b menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua *class*, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah *support vector*. Usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada SVM.

Misal data yang tersedia dinotasikan sebagai $\vec{x}_i \in \Re^d$ sedangkan label masing-masing dinotasikan $y_i \in$ {-1,+1} untuk $i = 1,2, \dots, l$ dimana l adalah banyaknya data. Diasumsikan kedua kelas -1 dan +1 dapat terpisah secara sempurna oleh *hyperplane* berdimensi d, yang didefinisikan:

$$\vec{w}.\vec{x} + b = 0 \tag{3.1}$$

Pattern $\vec{x}$ yang termasuk *class* -1 (sampel negatif) dapat dirrumuskan sebagai *pattern* yang memenuhi pertidaksamaan

$$\vec{w}.\vec{x} + b \leq -1 \tag{3.2}$$

Sedangakan *pattern* $\vec{x}$ yang termasuk *class* +1 (sampel positif)

$$\vec{w}.\vec{x} + b \geq +1 \tag{3.3}$$

Margin terbesar dapat ditemukan dengan mengoptimalkan nilai jarak antara *hyperplane* dan titik terdekatnya, yaitu $1/\|\vec{w}\|$. Hal tersebut dapat dirumuskan sebagai *Quadratic Proraming (QP) problem*, yaitu mencari titik minimal persamaan (3.4), dengan memperhatikan *constraint* persamaan (3.5).

$$\min_{\vec{w}} \tau(w) = \frac{1}{2} \|\vec{w}\|^2 \tag{3.4}$$

$$y_i(\vec{w}.\vec{x}_i + b) - 1 \geq 0, \qquad \forall_i \tag{3.5}$$

data *input* dinotasikan x_i, adalah keluaran dari data $x_i, \vec{w}, b$ adalah parameter-parameter yang di cari nilainya. Formulasi di atas, ingin meminimalkan fungsi tujuan (*obyektif function*) $\frac{1}{2}\|\vec{w}\|^2$ atau memaksimalkan kuantitas $\|\vec{w}\|^2$ dengan memperhatikan pembatas sebgaimana persamaan 3.2 dan 3.3. Bila output data $y_i = +1$, maka pembatas menjadi $\vec{w}.\vec{x} + b \geq +1$. Sebaliknya bila $y_i = -1$, pembatas menjadi $\vec{w}.\vec{x} + b \leq -1$.

Permasalahan tersebut dapat dipecahkan dengan berbagai teknik komputasi, di antaranya *Lagrange Multiplier*.

$$L(w,b,\alpha) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} \alpha_i \left(y_i\left((\vec{x}_i.\vec{w} + b\right) - 1\right)\right) \tag{3.6}$$

$$L(w,b,\alpha) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (x_i.\vec{w}) - b\sum_{i=1}^{l} \alpha_i y_i + \sum_{i=1}^{l} \alpha_i \tag{3.7}$$

Tambahan konstrain, $\alpha_i \geq 0$ (nilai dari koefisien *lagrange*). Meminimumkan L terhadap w dan b.

$$\frac{\partial}{\partial b} L(w,b,\alpha) = 0 \tag{3.8}$$

$$\frac{\partial}{\partial w} L(w,b,\alpha) = 0 \tag{3.9}$$

Dari persamaan 3.8 dan persamaan 3.9 diperoleh persamaan berikut:

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{3.10}$$

$$\vec{w} = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{3.11}$$

Nilai $\vec{w}$ sering kali besar, tetapi nilai α_i terhingga. Untuk itu, formula *lagrangian* Lp (*primal problem*) diubah kedalam *dual problem*. Dengan mensubsitusikan persamaan 3.11 ke LP diperoleh *dual problem* L_d dengan konstrain berbeda.

$$L_d(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{i=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i \vec{x}_j \tag{3.12}$$

Dimana α_i adalah *Lagrange Multiplier* $\alpha_i \geq 0$. Nilai optimal dari persamaan 3.12 dapat dihitung dengan meminimalkan L terhadap $\vec{w}$ dan b, dan memaksimalkan L terhadap α_i, dengan memperhatikan sifat bahwa pada titik optimal *gradient* $L = 0$ persamaan 3.12 dapat dimodifikasi sebagai maksimalisasi problem yang hanya mengandung α_i, sebagaimana terlihat pada persamaan 3.13 dan 3.14 dibawah ini.

$$\max_{\alpha} L_d = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i \vec{x}_j \tag{3.13}$$

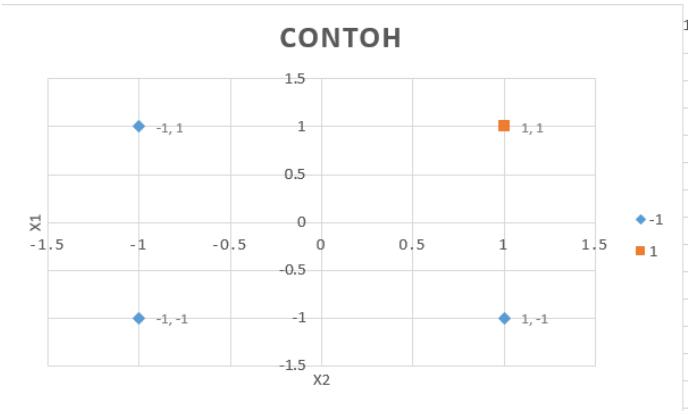
$$\alpha_i \geq 0 (i = 1,2,\ldots,l) \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{3.14}$$

Berdasarkan persamaan dia atas, maka akan diperoleh α_i yang kebanyakan bernilai positif yang disebut sebagai *support vector*.

Contoh:

Tabel 3.1 Contoh data

x1	x2	y
1	1	1
1	-1	-1
-1	1	-1
-1	-1	-1



Gambar 3.2 *Plot* Contoh Data

Gambar 3.2 menunjukkan terdapat 4 titik yang terdiri atas kelas -1 dan kelas 1. Titik-titik tersebut digunakan untuk mencari pemisah antara data positif dan negatif.

Diketahui:

$$\frac{1}{2}\|w\|^2 = \frac{1}{2}(w_1^2 + w_2^2)$$

syarat:

$$y_i(w.x_i + b) \geq 1, \quad i = 1,2,\dots,n$$

$$y_i(w_1.x_1 + w_2.x_2 + b) \geq 1$$

Sehingga dipatkan beberapa persamaan:

1. $(w_1 + w_2 + b) \geq 1$ untuk $y_1 = 1$, $x_1=1$, $x_2=1$
2. $(-w_1 + w_2 - b) \geq 1$ untuk $y_2 = -1$, $x_1=1$, $x_2=-1$
3. $(w_1 - w_2 - b) \geq 1$ untuk $y_3 = -1$, $x_1=-1$, $x_2=1$
4. $(w_1 + w_2 - b) \geq 1$ untuk $y_1 = -1$, $x_1=-1$, $x_2=-1$

Berdasarkan 4 persamaan di atas maka akan dicari nilai dari setiap variabel. Menjumlahkan persamaan 1 dan 2 maka didapat:

$$(w_1 + w_2 + b\ ) \geq 1$$

$$(-w_1 + w_2 - b\ ) \geq 1 \quad +$$

$$2w_2 = 2$$

$$w_2 = 1$$

Menjumlahkan persamaan 1 dan 3:

$$(w_1 + w_2 + b\ ) \geq 1$$

$$(w_1 - w_2 - b\ ) \geq 1 \quad +$$

$$2w_1 = 2$$

$$w_1 = 1$$

Menjumlahkan persamaan 2 dan 3:

$$(-w_1 + w_2 - b\ ) \geq 1$$

$$(w_1 - w_2 - b\ ) \geq 1 \quad +$$

$$-2b = 2$$

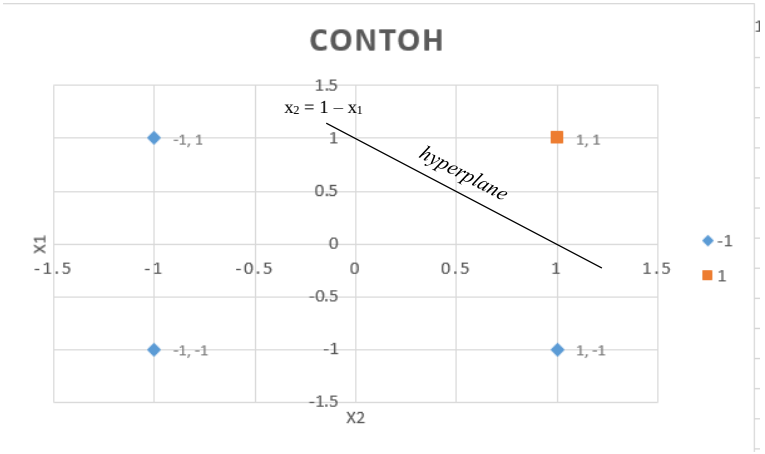
$$b = -1$$

Berdasarkan hasil di atas didapatkan persamaan:

$w_1.x_1 + w_2.x_2 + b = 0$

$x_1 + x_2 - 1 = 0$

$x_2 = 1 - x_1$



Gambar 3.3 *Hyperplane*

Gambar 3.3 menunjukkan suatu *hyperplane* pada contoh data. *Hyperplane* tersebutlah yang digunakan untuk memisah kelas data anatar positif dan negatif. Contoh diatas diketahui bahwa pengolahan data berdasarkan bentuk *data frame*, hal tersebut yang peneliti lakukan dalam mengaplikasikan metode SVM pada data bioinformatika dengan menggunakan program R.

3.1.2. Soft Margin

Penjelasan di atas berdasarkan asumsi bahwa kedua belah kelas terpisah secara sempurna oleh *hyperplane*. Dua buah kelas tidak selalu terpisah secara sempurna. Hal tersebut menyebabkan *constraint* pada persamaan 3.5 tidak terpenuhi, sehingga optimasi tidak terpenuhi dilakukan. Untuk mengatasi masalah ini, SVM dirumuskan ulang dengan menggunakan teknik *softmargin*. Persamaan 3.5 dimodifikasi menggunakan *softmargin* dengan memasukkan variabel *slack* ξ_i ($\xi_i > 0$) sebagai berikut:

$$y_i(\vec{w}.\vec{x}_i + b) - 1 \geq 1 - \xi_i \ , \forall_i \tag{3.15}$$

Dengan demikian persamaan 3.4 diubah menjadi:

$$\min_{\vec{w}} \tau(w) = \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^{l} \xi_i \tag{3.16}$$

Parameter C dipilih untuk mengontrol *trade off* antara margin dan *error* klasifikasi ξ. Nilai C yang besar berarti akan memberikan penalti yang lebih besar terhadap *error* klasifikasi tersebut.

3.1.3. SVM *Multiclass*

SVM saat pertama kali diperkenalkan oleh Vapnik, hanya dapat mengklasifikasikan data ke dalam dua kelas (klasifikasi biner). Penelitian lebih lanjut mengembangkan SVM sehingga bisa mengklasifikasi data yang memiliki lebih dari dua kelas terus dilakukan. Terdapat pilihan untuk mengimplementasikan *multiclass* SVM yaitu dengan menggabungkan beberapa SVM biner atau menggabungkan semua data yang terdiri dari beberapa kelas ke dalam sebuah bentuk permasalah optimasi. Pendekatan yang kedua permasalahan optimasi yang harus diselesaikan jauh lebih rumit. Ada dua metode yang umum digunakan untuk mengimplementasikan multi *class* SVM dengan pendekatan yaitu metode "*one-against-all*" dan metode " *one-against-one*".

a. One-Against-All

Metode *One-Against-All* mebangun k buah model SVM biner dengan k adalah jumlah kelas. Model SVM ke-m dilatih dengan semua contoh kelas dengan m label positif dan yang lainnya label negatif. Berdasarkan hal tersebut data latih $l\ (x_1,y_1), \ldots, (x_l,y_l)$ dimana $\vec{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, l$ dan $y_i \in \{1, 2, \ldots, k\}$ merupakan kelas dari x_i, permasalahan tersebut dapat diselesaikan dengan:

$$\min_{w^m, b^i, \xi^i} (w) \frac{1}{2} (w^m)^T w^m + C \sum_{i=1}^{l} \xi_i^m \qquad (3.24)$$

Dengan $(w^m)^T \phi(x_m) + b^m \geq 1 - \xi_i^m \rightarrow y_i = m$ (3.17)

$(w^m)^T \phi(x_m) + b^m \leq -1 + \xi_i^m \rightarrow y_i \neq m$ (3.18)

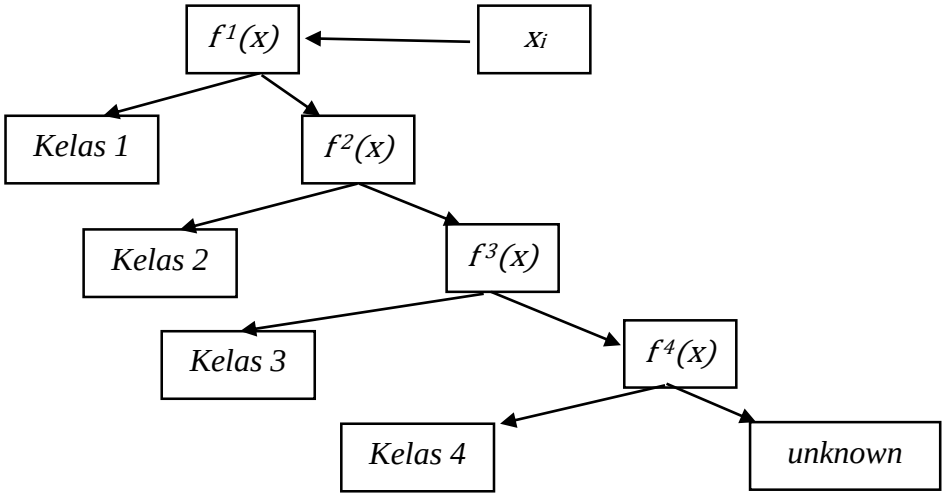
$\xi_i^m \geq 0, \quad i = 1, \ldots, l$

Contoh: terdapat permasalahan klasifikasi dengan 4 buah kelas. Empat buah model SVM biner digunakan untuk pelatihan seperti pada tabel 3.1 dan penggunannya dalam mengklasifikasi data baru dapat dilihat pada gambar 3.4.

Tabel 3.1 Contoh SVM dengan Metode *One-against-all*

$Y_i = 1$	$Y_i = -1$	Hipotesis
Kelas 1	Bukan Kelas 1	$f_1(x) = (w_1)x + b_1$
Kelas 2	Bukan Kelas 2	$f_2(x) = (w_2)x + b_2$
Kelas 3	Bukan Kelas 3	$f_3(x) = (w_3)x + b_3$
Kelas 4	Bukan Kelas 4	$f_4(x) = (w_4)x + b_4$

(3.20)



Gambar 3.4 Contoh Klasifikasi dengan Metode *One-Against-All*

b. One-Against-One

Metode *One-Against-One* membangun (k(k-1)/2) buah model klasifikasi biner (k adalah junlah kelas). Setiap model klasifikasi dilatih pada data dari dua kelas. Data latih dari kelas ke-m dan dan ke-n, diselesaikan dengan:

$$\min_{w^{mn},b^{mn},\xi^{mn}} \frac{1}{2} (w^{mn})^T w^{mn} + C \sum_{i=1}^{l} \xi_i^{mn} \qquad (3.21)$$

Dengan $(w^{mn})^T \phi(x_i) + b^{mn} \geq 1 - \xi_i^{mn} \rightarrow y_i = m$ (3.22)

$(w^{mn})^T \phi(x_i) + b^{mn} \leq -1 - \xi_i^{mn} \rightarrow y_i \neq n$ (3.23)

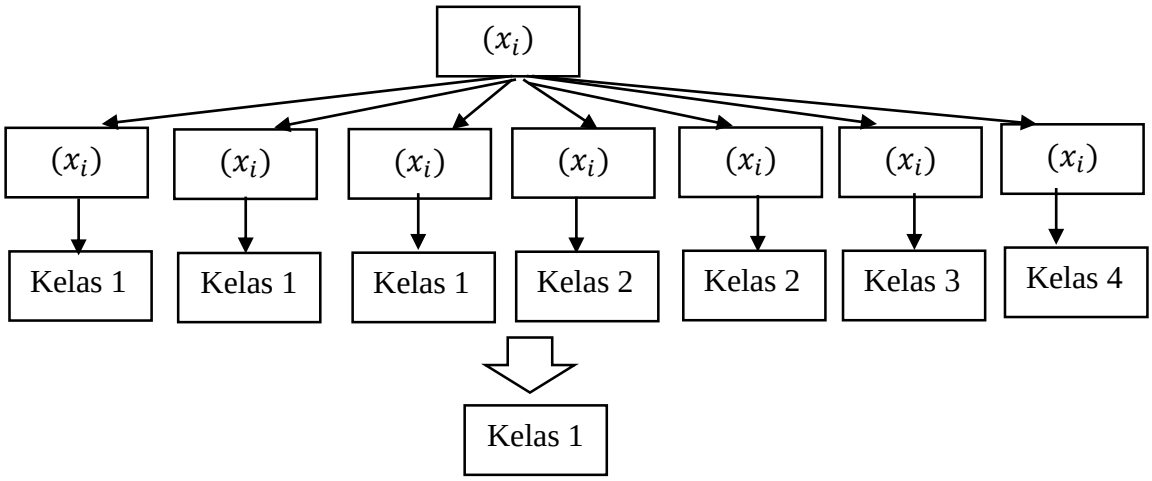
$\xi_i^{mn} \geq 0,$

Contoh: terdapat permasalahan klasifikasi dengan 4 buah kelas. Terdapat beberapa metode untuk melakukan pengujian setelah keseluruhan (k(k-1)/2) model klasifikasi selesai dibangun. Salah satunya adalah metode voting.

Tabel 3.2 Contoh SVM biner dengan Metode *One-against-one*

$Y_i = 1$	$Y_i = -1$	Hipotesis
Kelas 1	Kelas 2	$f_{12}(x) = (w_{12})x + b_{12}$
Kelas 1	Kelas 3	$f_{13}(x) = (w_{13})x + b_{13}$
Kelas 1	Kelas 4	$f_{14}(x) = (w_{14})x + b_1$
Kelas 2	Kelas 3	$f_{23}(x) = (w_{23})x + b_{23}$
Kelas 2	Kelas 4	$f_{24}(x) = (w_{24})x + b_{24}$
Kelas 3	Kelas 4	$f_{34}(x) = (w_{34})x + b_{34}$

(3.24)



Gambar 3.5 *Contoh klasifikasi dengan metode One-against-one*

Misal data x disubtitusikan kedalam fungsi $(w^{mn})^T \phi(x) + b$ kemudaian hasilnya menyatakan x adalah kelas i, maka kelas i terhitung 1 dan ditambahkan lagi ketika ada yang masuk ke kelas tersebut. Metode voting tersebut mengambil jumlah terbanyak dari kelas data x.

3.2. *Confusion Matrix*

Cara yang paling umum untuk menunjukkan hasil klasifikasi terutama pada data *multiclass* ialah menyajikan dalam bentuk *confusion matrix* atau juga dikenal dengan tabel kontingensi. Misal $x_{r,c}$ dari *confusion matrix* $C \in \mathbb{N}^{l \times l}$ dengan r menunjukkan kelas prediksi dan c menunjukkan kelas yang benar, maka diagonal pada tabel tersebut menunjukkan hasil prediksi yang benar dan yang diluar diagonal merupakan hasil prediksi yang salah.

Tabel 3.4 *confusion matrix*

Prediksi (r)	Asli (c)			
	kelas 1	kelas 2	…	kelas n
kelas 1	x_{11}	x_{12}	...	x_{1n}
kelas 2	x_{21}	x_{22}	...	x_{2n}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
kelas n	x_{n1}	x_{n2}	...	x_{nn}

3.3. ROC (*Receiver Operating Characteristic*) *curve*

ROC (*Receiver Operating Characteristic*) *curve* merupakan suatu metode yang dapat digunakan untuk menilai kinerja suatu pengujian (Bolstad,2004). Pengujian tersebut digambarkan dalam suatu kurva dengan axis vertikal merupakan tingkat *true* postif (*sensitivity*), artinya ialah suatu kelas yang diperidiksi masuk ke kelas positif dan hasilnya benar. Label axis horizontal merupakan tingkat *false* positif(1-*sensitivity*), artinya ialah kelas data yang diprediksi masuk ke kelas positif dan hasilnya salah. Cara tersebut akan memberikan hasil 100% jika tidak ada yang yang salah.

3.4. DNA dan *Gene Expression*

Material *genetic* yang berisi petunjuk dari suatu organisme dikenal dengan *deoxyribonucleic acid* (DNA). DNA terdiri dari beberapa nukleutida, masing nukleotida itu sendiri terdiri dari 3 komponen: basa, gula, dan pospat. Nukleutida tersebut tergabung dalam rantai panjang. Inti dari dari rantai tersebut adalah gula dan pospat, sementara basa-nya menggatung pada setiap gula. Terdapat 4 basa berbeda yaitu: *adenine*, *cytosine*, *guanine* dan *thymine* yang biasa dikenal dengan huruf masing-masing A, C, G, dan T. Molekul DNA terdiri dari dua rantai polynucleotide komplementer yang disatukan menggunakan ikatan hydrogen. Basa A dan T terikat bersama, seperti C dan G. Dengan cara ini, dapat dikatakan A adalah komplemen dari T, dan C komplemen dari G. Dua untaian gula dan pospat membenuk struktur *double helix*. Untaian DNA tersebut biasanya berisikan jutaan nukleotida.

RNA, *ribonucleic acid*, berebeda dengan DNA dalam beberapa hal. Khususnya pada gula yang *ribose* bukan *deoxyribose* dan basa-nya *uracil* (U) yang menggantikan *thymine*. U komplementer dengn A. Berbeda dengan DNA, molekul-molekul RNA beruntai tungggal dan hanya berisikan 75-5000 nucleutida. Sel mengandung beberapa jenis RNA: *massanger* RNA (mRNA), *transfer* RNA (tRNA) dan *ribosomal* RNA (rRNA).

Gen meupakan suatu *sequence* DNA yang mengkodekan protein. Protein dapat mengendalikan sifat fisik dari sel, misalnya mata atau rambut. Tali DNA mengandung banyak gen yang berbeda. Protein mengurutkan 20 jenis berbeda dari *amino acid*. Setiap *amino acid* di sandikan dengan suatu urutan dari 3 *base* yang disebut dengan *codons*. Terdapat 64 kemungkinan untuk ketiganya (*codons*), ada suatu kelebihan karena beberapa sandi codons berbeda terdapat kesamaan *amino acid*. Ada 3 codons yang tidak disandi ke *amino acid*. Hal tersebut merupakan penanda berhenti.

Proses sintesis protein dari DNA terjadi dalam 2 tahap, yaitu *transcription* dan *translation* yang dikenal dengan *central dogma of*

molecular biology. Tahap pertama, *transcription* adalah transfer informasi dari molekul DNA beruntai ganda ke mRNA beruntai tunggal. *Transcription* dimulai di daerah *sequence* yang dikenal sebagai *promotor sites* dan berakhir di daerah yang dikenal sebagai *situs terminator*. Tahap kedua *translation* ialah proses penerjemahan mRNA menjadi protein. Pada tahap in digunakan tRNA dan rRNA. Codon AUG menandai lokasi *translation* dapat dimulai. Molekul tRNA melampirkan *amino acid* ke rantai sebagai suatu molekul rRNA yang bergerak sepanjang mRNA. Proses tersebut berlanjut hingga salah satu codon penanda berhenti tercapai.

3.5. *Microarray*

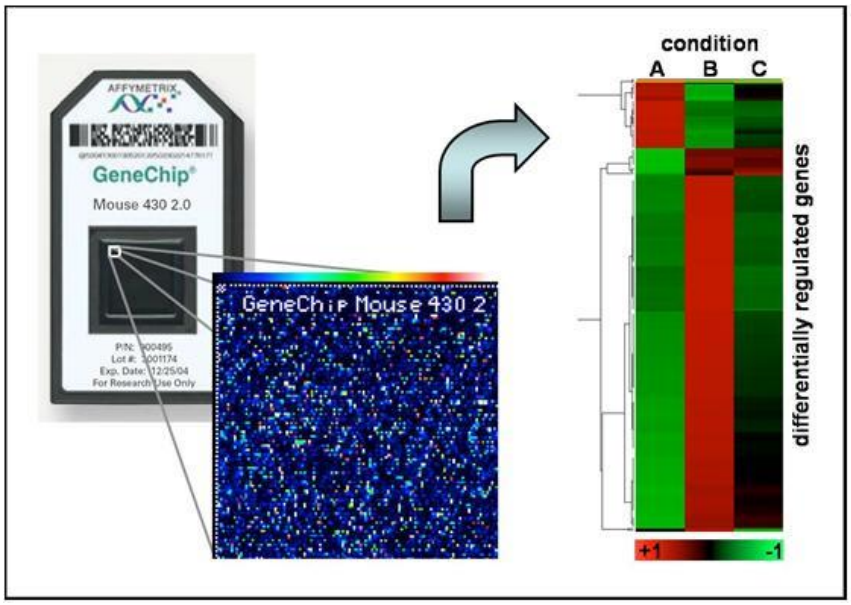
Microarray telah banyak digunakan pada penelitian biomedis. *Microarray* adalah suatu alat yang dirancang untuk mengukur *expression levels* dari sekian ribu gen dalam sebuah penyakit atau tipe sel.(Bolstad,2004). Biologi molekuler dan biomedis menggunakan *microarray* untuk mengetahui dan meneliti perbedaan penyakit. *Microarray* dapat memeriksa ribuan gen dalam waktu yang sama serta membantu mengidetifikasi gen yang terlihat pada sel yang berbeda dan mencari hubungan antara masing-masing gen.

Kebanyakan sel yang ada dalam tubuh manusia berisikan gen yang sama, akan tetapi tidak semua gen yang terpakai di dalam setiap sel. Ada beberapa gen yang aktif atau muncul ketika dibutuhkan saja. Teknologi *microarray* digunakan untuk mengetahui perbedaan gen yang aktif atau tidak pada suatu sel.

Mencocokkan DNA nukleutida ke gen berbeda yang diletakkan pada satu mikroskop yang disebut microarray. mRNA kemudian diekstraksi dari sel yang disintesis ke DNA komplementer dengan *enzyme reverse* transcriptase dan diberikan label dengan tanda yang bercahaya kemudian dihibridisasi, cahaya akan membantu untuk memnunjukkan gen yang aktif pada sel. Gen yang muncul berbeda antara susunan orang sakit dan susunan orang sehat dapat menunjukkan penyebab penyakit.

Microarray DNA Oligonucleotida lebih lanjut dapat menjadi dua subkelompok: long oligonucleotide *arrays*, yang *probe*nya terdiri dari 60-mer atau 50-mer sekuens DNA (contoh Illumina Beadarray), dan short oligonucleotide arrays yang menggunakan 25-mer (mis., Affymetrix GeneChip) atau 30-mer dari desain urutan *probe*.

Affymetrix genechip menggunakan teknologi seperti *chip silicon* komputer. Material *silicon affymetrix* dilindungi dengan menutup dan menerapkan *photolithographic process* untuk mengendalikan sintesis oligonukleotida pada permukaan kaca/plastic. Perancangan *probe*, menggunakan 25-mer gen spesifik olignukleotida, secara lebih khusus, *probe* set dibentuk dengan 11 sampai 20 pasang *probe* berbeda yang digunakan untuk mencocokan gen-gen berbeda. Desain pasangan *probe* yaitu *mismatch* (MM) dan *perfect match*(PM) *probe*. *Probe* MM digunakan untuk mengendalikan ikatan-ikatan non-spesifik selama hibridisasi. Salah satu fitur khusus dari *array GeneChip* adalah bahwa setiap pasangan *probe* terpasang pada lokasi yang telah ditentukan pada permukaan array.



Gambar 3.6 *Affymetrix Genechip*

Keterangan:

Titik merah = gen tersebut diekspresikan hanya dalam kondisi aerobik

Green spot = gen tersebut diekspresikan hanya dalam kondisi anaerobik

Titik kuning = gen tersebut dinyatakan dalam kedua kondisi

Bintik hitam = tidak ada ekspresi gen dalam kondisi baik

3.6. *Preprocessing*

Preprocessing merupakan tahap untuk penyesuaian serta konversi data *affybatch* ke dalam bentuk *expression set*. Tidak hanya melakukan konversi data *affybatch*, dalam tahap *preprocessing* dilakukan proses *background correction*, *normalization* dan *summarization*. Istilah *background correction* mengacu pada penyesuaian berbagai macam metode, serta yang harus dilakukan meliput:

1. Memperbaiki *background noise* dan efek pengolahan
2. Menyesuaikan hibridisasi pengikat DNA non-spesifik pada *array*
3. Menyesuaikan estimasi ekspresi sehingga bersekala tepat atau berhubungan linear

Penting untuk dicatat bahwa definisi tersbut merupakan hal yang umum digunakan. *Background correction* secara umum mengacu pada penegrtian pertama.

Tahap setelah *background correction* yaitu *normalization*. *Normalization* adalah proses penghapusan akhiran non-biologis yang tidak diinginkan diantara chip dalam eksperimen *microarray*. Tahap setelah hal tersebut ialah tahap dilakukan *summarization*, yaitu untuk mengukur *gene expression*. (Bolstad, 2004)

3.7. *Filtering*

Filtering data *microarray* adalah proses pemilihan subset dari *probe* yang tersedia untuk pengecualian atau penyertaan dalam analisis. Program R dalam penelitian ini menggunakan tambahan *package* genefilter untuk melakukan penyaringan. Fungsi pertama yang digunakan pada tahap *filtering*

pada penelitian ini ialah `nsFilter`, *function* tersebut menyediakan suatu opsi *one-stop shop* (serba ada) untuk berbagai pilihan *filtering* (penghapusan) fitur dari *expression set*. Fitur *filtering* dapat menunjukkan beberapa varian kecil atau ketimpangan data secara konsisten di seluruh sampel, hal ini dapat berguna untuk analisis selanjutnya (Bourgon,2010).

BAB IV
METODE PENELITIAN

4.1.Jenis dan Sumber Data

Jenis data dalam penelitlian ini berupa data sekunder yang berasal dari penelitian Finkelstein dengan judul "*Novel mutations target distinct subgroups of medulloblastoma*" tersimpan dalam *database chip microarray* yang terdapat di instansi yang bergerak dibidang genetika atau *database* kumpulan data genetik yaitu NCBI dengan kode akses GSE37418.

4.2.Tempat dan Waktu Penelitian

Penelitian dilakukan di kampus terpadu Universitas Islam Indonesia yang ada di Yogyakarta. Waktu pengambilan data yang dilakukan penulis adalah pada bulan Agustus 2017.

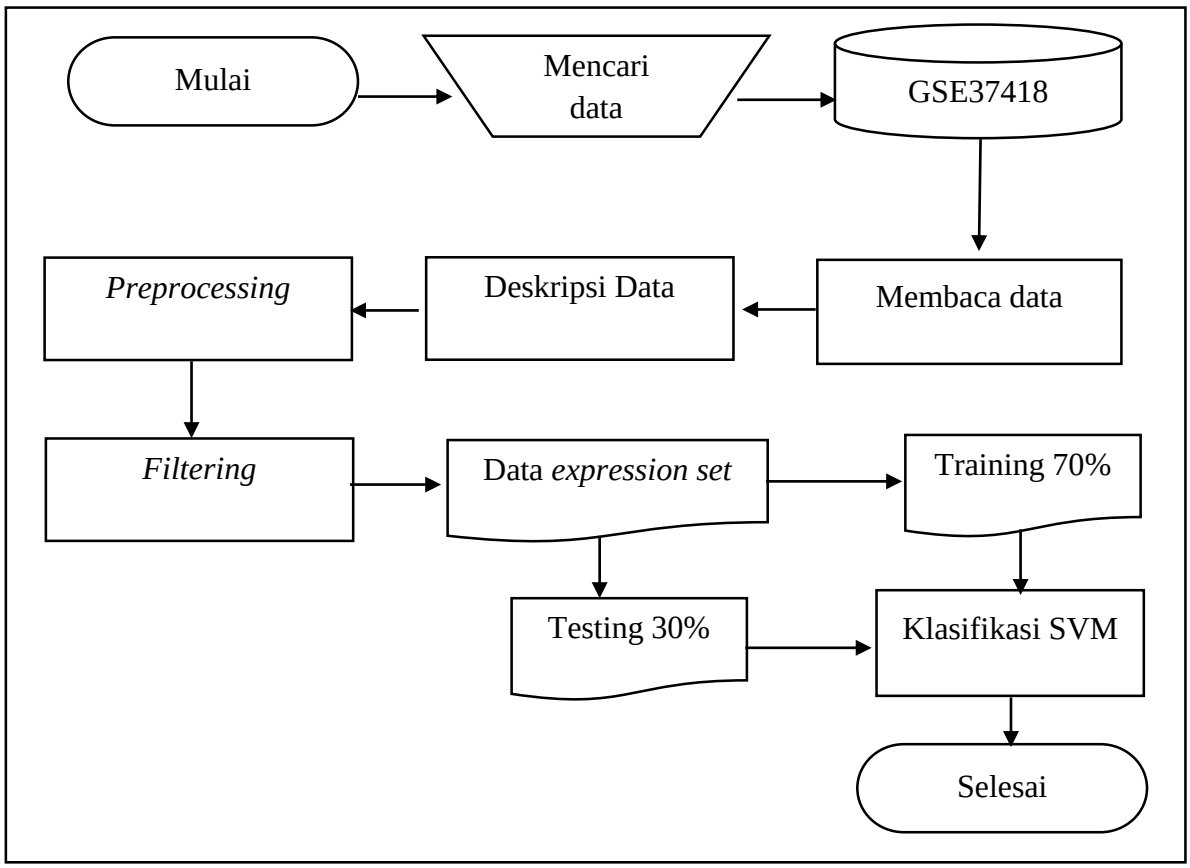
4.3.Variabel Penelitian

Variabel penelitian yang digunakan peneliti ialah sebanyak 54675 yang merupakan kumpulan *probe* dari data *expression set* dari GSE37418 serta 3 variabel tambahan dari *phenotype* data yakni etnis, jenis kelamin dan *subgroup* kanker *medulloblastoma* yang menjadi faktor dalam melakukan klasifikasi.

4.4.Metode Analisis Data

Metode yang digunakan dalam penelitian ialah klasifikasi dengan menggunakan algoritma SVM. *Software* yang digunakan ialah R studio atau Program R dengan R version 3.4.2. Terdapat beberapa metode yang dilakukan untuk implementasi metode statistik SVM dalam data bioinformatika mulai dari pembacaan data oleh program, *Preprocessing*, *filtering* yang kemudian merubah data kedalam bentuk matriks atau tabel.

Langkah atau tahapan yang dilakukan dalam penelitian ini dapat digambarkan dengan gamabr berikut:



Gambar 4.1 *Flowchart* Penelitian

BAB V
PEMBAHASAN

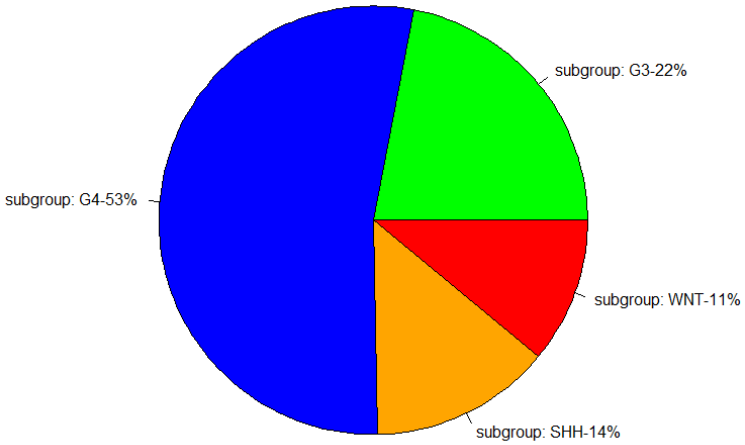
Medulloblastoma terbentuk karena terjadi kesalahan pada fungsi sel yang mengendalikan pertumbuhan dan kematian sel. Alasan hal tersebut terjadi masih belum dipahami, bagaimanapun, para ilmuwan membuat kemajuan signifikan dalam memahami apa yang terjadi di dalam sel-sel yang mengubah sel otak normal menjadi kanker yang tumbuh. Perubahan tersebut teridentifikasi pada gen dan kromosom (sel DNA) yang berperan dalam perkembangan kanker *medulloblastoma*.

Data GSE37418 merupakan kumpulan sampel *gene expression* dari penderita *pediatric medulloblastoma*. 76 data pasien penderita *medulloblastoma* pada sampel memiliki karakteristik yang berbeda-beda seperti etnis, jenis kelamin, dan yang lainnya.

5.1.Deskripsi Data

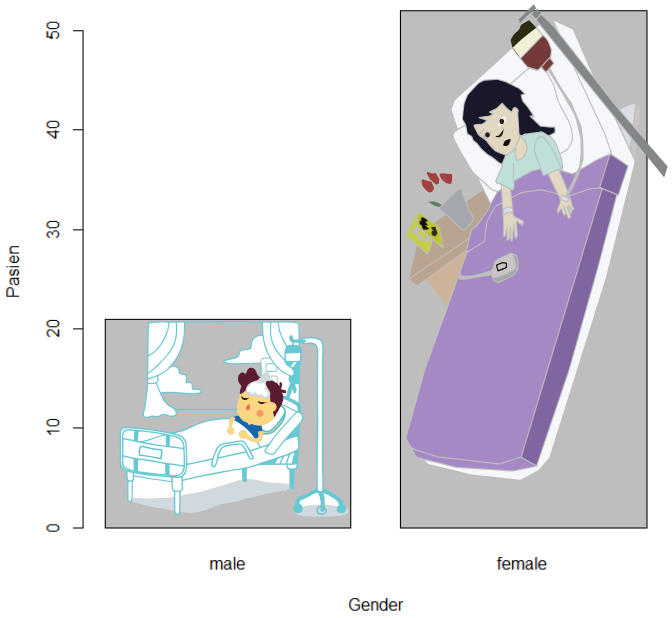
Gambaran secara umum dari penderita *medulloblastoma* pada penelitian, akan disajikan dengan menggunakan analisis dekriptif. Data GSE37418 terdapat 73 pasien dengan 4 subtipe yang sama dengan disebutkan Upadhyay dan 3 pasien lain tidak memiliki subtype (*outlier*), oleh karena itu data yang digunakan ialah pasien yang tergolong ke dalam 4 subtipe yaitu WNT, SHH, *Subgroup* 3, dan *Subgroup* 4.

Gambar 5.1 menunjukkan jika kebanyakan pasien penderita tergolong pada *subgroup* 4, setengah (53%) dari pasien tergolong ke *subtype subgroup* 4, 22% pasien tergolong ke *subgroup* 3, pasien dengan golongan SHH 14% dan 11% pasien tergolong WNT. Perlakuan untuk setiap golongan berbeda-beda dalam proses penyembuhan .

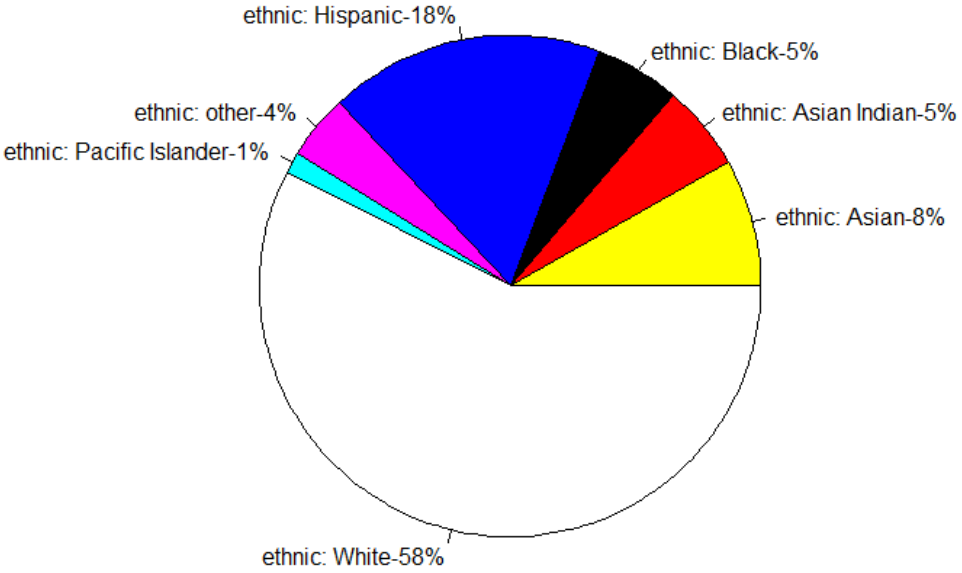


Gambar 5.1 *Pie Chart* Subtipe *Medulloblastoma*

Pasien berjenis kelamin perempuan lebih banyak ditemukan pada sampel yang ditunjukkan pada gambar 5.2 dibandingkan dengan pasien berjenis kelamin laki-laki. Kebanyakan pasien berumur 9 tahun, hal tersebut seusai dengan artikel yang dituliskan ABTA, penderita *medulloblastoma* lebih cendrung pada anak dibawah 10 tahun.



Gambar 5.2. Jenis Kelamin



Gambar 5.3. Etnis Pasien

Gambar 5.3 diatas menunjukkan jumlah masing-masing etnis pasien. Etnis kulit putih paling banyak yaitu 58% dalam sampel, hal tersebut dikarenakan pengambilan sampel dilakukan di Amerika yang kebanyakan penduduknya merupakan kulit putih. 18% penderita berasal dari etnis hispanik yang merupakan etnis terbesar di Amerika, kemudian Asia 5% pasien. 5% pasien masing masing dari etnis Indian dan Black, 4% pasien other dan 1% pasien dari etnis pasiific islander.

5.2. Pengolahan Data *Bioinformatics*

Data *bioinformatics* dalam pengolahannya menggunakan R memerlukan *package* tamabahan dari *biocondutor*. Data yang dapat diolah oleh R bersifat vektor, matrix, data *frame*, atau yang lainnya, akan tetapi R memerlukan tambahan *package* untuk membaca data *affymetrix* atau *expression set*. Proses mendapatkan data bioinformatics pada R diperlukan *package* `GEOquery` dan untuk membaca data tersebut dipelukan *package* `affy`.

Data yang didapat dengan *package* `GEOquery` masih dalam bentuk *affymetrix*, data tersebut dijabarkan agar terdeteksi dalam program R dengan menggunakan function `ReadAffy()`. Data *affymetrix* menyimpan banyak informasi yaitu *expression set* dari pasien dan *phenotype*-nya. Hal tersebut didapat dengan query seperti pada lampiran untuk mendapatkan hal seperti berikut:

Tabel 5.1 *pheno data*

title	...	channel_c	source_na	organism_ch1	characteristics_ch1	characteristics_ch1.1	characteristics_ch1.2	characteristics_ch1.3	characteristics_ch1.4	...	treatment_protocol_ch1
GSM918578	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 8yrs 4mos	Sex: Male	ethnic: White	...	m stage: MB-CL
GSM918579	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 8yrs 3mos	Sex: Male	ethnic: Black	...	m stage: MB-CL
GSM918580	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: WNT	age: 6yrs 6mos	Sex: Female	ethnic: White	...	m stage: MB-CL
GSM918581	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 9yrs 2mos	Sex: Male	ethnic: White	...	m stage: MB-CL
GSM918582	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: SHH	age: 8yrs 6mos	Sex: Male	ethnic: Asian	...	m stage: MB-DN
GSM918583	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 8yrs 1mos	Sex: Female	ethnic: White	...	m stage: MB-CL
GSM918584	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 8yrs 8mos	Sex: Male	ethnic: White	...	m stage: MB-CL
GSM918585	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 9yrs 0mos	Sex: Female	ethnic: White	...	m stage: MB-CL
GSM918586	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G3	age: 10yrs 10mos	Sex: Male	ethnic: White	...	m stage: MB-AN
GSM918587	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G3	age: 9yrs 8mos	Sex: Male	ethnic: other	...	m stage: MB-CL
GSM918588	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G3	age: 4yrs 11mos	Sex: Male	ethnic: White	...	m stage: MB-CL
GSM918589	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G3	age: 5yrs 2mos	Sex: Male	ethnic: Hispanic	...	m stage: MB-CL
GSM918590	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 8yrs 9mos	Sex: Male	ethnic: Hispanic	...	m stage: MB-AN
GSM918591	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 9yrs 4mos	Sex: Male	ethnic: White	...	m stage: MB-CL
GSM918592	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 14yrs 8mos	Sex: Male	ethnic: Hispanic	...	m stage: MB-CL
GSM918593	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: WNT	age: 9yrs 1mos	Sex: Female	ethnic: White	...	m stage: MB-CL
GSM918594	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G3	age: 4yrs 8mos	Sex: Male	ethnic: White	...	m stage: MB-AN
GSM918595	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 7yrs 9mos	Sex: Female	ethnic: Asian Indian	...	m stage: MB-CL
GSM918596	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 3yrs 4mos	Sex: Female	ethnic: White	...	m stage: MB-CL
GSM918597	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 5yrs 5mos	Sex: Male	ethnic: White	...	m stage: MB-CL
GSM918598	...	RNA	1	Medulloblastoma	Homo sapiens	subgroup: G4	age: 11yrs 9mos	Sex: Male	ethnic: White	...	m stage: MB-CL

Dari tabel diatas dapat dilihat ID dari pasien diberikan dengan kode GSM yang mencakup *subtype medulloblastoma* yang diderita, status publikasi data, organisme data, hingga lokasi data diambil. Keseluruhan isi pada tabel 5.1 terdapat 73 pasien dengan 37 variabel. Tahapan di atas program R masih terbatas dalam mengakses data karena data masih berbentuk *Affybatch*.

5.2.1. *Preprocessing*

Tahap ini memerlukan *function* `threestep()` dari *package* `affyPLM`. Function threestep() selain melakukan konversi data *affybatch* dalam function tersebut dilakukan proses *background correction*, *normalization* dan *summarization*.

Metode *background correction* dalam penelitian ini menggunakan RMA.2. RMA merupakan singkatan *dari Robust multi-*

array average, metode *background correction* RMA bertujuan untuk mengetahui distribusi intesitas *probe*. Tahapan setelah *background correction* ialah melakukan *normalization* dengan metode `quantile` unutuk menyetarakan data. Tahap setelah dilakkukan *background correction* dan *normalizatition* dilanjutkan ke tahap *summarization* dengan metode `median.polish`.

5.2.2. *Filtering*

Filtering data *microarray* adalah proses pemilihan subset dari *probe* yang tersedia untuk pengecualian atau penyertaan dalam analisis. Program R menggunakan tambahan *package* `genefilter` untuk melakukan penyaringan. Fungsi pertama yang digunakan pada tahap *filtering* pada penelitian ini ialah `nsFilter`, function ini menyediakan suatu opsi *one-stop shop* (serba ada) untuk berbagai pilihan *filtering* (penghapusan) fitur dari *expression set*. Fitur *filtering* dapat menunjukkan beberapa varian kecil atau ketimpangan data secara konsisten di seluruh sampel, hal ini dapat berguna untuk analisis selanjutnya (Bourgon,2010).

Fungsi `nsFilter` merupakan salah stu fungsi yang digunakan dalam penelitian, Fungsi tersebut menunjukkan jika dalam *filtering* digunakan *entrez* (database yang menyimpan terkait gen). Fungsi pada lampiran.7 berguna untuk menghapus `dupEntrez` yang artinya mengeluarkan variable dengan nilai IQR(*interquartile range*) tertinggi. Data yang didapat setelah prooses tersebut berdimensi 10251x 73, untuk mereduksi dimensi data dilakukan metode *filtering* kembali dengan menggunakan function `genefilter`.

Klasifikasi data pada penelitian ini diasumsikan berdistribusi normal yang melakukan *filtering* dengan fungsi Anova dengan nilai signifikansi 0.001. Dimensi data yang didapat setelah proses *filtering* menggunakan `genefilter` menjadi 5226x73.

5.2.3. Mengolah data

Tahap *Preprocessing* dan *filtering* yang telah dijabarkan menghasilkan suatu data *expressionset*. Tahap selanjutnya yaitu merubah data *expression set* menjadi matrix atau *data frame* agar dapat diolah dengan menggunakan *package* yang umum dari R. fungsi yang digunakan dalam peneiltian ini yaitu `exprs` dari *package biobase* dari Bioconductor.

Function `exprs` berfungsi untuk merubah data *expression set* menjadi bentuk matrix. Data yang berebentuk matrix atau data frame merupakan syarat untuk mengolah data dengan *package* `e1071`, dengan menggunakan `e1071` akan dilakukan klasifikasi SVM.

5.3. Klasifikasi Data Gen

Klasifikasi data pada umumnya menggunakan 2 data yakni data latih dan data uji. Suatu kumpulan data dibagi menjadi 2 pada kasus ini dengan porsi data latih 70% dan data uji 30%. Masing-masing data dari data latih dan data uji diambil secara random dengan mengutamakan semua klasifikasi termasuk kedalam 2 data tersebut. Diketaui data *medulloblastoma* dari GSE37418 terdapat 4 kelas yakni SHH,WNT, *subgroup* 3 dan *subgroup* 4 seperti gambar 5.1, data GSE37418 terdiri dari 73 sample yang digunakan, maka 58(80%) sampel yang diambil harus terdapat 4 kelas dan 15(20%) sampel data uji juga harus terdiri dari 4 kelas dari klasifikasi medulloblastoma.

Pengaplikasian SVM dengan menggunakan *package* e1071 dapat dilakukan *tune* untuk mengetahui nilai *cost* terbaik untuk model. Semakin kecil nilai *cost* maka akan semakin baik hasilnya. *Cost* yang digunakan benilai antara 0.1, 0.01, 0.001, 1, 10, dan 100. Berdasarkan nilai-nilai tersebut didapat *cost* terbaik untuk data ialah 0.1(lampiran 13), oleh karena itu untuk membangun model digunakan *cost* 0.1.

Model yang dihasilkan dengan menggunakan data latih selanjutnya dilihat hasil dari klasifikasi terhadap data tersebut dan tidak terdapat

kesalahan seperti tabel5.2 berikut yang menunjukkan hasil klasikasi data latih.

5.3.1. *Confusion matrix*

Efektifitas suatu klasifikasi diuji dengan suatu pengukuran evaluasi. Pengukuran tersebut didapatkan dalam *confusion matrix* (Chin, 2010).

Tabel 5.2 *Confusion Matrix* Data Latih

Predictions	Data latih			
	subgroup: G3	subgroup: G4	subgroup: SHH	subgroup: WNT
subgroup: G3	11	0	0	0
subgroup: G4	0	27	0	0
subgroup: SHH	0	0	7	0
subgroup: WNT	0	0	0	5

Setiap prediksi berdasarkan model pada data latih menghasilkaan klasifikasi yang sesuai dengan kelasnya, setiap *subgroup* tidak masuk ke klasifikasi yang salah. Hasil tersebut didapatkan karena data yang digunakan untuk membangun model SVM sama dengan data yang diujikan. Berdasarkan hasil tersebut data uji digunakan untuk mengetahui model yang dibangun mampu memprediksi data yang baru. Data uji diasumsikan sebagi data yang baru dan data uji tidak digunakan dalam membagun model SVM. Prediksi yang didapat dengan data uji ditunjukkan oleh tabel 5.3.

Tabel 5.3 *Confusion matrix* Data Uji

Predictions	Data uji			
	subgroup: G3	subgroup: G4	subgroup: SHH	subgroup: WNT
subgroup: G3	4	0	0	0
subgroup: G4	1	12	0	0
subgroup: SHH	0	0	3	0
subgroup: WNT	0	0	0	3

Data uji menunjukkan jika model yang dibangun memiliki akurasi yang bagus yakni 95,6%. *Confusion matrix* data uji dari 15 sampel terdapat 1 kesalahan yaitu pasien yang diprediksi klasifikasi *subgroup* G4 pada kenyataanya pasien tersebut merupakan klasifikasi *subgroup* G3. Hasil dari model berdasarakan *confusion matrix* memiliki hasil yang bagus, akan tetapi untuk menguji validasi lebih lanjut digunakan ROC.

5.3.2. *ROC curve*

Akurasi dari model sebelumnya menggunakan data uji mendapatkan nilai akurasi 95% dan hasil ini akan ditinjau lagi dengan melihat *area under ROC curve* (AUC). Metode AUC pada awalnya digunakan untuk memprediksi sinyal dan digunakan untuk membuat keputusan dalam bidang medis.

Nilai *multi-class* AUC ialah 0.98, nilai tersebut dapat dikatakan tinggi karena nilai maksimumnya 1. Nilai akurasi akurasi model didapat dari *confusion matrix* 0.95(95%) dan nilai *multi-class* AUC 0.98. Nilai AUC yang tinggi menunjukkan jika model yang diperoleh baik karena nilai AUC menunjukkan akurasi prediksi dari model.

5.4. Model Klasifikasi

Hasil klasifikasi pasien *medulloblastoma* pada data GSE3418 memberikan bentuk dengan nilai AUC %. Pada data tersebut dikethui setelah melakukan *filtering* didapatkan 5226 *probe* yang mejadi variabel pada penelitian dana ditambahkan 3 variabel karakteristik dari sampel yakni jenis kelamin, etnis, dan kelas pasien. Berdasarkan model, berikut bobot 10 variabel dari 5229 variabel yang ditunjukkan ($\vec{w}$).

Tabel 5.5 Bobot *Probe ID*

probe id	w
201049_s_at	0.009322832
200763_s_at	0.009145843
200819_s_at	0.009108242
201665__at	0.009061992
200062_s_at	0.009028423
200013_at	0.008993589
215963__at	0.008993333
208645_s_at	0.008926398
208692_at	0.008914243
200010_at	0.008872732

Tabel 5.5 menunjukkan bobot dari *probe*, bobot tertingi ialah pada *probe id* 201049_s_at, *probe id* dapat menununjukkan gen name dan ontology dari gen tersebut. Pada kasus ini diambil 10 *probe* dengan nilai bobot yang tinggi, setiap *probe* tersusun dari bobot yang tinggi ke bobot yang rendah, dari 10 *probe* tersebut didapatkan informasi seperti pada tabel 5.5.

Tabel 5.5 menunjukkan nama gen dari *probe* tersebut berasal serta *gen ontology*-nya. *Gene ontology* merupakan kerangka untuk model biologi. *Gene Ontology* didefinisikan sebagai konsep atau kelas yang digunakan untuk menggambarkan fungsi gen serta hubungan antar konsepnya. Terdapat 3 fungsi yaitu BP(*biological process*), MF(*molecular function*), dan CC (*Celullar Component*).

Tabel 5.5 Tabel *Ontology*

*PROBE*ID	SYMBOL	GENENAME	ENTREZID	ONTOLOGY
201049_s_at	RPS18	ribosomal protein S18	6222	BP
201049_s_at	RPS18	ribosomal protein S18	6222	CC
201049_s_at	RPS18	ribosomal protein S18	6222	MF
200763_s_at	RPLP1	ribosomal protein lateral stalk subunit P1	6176	BP
200763_s_at	RPLP1	ribosomal protein lateral stalk subunit P1	6176	CC
200763_s_at	RPLP1	ribosomal protein lateral stalk subunit P1	6176	MF

*PROBE*ID	SYMBOL	GENENAME	ENTREZID	ONTOLOGY
200819_s_at	RPS15	ribosomal protein S15	6209	BP
200819_s_at	RPS15	ribosomal protein S15	6209	CC
200819_s_at	RPS15	ribosomal protein S15	6209	MF
201665_x_at	RPS17	ribosomal protein S17	6218	BP
201665_x_at	RPS17	ribosomal protein S17	6218	CC
201665_x_at	RPS17	ribosomal protein S17	6218	MF
200062_s_at	RPL30	ribosomal protein L30	6156	BP
200062_s_at	RPL30	ribosomal protein L30	6156	CC
200062_s_at	RPL30	ribosomal protein L30	6156	MF
200013_at	RPL24	ribosomal protein L24	6152	BP
200013_at	RPL24	ribosomal protein L24	6152	CC
200013_at	RPL24	ribosomal protein L24	6152	MF
215963_x_at	RPL3	ribosomal protein L3	6122	BP
215963_x_at	RPL3	ribosomal protein L3	6122	CC
215963_x_at	RPL3	ribosomal protein L3	6122	MF
208645_s_at	RPS14	ribosomal protein S14	6208	BP
208645_s_at	RPS14	ribosomal protein S14	6208	CC
208645_s_at	RPS14	ribosomal protein S14	6208	MF
208692_at	RPS3	ribosomal protein S3	6188	BP
208692_at	RPS3	ribosomal protein S3	6188	CC
208692_at	RPS3	ribosomal protein S3	6188	MF
200010_at	RPL11	ribosomal protein L11	6135	BP
200010_at	RPL11	ribosomal protein L11	6135	CC
200010_at	RPL11	ribosomal protein L11	6135	MF

Tabel diatas menunjukkan jika *probe* berasal dari gen ribosomal protein serta termasuk kedalam ketiga fungsi yang ada dalam fungsi *gene ontology*. Merujuk pada situs NCBI, setiap *probe* memiliki pengaruh yang berbeda. Keterengan terkait peran dan fungsi setiap *probe* disajikan pada tabel 5.6. Tabel 5.6 memnunjukkan 10 *probe* yang memiliki bobot tertinggi merupakan protein yang terletak pada sitoplasma. *Probe* 200819_s_at pernah ditemukan aktif pada berbagai tumor seperti insulinomas, kanker kerongkongan, dan kanker usus besar, *probe* tersebut merupakan salah 1 probe yang memiliki bobot tinggi pada penyakit *medulloblastoma*.

Tabel 5.6 Keterangan *Probe ID*

No	*Probe* ID	Simbol	Keterangan
1	201049_s_at	RPS18	Gen ini mengkodekan protein ribosom yang merupakan komponen subunit 40S. Protein tersebut termasuk dalam famili protein ribosom S13P. Terletak di sitoplasma. Produk gen E. coli ortholog (protein ribosom S13) terlibat dalam pengikatan fMet-tRNA, dan inisiasi penerjemahan.
2	200763_s_at	RPLP1	Gen ini mengkodekan fosfoprotein ribosom yang merupakan komponen subunit 60S. Protein yang merupakan ekuivalen fungsional protein ribosom E. coli L7 / L12 termasuk dalam protein protein ribosom L12P. hal ini memainkan peran penting dalam tahap perpanjangan sintesis protein.
3	200819_s_at	RPS15	Gen ini mengkodekan protein ribosom yang merupakan komponen subunit 40S. Protein tersebut termasuk dalam famili protein ribosom S19P. Terletak di sitoplasma. Gen ini telah ditemukan aktif di berbagai tumor seperti insulinomas, kanker kerongkongan, dan kanker usus besar
4	201665_x_at	RPS17	Gen ini mengkodekan protein ribosom yang merupakan komponen subunit 40S. Protein tersebut termasuk dalam famili S17E protein ribosom dan terletak di sitoplasma. Mutasi pada gen ini menyebabkan anemia Diamond-Blackfan 4.

No	*Probe* ID	Simbol	Keterangan
5	200062_s_at	RPL30	Gen ini mengkodekan protein ribosom yang merupakan komponen subunit 60S. Protein tersebut termasuk dalam famili protein ribosom L30E. Terletak di sitoplasma. Gen ini ditranskripsikan dengan gen RNA nukleolar U72 kecil, yang terletak pada intron keempatnya.
6	200013_at	RPL24	Protein tersebut termasuk dalam famili protein ribosom L24E. Terletak di sitoplasma. Gen ini telah disebut sebagai protein ribosom L30 karena protein yang dikodekan memiliki identitas asam amino dengan protein ribosom L30 dari S. cerevisiae; Namun, nama resminya adalah protein ribosom L24.
7	215963_x_at	RPL3	Protein tersebut termasuk dalam famili protein ribosom L3P dan terletak di sitoplasma. Gen ini ditranskripsikan dengan beberapa gen RNA nukleolar kecil, yang terletak di beberapa intron gen ini.
8	208645_s_at	RPS14	Pada sel ovarium hamster Cina, mutasi pada gen ini dapat menyebabkan resistensi terhadap *emetine* suatu inhibitor sintesis protein.

No	*Probe* ID	Simbol	Keterangan
9	208692_at	RPS3	Studi tentang protein in pada tikus telah menunjukkan bahwa protein ini memiliki peran ekstrrabosom sebagai endonuklease yang terlibat dalam perbaikan kerusakan DNA akibat sinar UV. Protein ini tampaknya terletak di sitoplasma dan nukleus tetapi tidak di dalam nukleolus.
10	200010_at	RPL11	Protein ini termasuk dalam famili protein ribosom L5P. Terletak di sitoplasma. Protein ini berasosiasi dengan rRNA 5S atau varian transkrip yang disambung yang mengkodekan isoform yang berbeda telah dapat ditemukan untuk gen ini.

BAB VI
KESIMPULAN DAN SARAN

6.1. Kesimpulan

Berdasarkan hasil penelitian didapatkan kesimpulan:

1. Data bioinformatika merupakan data biologi yang disajikan dalam bentuk *gene expression* yang disimpan dalam *microarray*. Data tersebut diolah dengan tahapan pembacaan data yang kemudian dilakukan *preprocessing*. Tahap *preprocessing* memliki 3 proses yaitu *background correction* dengan metode RMA, *normalization* dengan metode quantile, *summarization* dengan metode median polish. Hasil yang didapatkan setelah tahap tersebut ialah data *expression set*. Berdasarakan data tersebut dilakukan proses *filtering* untuk menyaring data. Tahap filtering dilakukan 2 kali yaitu dengan `nsFilter` yang mengeluarkan variabel dengan nilai IQR(*interquartile range*) dan `genefilter` yang menyeleksi probe berdasarkan ANOVA dengan nilai signifikansi 0,001. Tahap setelah itu adalah merubah data kedalam bentuk matriks yang kemudian dilakukan anlisis SVM.
2. Hasil klasifikasi pada data uji menunjunukkan ketepatan prediksi 95% serta didukung dengan nilai AUC *multiclass* 98%. Nilai akurasi yang tinggi menunjukkan model SVM yang terbentuk dapat memrediksi dengan baik.
3. Klasifikasi SVM prinsipnya dibuat untuk mengklasifikasikan data berdasarkan garis hyperplane. Garis tersebut memisahkan kelas data, dan memiliki bobot (w) yang memberikan jarak antara kelas satu dengan kelas lainnya. Data GSE37418 memiliki variabel yang banyak, dalam pembentukan model digunakan 10 variabel saja. 10 variabel tersebut memiliki nilai bobot (w) tertinggi (Tabel 5.4). Tabel 5.6 ditunjukkan bahwa setiap probe set memiliki peran dan *probe* set 200819 s_at merupakan gen yang aktif pada berbagai tumor seperti insulinomas, kanker kerongkongan, dan kanker usus besar, serta pada *probe* 201665 x_at dapat menyebabkan anemia Diamond-Blackfan 4 jika terjadi mutasi. Gen-gen yang memiliki bobot tinggi pada *medulloblastoma* merupakan gen yang juga aktif pada penyakit lain.

6.2. Saran

1. Hasil klasifikasi SVM telah menunjukkan akurasi yang tinggi pada penelitian ini. Model SVM yang terbentuk memberikan bobot pada setiap varibelnya (*probe*), terdapat nilai bobot yang tinggi pada *probe* tabel 5.6 yang dapat menjadi target dalam melakukan treatmen pada pasien.
2. Guna mendapatkan hasil yang lebih baik pada penelitian selanjutnya, perlu dialkukan uji normalitas pada sampel ketika melakukan filtering.

DAFTAR PUSTAKA

ABTA. 2015. *Medulloblastoma*. http://www.abta.org/secure/ medulloblastoma-brochure.pdf diakses pada 10 Oktober 2017.

Alshamlan, Hala M., Badr, Ghada H. dan Alohali, Yousef. 2013. *A Study of Cancer Microarray Gene Expression Profile:Objectives and Approaches*. London, U.K. Proceeding of the World Congress on Engineering 2013 Vol II. ISBN: 978-988-19252-8-2.

Anonim, "Big Data". https://www.ibm.com/big-data/us/en/ diakses pada 14 Februari 2017

Anonim, "GenBank", https://www.ncbi.nlm.nih.gov/genbank/statistics/ diakses pada 14 Februari 2017

Anonim, "*Microarray*", diakses dari https://www.ncbi.nlm.nih.gov/*probe*/docs/techmicroarray/ pada 14 Februari 2017

Aprijani, Dwi Astuti dan Elfaizi , M. Abdushshomad, 2004, "Bioinformatika: Perkembangan, Disiplin Ilmu dan Penerapannya di Indonesia", diakses dari ftp://202.125.94.81/pub/linux/docs/v06/Kuliah/SistemOperasi/2003/50/Bioinformatika.pdf pada 14 februari 2017.

Attwood, T.K. dan Parry-Smith, D.J. (1999), Introduction to Bioinformatics, Harlow: Pearson Education.

Bolstad M, Benjamin. 2004. "*Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*". https://pdfs.semanticscholar.org/a0e2/34479d90b24f59791b3d52bbf2cb27d90acf.pdf diakses pada 3 November 2017.

Brown , Michael P. S. William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S. Furey, Manuel Ares, Jr., dan David Haussler. 1999. "*Knowledge-based analysis of microarray gene expression data by using support vector machines*". Proc Natl Acad Sci U S A. **97(1)**:262-7.

Buhler, Lukas K & Rashidi, Hooman H., 2005, "*Bioinformatics Basic: Aplications in biological science and medicine*" , Boca Raton, FL. CRC Press.

Carrillo ,Henry . Brodersen , Kay H.dan Castellanos, Jose A. 2014. "*Probabilistic performance evaluation for multiclass classification using the posterior balanced accuracy*. https://kaybrodersen.github.io/publications/Carrillo_ 2014_ROBOT2013.pdf diakses pada14 desember 2017.

Furey, Terrence S., Duffy, Nigel., Cristianini Nello., Bednarski, David., Schummer, Michel. dan Haussler, David. *Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data*. Bioinformatics. 2000, **16(10)**:906-914.

Hsu, Chih-Wei, Chih-Chung Chang. dan Chih-Jen Lin.. 2010. *A Practical Guide to Support vector Classification*. Taiwan: Department of Computer Science National Taiwan University.

Hsu, Chih-Wei, Chih-Jen Lin. *A Comparison of Methods for Multi-class Support Vector Machines*. IEEE Transactions on Neural Networks, **13(2)**:415-425.2002.

Kilpelainen, Sini. 2008. *Microarray Data Analysis of Dyslexia Candidate Genes*. http://www2.math.su.se/matstat/reports/seriec/2008/rep1/report.pdf diakses pada 3 November 2017.

Lesk, Arthur M. , "*Bioinfomatics*" https://www.britannica.com/science/ bioinformatics diakses pada 3 Juni 2017.

Luo,Yiqiang.2007. *Comparison Between Affymetrix And Illumina Gene Expression Microarray Platforms*. Thesis. McMaster University.

Milgram , Johnathan . 2006. "*One Against One" or "One Against All": Which One is Better for Handwriting Recognition with SVMs?*. https://hal.archives-ouvertes.fr/inria-00103955/document diakses pada 3 November 2017

Mohammed,Mohssen, Khan, Muhammad Badruddin , dan Bashier, Eihab Bashier Mohammed. 2017. *Machine Learning: Algorithms and Applications*. CRC press. New York.

Northcot. Paul .A, Andrey Korshunov, Hendrik Witt, Thomas Hielscher, Charles G. Eberhart, Stephen Mack, Eric Bouffet, Steven C. Clifford, Cynthia E. Hawkins, Pim French, James T. Rutka, Stefan Pfister, dan Michael D. Taylor. 2011. *Medulloblastoma comprises four distinct molecular variants* 2011. JOURNAL OF CLINICAL ONCOLOGY. **10(11)**: 1400, 1415, dan 1424

Nugroho, Anto Satriyo, Arief Budi Witarto dan Dwi Handoko. 2003. *Suport Vector Machines : Teori Aplikasinya dalam Bioinformatika*. ilmukomputer.com.

Rossi A., Caracciolo V., Russo G, dan Reiss K, Giordano A. 2008. *Medulloblastoma: from molecular pathology to therapy*. Clin Cancer Res. **14(4)**:971–976

Trimarsanto, Hidayat. Bioinformatika. Diakses dari http://www.bioinformatika.org /Beranda pada 14 Februari 2017.

Upadhay, Amit Kumar. 2014. *Pediatric Medulloblastoma: Molecular biology, correlation with histopathological and clinical outcome*. http:// dspace.sctimst.ac.in/jspui/bitstream/123456789/2637/1/6307.pdf diakses pada 24 Oktober 2017.

Sembiring, Krisantus. 2007. *Tutorial SVM Bahasa Indonesia*. Bandung : ITB http://sutikno.blog.undip.ac.id/files/2011/11/tutorial-svm-bahasa-indonesia-oleh-krisantus.pdf diakses pada 20 Desember 2017.

Yi Liu dan Y.F. Zheng. *One-against-all multi-class svm classification using reliability measures*. In Neural Networks, 2005. IJCNN ’05. Proceedings. 2005 IEEE International Joint Conference. **2(31)**:849 – 854.

LAMPIRAN

Lampiran 1 ***Package Bioinformatics***

```
> Usepackages <- function(x){
+   for( i in x ){
+     #  require returns TRUE invisibly if it was able
to load package
+     if( ! require( i , character.only = TRUE ) ){
+       #  If package was not able to be loaded then
re-install
+       install.packages( i , dependencies = TRUE )
+       source('http://bioconductor.org/biocLite.R')
+       biocLite(i)
+       #  Load package after installing
+       require( i , character.only = TRUE ) } } }
> package <- c('affy','GEOquery','Biobase',
+              'simpleaffy','affyPLM','hgu133plus2.db',
+
'hgu133acdf','hgu133a.db','hgu133plus2cdf',
+              'genefilter','AnnotationDbi')
> Usepackages (package)
```

Lampiran 2 ***Input data***

```
> setwd("D:/Medulloblastoma_brain_tumor")
> #input data
> gse<- list.celfiles("D:/Medulloblastoma_brain_tumor/
  GSE37418_RAW", full.names=T)
> gse37418<- ReadAffy(filenames=gse)
> #get pheno data
```

```
> gset <- getGEO(GEO="GSE37418")

> data.gse <- exprs(gset[[1]])

> pheno <- pData(phenoData(gset[[1]]))

> #save pheno in txt file

> WriteMatrixToFile <- function(tmpMatrix, tmpFileName,
blnRowNames, blnColNames)

+ { output <- file(tmpFileName, "at")

+   utils::write.table(tmpMatrix, output, sep = "\t", quote
= FALSE,

+                      row.names = blnRowNames, col.names =
blnColNames)

+   close(output)}

> WriteMatrixToFile(tmpMatrix=pheno,
tmpFileName="pheno1.txt",

+                   blnRowNames=TRUE, blnColNames=TRUE)

> ## PHENO DATA

>  p1  <-  read.AnnotatedDataFrame(file.path("pheno1.txt"),
  sep= "\t", header = TRUE)

> phenoData(gse37418) <- p1

> pheno =  pData(phenoData(gse37418)

> varMetadata(p1)
```

Lampiran 3 *subset sample*

```
> newgse37418<-get.array.subset.affybatch(gse37418,
"characteristics_ch1", c("subgroup: SHH", "subgroup: G4",
"subgroup: G3", "subgroup: WNT"))

> #selected pheno sample to use\\

> selected<- c('GSM918578', 'GSM918579', 'GSM918580',
'GSM918581', 'GSM918582', 'GSM918583', 'GSM918584',
'GSM918585', 'GSM918586', 'GSM918587', 'GSM918588',
'GSM918589', 'GSM918590', 'GSM918591', 'GSM918592',
'GSM918593', 'GSM918594', 'GSM918595', 'GSM918596',
```

```
'GSM918597', 'GSM918598', 'GSM918599', 'GSM918600',
'GSM918601', 'GSM918602', 'GSM918603', 'GSM918604',
'GSM918605', 'GSM918606', 'GSM918607', 'GSM918608',
'GSM918609', 'GSM918610', 'GSM918611', 'GSM918612',
'GSM918613', 'GSM918614', 'GSM918615', 'GSM918616',
'GSM918617', 'GSM918619', 'GSM918620', 'GSM918621',
'GSM918622', 'GSM918623', 'GSM918624', 'GSM918625',
'GSM918626', 'GSM918627', 'GSM918629', 'GSM918630',
'GSM918631', 'GSM918632', 'GSM918633', 'GSM918634',
'GSM918635', 'GSM918636', 'GSM918637', 'GSM918638',
'GSM918639', 'GSM918640', 'GSM918641', 'GSM918642',
'GSM918643', 'GSM918645', 'GSM918646', 'GSM918647',
'GSM918648', 'GSM918649', 'GSM918650', 'GSM918651',
'GSM918652', 'GSM918653')

> p1<-p1[selected,]

> phenoData(newgse37418) <- p1

> pheno = pData(phenoData(newgse37418))
```

Lampiran 4 *Piechart*

```
> des <-table(pheno$characteristics_ch1)

> percent <- round(des/sum(des)*100)

> des <- as.data.frame(des)

> lbls <- paste(des$Var1,'-',percent, '%', sep='')

> pie(des$Freq, label= lbls, col=
c('green','blue','orange','red'))
```

Lampiran 5 *Barplot gender*

```
> des2 <-table(pheno$characteristics_ch1.2)

> des2 <- as.data.frame(des2)

> barplot(des2$Freq, xlab= 'Gender',ylab='Pasien',
names.arg= c("male","female"))
```

Lampiran 6 *ethnic*

```
> des3 <-table(pheno$characteristics_ch1.3)

> percent3 <- round(des3/sum(des3)*100)

> des3 <- as.data.frame(des3)
```

```
> lbls3 <- paste(des3$Var1,'-',percent3, '%', sep='')

> pie(des3$Freq, label= lbls3, col= c(7,2,1,4,6,5,'white'))
```

Lampiran 7 ***Preprocessing data***

```
> rma_data<- threestep(newgse37418, background.method =
"RMA.2", normalize.method="quantile",

+                      summary.method="median.polish")

> dim(rma_data)

Features  Samples

   54675       73
```

Lampiran 8 ***Filtering***

```
> filter <- nsFilter(rma_data, require.entrez =T,
remove.dupEntrez = T,var.cutoff = 0.5, feature.exclude =
"^AFFX")

> log <-filter$filter.log

> eset <- filter$eset

> featureNames(eset) <- make.names(featureNames(eset))

> View(eset)

> dim(eset)

Features  Samples

   10251       73

> f1 <- Anova(eset$characteristics_ch1, p=0.001)

> flist <- filterfun(f1)

> ffilter<- genefilter(exprs(eset), flist )

> summary(ffilter)

   Mode   FALSE    TRUE

logical    5025    5226

> esetf<- eset[ffilter,]

> featureNames(esetf) <- make.names(featureNames(esetf))
```

```
> View(esetf)
> dim(esetf)
Features  Samples
    5226       73
```

Lampiran 9 *Make ExpressionSet*

```
> exp = exprs(esetf)
> as(esetf,"ExpressionSet")
ExpressionSet (storageMode: lockedEnvironment)
assayData: 5226 features, 73 samples
  element names: exprs, se.exprs
protocolData: none
phenoData
  sampleNames: GSM918578 GSM918579 ... GSM918653 (73 total)
  varLabels: title geo_accession ... data_row_count (37
total)
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
Annotation: hgu133plus2
> all(rownames(pData)==colnames(exp))
[1] TRUE
> newData<- ExpressionSet(assayData = exp,
+                         phenoData = p1,
+                         annotation = "hgu133plus2" )
```

Lampiran 10 *Packages* SVM

```
> analisis <- c('e1071','pROC')

> Usepackages(analisis)
```

Lampiran 11 Penyusunan data set

```
> data   = as.data.frame (t(exprs(newData)))

> gender  = as.factor(newData$characteristics_ch1.2)

> ethnic = as.factor(newData$characteristics_ch1.3)

> dataY = as.factor(newData$characteristics_ch1)

> datause = as.data.frame(cbind(data,ethnic,gender,dataY))

> ratio = 7/10

> G3 = datause[dataY=='subgroup: G3',]

> G4 = datause[dataY=='subgroup: G4',]

> SHH = datause[dataY=='subgroup: SHH',]

> WNT = datause[dataY=='subgroup: WNT',]

> sampelg3 = sample(nrow(G3), size = floor(ratio*nrow(G3)))

> sampelg4 = sample(nrow(G4), size = floor(ratio*nrow(G4)))

> sampelshh = sample(nrow(SHH), size =
floor(ratio*nrow(SHH)))

> sampelwnt = sample(nrow(WNT), size =
floor(ratio*nrow(WNT)))

> sampelg3 = c(7 ,3 ,4 ,8 ,1, 15, 10, 6, 12, 9, 14)

> sampelg4 = c(30, 19, 31, 9, 34, 2, 12, 3, 17, 6, 4, 14,
27, 38, 13, 29, 10, 20, 25, 22, 33, 16, 36, 21, 11, 39, 26)

> sampelshh = c(10, 2, 8, 6, 7, 4, 3)

> sampelwnt = c(4, 6, 5, 3, 2)

> trainG3= G3[sampelg3,]

> trainG4= G4[sampelg4,]

> trainSHH= SHH[sampelshh,]

> trainWNT= WNT[sampelwnt,]
```

```
> testG3= G3[-sampelg3,]

> testG4= G4[-sampelg4,]

> testSHH= SHH[-sampelshh,]

> testWNT= WNT[-sampelwnt,]

> datatrain = rbind(trainG3, trainG4, trainSHH, trainWNT)

> datatest = rbind(testG3, testG4, testSHH, testWNT)
```

Lampiran 12 *Tuning*

```
> tunaslin <- tune(svm, dataY~. , data = datatrain,
kernel="linear",types = "C-clasification",ranges= list( cost
= c(0.1,0.01, 0.001,1 , 10 , 100)))

> summary(tunaslin)

Parameter tuning of ‘svm’:

- sampling method: 10-fold cross validation

- best parameters:

 cost

  0.1

- best performance: 0.06

- Detailed performance results:

   cost error dispersion

1 1e-01  0.06 0.09660918

2 1e-02  0.06 0.09660918

3 1e-03  0.06 0.09660918

4 1e+00  0.06 0.09660918

5 1e+01  0.06 0.09660918

6 1e+02  0.06 0.09660918
```

```
> model <- svm(dataY~. , datatrain, kernel = "linear",
  cost=0.1,scale=F,     types     =     "C-clasification",
  decision.value=T)

> predictions <- predict(model, datatest)
```

Lampiran 13*Confusion matrix*

```
> table(predictions,datatest$dataY)

predictions     subgroup: G3 subgroup: G4 subgroup: SHH subgroup: WNT
  subgroup: G3             4            0             0             0
  subgroup: G4             1           12             0             0
  subgroup: SHH            0            0             3             0
  subgroup: WNT            0            0             0             3

> mean(predictions == datatest$dataY)

[1] 0.9565217

> predictions <- predict(model, datatrain)

> table(predictions,datatrain$dataY)

predictions     subgroup: G3 subgroup: G4 subgroup: SHH subgroup: WNT
  subgroup: G3            11            0             0             0
  subgroup: G4             0           27             0             0
  subgroup: SHH            0            0             7             0
  subgroup: WNT            0            0             0             5

> mean(predictions == datatrain$dataY)

[1] 1
```

Lampiran 14 Bobot

```
> w <- t(model$coefs) %*% model$SV

> w <- apply(w, 2, function(v){sqrt(sum(v^2))})

> w <- sort(w, decreasing = T)

> x<- as.data.frame(w)

> View(x)
```

```
> all(colnames(w)==colnames(datatrain))
[1] TRUE
```

Lampiran 15 AUC

```
> forauc <- as.numeric(predict(model,datatest))
> auc <- multiclass.roc(datatest$dataY,forauc)
> auc$auc
Multi-class area under the curve: 0.9833
```

Lampiran 16 Gene name dan gene ontology

```
> id <- substring(as.character(head(rownames(x), n=10)),2)
> select(hgu133plus2.db, id, "GENENAME")
'select()' returned 1:1 mapping between keys and columns
       PROBEID                                      GENENAME
1   1559213_at long intergenic non-protein coding RNA 1419
2   201049_s_at                      ribosomal protein S18
3   200819_s_at                      ribosomal protein S15
4   200763_s_at  ribosomal protein lateral stalk subunit P1
5   201665_x_at                      ribosomal protein S17
6   221798_x_at                       ribosomal protein S2
7   208645_s_at                      ribosomal protein S14
8   215963_x_at                       ribosomal protein L3
9   213414_s_at                      ribosomal protein S19
10   200013_at                      ribosomal protein L24
> genname <- select(hgu133plus2.db, id,
  c("SYMBOL","GENENAME"), "PROBEID")
> ontology <- select(hgu133plus2.db, id,
  c("SYMBOL","GENENAME","ENTREZID","ONTOLOGY"), "PROBEID")
> genname
```

```
       PROBEID SYMBOL                                      GENENAME
1  201049_s_at  RPS18                        ribosomal protein S18
2  200763_s_at  RPLP1 ribosomal protein lateral stalk subunit P1
3  200819_s_at  RPS15                        ribosomal protein S15
4  201665_x_at  RPS17                        ribosomal protein S17
5  200062_s_at  RPL30                        ribosomal protein L30
6    200013_at  RPL24                        ribosomal protein L24
7  215963_x_at   RPL3                         ribosomal protein L3
8  208645_s_at  RPS14                        ribosomal protein S14
9    208692_at   RPS3                         ribosomal protein S3
10   200010_at  RPL11                        ribosomal protein L11
> ontology
        PROBEID SYMBOL                                      GENENAME ENTREZID
ONTOLOGY
1  201049_s_at  RPS18                        ribosomal protein S18     6222       BP
2  201049_s_at  RPS18                        ribosomal protein S18     6222       CC
3  201049_s_at  RPS18                        ribosomal protein S18     6222       MF
4  200763_s_at  RPLP1 ribosomal protein lateral stalk subunit P1     6176       BP
5  200763_s_at  RPLP1 ribosomal protein lateral stalk subunit P1     6176       CC
6  200763_s_at  RPLP1 ribosomal protein lateral stalk subunit P1     6176       MF
7  200819_s_at  RPS15                        ribosomal protein S15     6209       BP
8  200819_s_at  RPS15                        ribosomal protein S15     6209       CC
9  200819_s_at  RPS15                        ribosomal protein S15     6209       MF
10 201665_x_at  RPS17                        ribosomal protein S17     6218       BP
11 201665_x_at  RPS17                        ribosomal protein S17     6218       CC
12 201665_x_at  RPS17                        ribosomal protein S17     6218       MF
13 200062_s_at  RPL30                        ribosomal protein L30     6156       BP
14 200062_s_at  RPL30                        ribosomal protein L30     6156       CC
15 200062_s_at  RPL30                        ribosomal protein L30     6156       MF
16   200013_at  RPL24                        ribosomal protein L24     6152       BP
17   200013_at  RPL24                        ribosomal protein L24     6152       CC
```

18	200013_at	RPL24	ribosomal protein L24	6152	MF
19	215963_x_at	RPL3	ribosomal protein L3	6122	BP
20	215963_x_at	RPL3	ribosomal protein L3	6122	CC
21	215963_x_at	RPL3	ribosomal protein L3	6122	MF
22	208645_s_at	RPS14	ribosomal protein S14	6208	BP
23	208645_s_at	RPS14	ribosomal protein S14	6208	CC
24	208645_s_at	RPS14	ribosomal protein S14	6208	MF
25	208692_at	RPS3	ribosomal protein S3	6188	BP
26	208692_at	RPS3	ribosomal protein S3	6188	CC
27	208692_at	RPS3	ribosomal protein S3	6188	MF
28	200010_at	RPL11	ribosomal protein L11	6135	BP
29	200010_at	RPL11	ribosomal protein L11	6135	CC
30	200010_at	RPL11	ribosomal protein L11	6135	MF

Lampiran 17 Session Info

```
> sessionInfo()

R version 3.4.2 (2017-09-28)

Platform: x86_64-w64-mingw32/x64 (64-bit)

Running under: Windows 10 x64
```