UNIVERSITAS INDONESIA

METODE SELEKSI FITUR BERBASIS PERANKINGAN BOBOT SECARA MULTI STEP MENGGUNAKAN DEEP LEARNING UNTUK PENCARIAN BIOMARKER PADA DATA MICROARRAY

THESIS

MUKHLIS AMIEN
1406522102

FAKULTAS ILMU KOMPUTER
PROGRAM STUDI MAGISTER ILMU KOMPUTER
DEPOK
JUNI 2016

UNIVERSITAS INDONESIA

METODE SELEKSI FITUR BERBASIS PERANKINGAN BOBOT SECARA MULTI STEP MENGGUNAKAN DEEP LEARNING UNTUK PENCARIAN BIOMARKER PADA DATA MICROARRAY

THESIS
Diajukan sebagai salah satu syarat untuk memperoleh gelar
Master

MUKHLIS AMIEN
1406522102

FAKULTAS ILMU KOMPUTER
PROGRAM STUDI MAGISTER ILMU KOMPUTER
DEPOK
JUNI 2016

HALAMAN PERSETUJUAN

Judul	:	Metode Seleksi Fitur Berbasis Perankingan Bobot Secara Multi Step Menggunakan Deep Learning untuk Pencarian Biomarker pada Data Microarray
Nama	:	Mukhlis Amien
NPM	:	1406522102

Laporan Thesis ini telah diperiksa dan disetujui.

XX Januari 2016

Ito Wasito PhD.

Pembimbing Thesis

HALAMAN PERNYATAAN ORISINALITAS

Thesis ini adalah hasil karya saya sendiri,
dan semua sumber baik yang dikutip maupun dirujuk
telah saya nyatakan dengan benar.

Nama	:	**Mukhlis Amien**
NPM	:	**1406522102**
Tanda Tangan	:	
Tanggal	:	**XX Januari 2016**

HALAMAN PENGESAHAN

Thesis ini diajukan oleh	:	
Nama	:	Mukhlis Amien
NPM	:	1406522102
Program Studi	:	Magister Ilmu Komputer
Judul Thesis	:	Metode Seleksi Fitur Berbasis Perankingan Bobot Secara Multi Step Menggunakan Deep Learning untuk Pencarian Biomarker pada Data Microarray

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Master pada Program Studi Magister Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Indonesia.

DEWAN PENGUJI

Pembimbing	:	Ito Wasito PhD.	(	)
Penguji	:	Prof. XXX	(	)
Penguji	:	Prof. XXXX	(	)
Penguji	:	Prof. XXXXXX	(	)

@todo

Jangan lupa mengisi nama para penguji.

Ditetapkan di	:	Depok
Tanggal	:	XX Januari 2016

KATA PENGANTAR

Pertama-tama saya ucapkan syukur kehadirat Allah SWT dan Junjungan kita Nabi Muhammad SAW, atas segala petunjuk dan karunianyalah thesis ini bisa terselesaikan. Saya mengucapkan terima kasih yang sebesar-besarnya atas bimbingan dan petunjuk dari pembimbing saya Bpk Ito Wasito PhD. Kepada istri saya, Catur Prastiasih yang sangat mendukung saya melanjutkan sekolah. Kepada teman-teman satu bimbingan yaitu Aris dan Arida yang telah memberikan ide dan masukan serta diskusi yang mendalam.

Kepada Fakultas Ilmu Komputer Universitas Indonesia, yang telah memberikan fasilitas lab yang sangat membantu saya dalam menyelesaikan thesis ini.

Kepada orang tua saya, terima-kasih banyak atas dukungan moral dan spiritual yang diberikan. Dan saudara-saudara saya di Batu.

Depok, 30 Desember 2009

Mukhlis Amien

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Indonesia, saya yang bertanda tangan di bawah ini:

Nama	:	Mukhlis Amien
NPM	:	1406522102
Program Studi	:	Magister Ilmu Komputer
Fakultas	:	Ilmu Komputer
Jenis Karya	:	Thesis

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia **Hak Bebas Royalti Noneksklusif (Non-exclusive Royalty Free Right)** atas karya ilmiah saya yang berjudul:

Metode Seleksi Fitur Berbasis Perankingan Bobot Secara Multi Step Menggunakan Deep Learning untuk Pencarian Biomarker pada Data Microarray

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Indonesia berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (database), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyatan ini saya buat dengan sebenarnya.

Dibuat di : Depok
Pada tanggal : XX Januari 2016

Yang menyatakan

(Mukhlis Amien)

ABSTRAK

Nama	:	Mukhlis Amien
Program Studi	:	Magister Ilmu Komputer
Judul	:	Metode Seleksi Fitur Berbasis Perankingan Bobot Secara Multi Step Menggunakan Deep Learning untuk Pencarian Biomarker pada Data Microarray

Data ekspresi gen pada percobaan microarray memiliki ciri khas yaitu jumlah sampel yang sedikit dengan dimensi fitur yang sangat besar. Algoritma *Deep Believe Network (DBN)* adalah bagian dari algoritma *deep learning* yang menerapkan teknik *unsupervised learning*, digunakan untuk membantu menganalisa data ekspresi gen. Algoritma seleksi fitur yang berbasis pada perankingan bobot secara multi-step digunakan untuk mendapatkan fitur gen yang paling penting. Pemilihan fitur yang paling informatif dari suatu kasus percobaan microarray, merupakan masalah yang ada pada pemrosesan data ekspresi gen. Sehingga diperlukan eksplorasi lebih lanjut untuk pemilihan fiturnya. Seleksi fitur gen, berdasarkan ranking bobot yang dihasilkan oleh *deep learning*, diharapkan dapat memecahkan masalah seleksi fitur tersebut. Sedangkan metode clustering yang dipakai adalah: Cluster Affinity Search Technique (CAST) , K-Means Clustering dan Hierarchical Clustering. Deep learning adalah metode pembelajaran mesin yang merupakan bagian dari algoritma neural network, metode deep learning yang sering dipakai adalah arsitektur deep believe network (DBN). Pendekatan unsupervised learning pada deep learning dan clustering, diharapkan dapat digunakan untuk membantu peneliti dalam menganalisa data ekspresi gen-nya.

Kata Kunci:
Microarray, ekspresi gen, Algoritma Clustering, feature selection, deep learning, unsupervised learning.

ABSTRACT

Name	:	Mukhlis Amien
Program	:	Magister Ilmu Komputer
Title	:	FEATURE SELECTION METHOD BASED ON MULTI STEP WEIGHT RANKING USING DEEP LEARNING TO SEARCH BIOMARKER IN LUNG CANCER MICROARRAY DATA SET

Microarray technology has made possible the profiling of gene expressions of the entire genome in a single hybridization experiment. Since microarray data acquire tens of thousands of gene expression values simultaneously. However, the number of sample usually small. Feature selection and clustering algorithm for microarray data analysis is useful to extract cluster structure and to reduce the high dimensional microarray data and reconstruct to lower dimensional with minimum error possible. Deep learning and clustering is a machine learning method. In this research we will investigate the effectiveness of clustering after or prior dimensionality reduction. The most common deep learning architecture used for dimensionality reduction is deep believe network (DBN) and stacked auto encoder (SAE). Pre training unsupervised learning and greedy layer wise training approach are expected for better dimensionality reduction in microarray datasets compared with other methods.

Keywords:
Microarray, ekspresi gen, Algoritma Clustering, feature selection, deep learning, unsupervised learning.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Data ekspresi gen pada percobaan *microarray* memiliki ciri khas yaitu dimensi fitur gen yang jauh lebih besar dibandingkan dengan sampel pasien. Masalah tersebut menyebabkan penerapan teknik pendeteksian penyakit genetis dengan menggunakan data ekspresi gen lebih sulit dilakukan, dikarenakan data ekspresi gen tersebut memiliki signifikansi yang berbeda-beda. Menurut penelitian Yoon et al. (2006) dan Bandyopadhyay et al. (2014) tidak semua ekspresi gen yang didapatkan dalam percobaan *microarray* tersebut adalah gen yang informatif, bahkan jumlah ekspresi gen yang informatif untuk kasus yang diinginkan misalnya untuk pengenalan sel kanker, sangat sedikit dibandingkan dengan keseluruhan ekspresi gen yang didapatkan dalam sebuah percobaan (Bandyopadhyay et al., 2014). Data ekspresi gen yang tidak informatif tersebut dapat mengganggu dan mengurangi performa secara signifikan pada teknik pengenalan pola penyakit yang diterapkan. Akan tetapi, beberapa gen yang informatif berpengaruh secara signifikan terhadap pengenalan pola tersebut. Sebagai contoh, untuk mendiagnosa kanker paru-paru, hanya dibutuhkan sekitar 50 gen saja dari 22 ribu gen yang didapatkan dalam percobaan. Gen-gen yang paling informatif ini disebut dengan *Biomarker* (Belinsky, 2004). Sehingga hanya dengan menggunakan data *Biomarker* yang ditemukan saja, sudah dapat digunakan untuk mengenali penyakit yang diderita oleh pasien.

Pada penelitian ini, akan dibangun sebuah teknik pencarian *Biomarker* dengan metode seleksi fitur gen. Metode ini menerapkan perankingan gen secara *multi step* terhadap model yang didapatkan pada proses *training*. Arsitektur yang digunakan adalah arsitektur *Deep Belief Network (DBN)* yang merupakan bagian dari metode *deep learning*. Metode perankingan yang digunakan adalah modifikasi dari algoritma seleksi fitur untuk *logistic regression* yang dilakukan oleh Shevade and Keerthi (2003). Akan tetapi metode ini memiliki masalah dalam mengeliminasi fitur jika diterapkan secara langsung pada model DBN, dikarenakan parameter bobot (W) dan bias (b) ditempatkan disetiap fitur dan model ini hanya memiliki satu layer dibandingkan dengan DBN yang memiliki banyak layer.

DBN merupakan jaringan *Restrictive Boltzmann Machine (RBM)* yang disusun secara bertingkat. Dimulai dengan memberikan bobot random diantara dua network, yang dapat dilatih dengan cara meminimalkan perbedaan antara data asli dengan data rekonstruksinya. *Gradien* didapatkan dengan *chain rule* untuk melakukan penurunan error dengan teknik *Contrastive Divergence (CD)*. Untuk dicari bobot (W) dan bias dengan *maximum likelihood learning* secara *greedy* pada tiap layernya (Hinton and Salakhutdinov, 2006).

Pada DBN, *hidden unit* yang paling sering aktif adalah *hidden unit* yang lebih penting dibandingkan dengan *hidden unit* yang jarang aktif, oleh karena itu *hidden unit* ini memiliki parameter bobot yang lebih besar dibandingkan dengan hidden unit yang jarang aktif pada saat proses *training* dilakukan. Pemilihan fitur dilakukan dengan meranking unit-unit yang memiliki bobot tertinggi dimulai dari *layer output* menuju *layer input* untuk mendapatkan fitur gen yang paling berpengaruh. Kemudian dilakukan eliminasi bobot pada *hidden unit* per layernya secara *multi step*. Selanjutnya akan dipilih sebanyak *top-n* gen dari hasil perankingan ini untuk dievaluasi apakah *Biomarker* yang ditemukan tersebut informatif atau tidak.

Tahapan berikutnya, fitur yang telah didapatkan akan digunakan sebagai data input pada *Multi Layer Perceptron* (MLP) dengan tujuan untuk melakukan evaluasi apakah gen *Biomarker* yang ditemukan dengan perankingan tersebut dapat memperbaiki hasil klasifikasi pasien sakit atau sehat. Untuk mengetahui keakuratannya, dilakukan perbandingan hasil eksperimen ini dengan hasil pada eksperimen lain pada literatur yang juga bertujuan untuk menemukan *Biomarker*.

1.2 Rumusan Masalah

Berdasarkan pada uraian pendahuluan diatas maka dapat dibuat rumusan permasalahan sebagai berikut: Dikarenakan karakteristik sedikitnya sampel dan besarnya fitur pada data ekspresi gen serta signifikansi pencarian Biomarker pada penyakit yang disebabkan oleh genetis, maka apakah metode seleksi fitur berbasis perankingan bobot secara multi step menggunakan deep learning untuk pencarian Biomarker tersebut dapat diterapkan?

1.3 Batasan Permasalahan

- Dataset yang digunakan adalah data ekspresi gen microarray untuk penyakit kanker paru-paru yang tersedia secara bebas dengan kode GSE10072
- Data yang digunakan adalah dataset yang sudah dilakukan pengolahan awal standar.
- Komputer yang digunakan adalah laptop lenovo core i7 dengan memory 8 Gb.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk:

- Membangun metodologi pencarian *Biomarker* pada dataset ekspresi gen percobaan *microarray*.
- Membuat algoritma perankingan gen secara multi step yang diterapkan pada arsitektur DBN.
- Melakukan evaluasi apakah *Biomarker* yang ditemukan oleh metode ini untuk dilakukan verifikasi dengan literatur.

1.5 Manfaat Penelitian

Hasil dari penelitian ini memiliki manfaat :

- Framework DBN untuk pencarian *Biomarker* ini dapat diterapkan untuk mendeteksi apakah seseorang memiliki resiko genetis penyakit kanker paru-paru.
- Mendapatkan fitur gen yang paling penting dan informatif pada kasus penyakit kanker paru-paru.
- Melakukan pendeteksian kanker paru-paru secara dini dengan data yang didapatkan dari profil gen pasien pada eksperimen*microarray*.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Berisi gambaran umum permasalahan dan metodologi apa yang akan diterapakan.

- Bab 2 TINJAUAN PUSTAKA
  Landasan teori dipakainya metodologi yang akan diterapkan dalam eksperimen ini.

- Bab 3 METODOLOGI PENELITIAN
  Penjelasan detail metodologi yang akan diterapkan dalam penelitian.

- Bab 4 PEMBAHASAN
  Pembahasan hasil dari eksperimen yang sudah dilakukan.

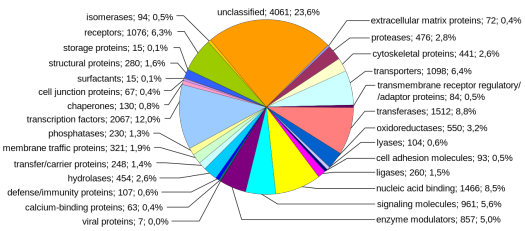
- Bab 5 KESIMPULAN DAN SARAN

BAB 2
TINJAUAN PUSTAKA

2.1 Ekspresi Gen

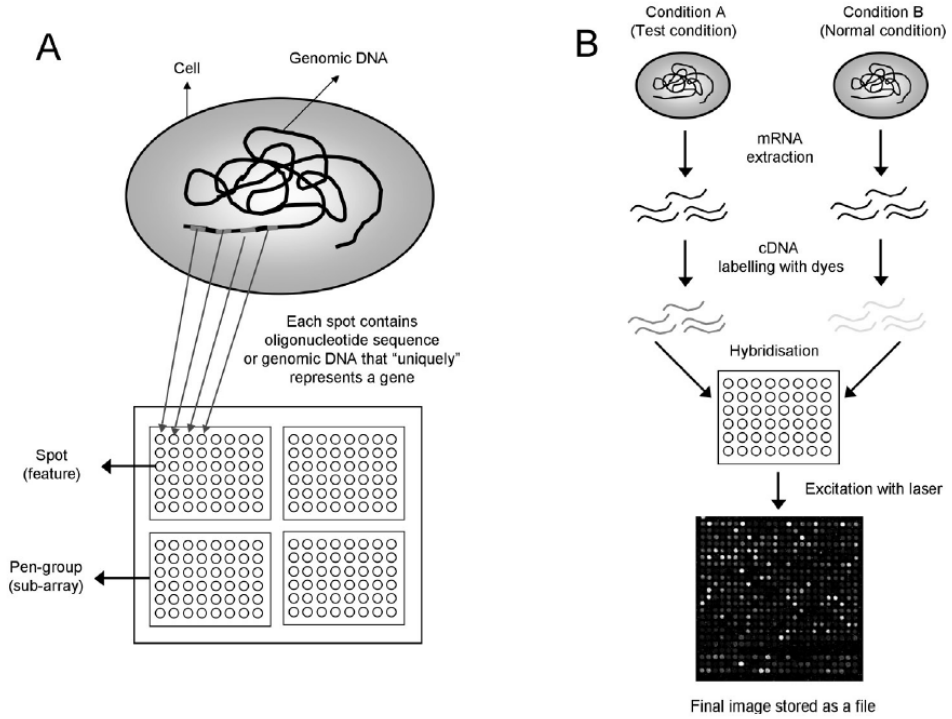
Percobaan *microarray*, mengukur tingkat aktivitas gen di dalam sebuah jaringan sel. Sehingga dapat memberikan informasi berdasarkan aktivitas di dalam jaringan yang bersangkutan. Data ini didapatkan dengan cara mengukur banyaknya mRNA yang diproduksi pada saat proses transkripsi DNA, dimana dapat diukur seberapa aktif atau seberapa berfungsinya gen tersebut dalam sebuah jaringan (Elloumi and Zomaya, 2011). Karena kanker berhubungan dengan berbagai macam aktivitas penyimpangan regulasi pada sel, maka data ekspresi gen pada kanker merefleksikan penyimpangan regulasi tersebut. Untuk menangkap keabnormalan ini, percbaan microarray, dimana dapat mengukur secara simultan dari level ekspresi ratusan bahkan ribuan ekspresi gen dapat digunakan untuk mengidentifikasi kanker. Percobaan microarray sering dipakai untuk membandingkan profil ekspresi gen pada sel yang terkena kanker, dibandingkan dengan sel yang normal pada berbagai macam percobaan. Percobaan microarray digunakan untuk mengidentifikasi ekspresi yang berbeda pada dua percobaan, yang biasanya berupa data tes dan data kontrol (Elloumi and Zomaya, 2011).

Ada 23.6% fungsi gen yang belum diketahui kegunaannya sampai saat ini, hal ini merupakan tantangan pada saat dilakukan proses pengenalan penyakit yang diderita oleh pasien. Dikarenakan ada kemungkinan gen yang sangat berpengaruh terhadap identifikasi penyakit, masih belum diketahui fungsinya. Oleh karena itu, pada proses klasifikasi penyakit dengan menggunakan machine learning, sering digunakan pengenalan secara *unsupervised learning* (Häggström, 2014).



Gambar 2.1: Ada 23,6% dari keseluruhan fungsi gen yang belum diketahui, sehingga pengetahuan tentang fungsi gen masih belum lengkap. (Häggström, 2014)

Data ekspresi gen yang masih mentah didapatkan dari percobaan di laboratorium menggunakan alat yang dinamakan dengan alat Genchip microarray. Data tersebut kemudian dilakukan pemrosesan awal untuk mendapatkan sebuah matriks ekspresi gen. Matriks ini memiliki data kolom dan baris, dimana kolom berisi data eksperimen, dan baris berisi nilai ekspresi pada tiap-tiap gen (Gambar 2.2) (Babu, 2004).



Gambar 2.2: Proses Keseluruhan Percobaan Microarray.(Yoon et al., 2006)

Pengukuran microarray direspresentasikan dengan tabel gen ekspresi, dimana bagian barisnya adalah fitur ekspresi gen, dan bagian kolom merepresentasikan pasien.

Table 1.A: Absolute measurement

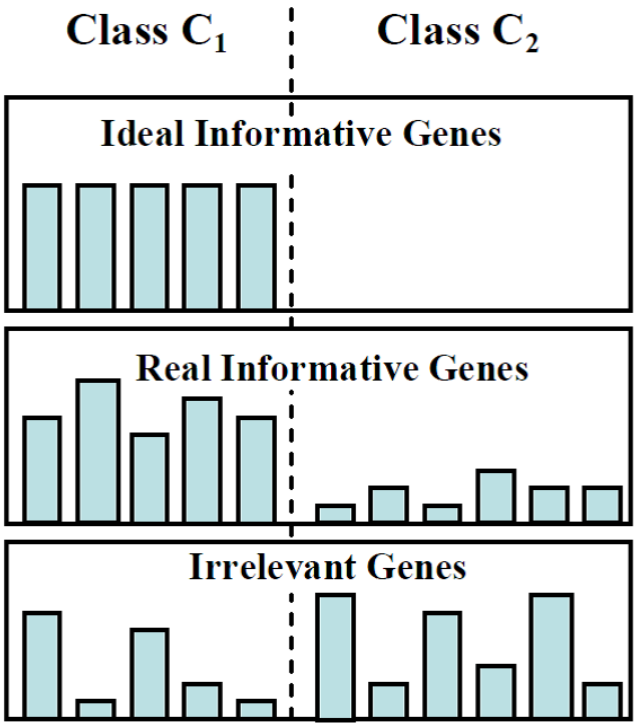
	C1	C2	C3	C4
Gene A	10	80	40	20
Gene B	100	200	400	200
Gene C	30	240	60	60
Gene D	20	160	80	80

Table 1.B: Relative measurement

	C1/C4	C2/C4	C3/C4
Gene A	0.50	4.00	2.00
Gene B	0.50	1.00	2.00
Gene C	0.50	4.00	1.00
Gene D	0.25	2.00	1.00

Gambar 2.3: Contoh data pengukuran percobaan microarray (Yoon et al., 2006)

Karena data microarray yang didapatkan dapat mencapai ribuan ekspresi dalam satu waktu secara simultan, maka data ini dapat sangat membantu dalam mengidentifikasi penyakit. Akan tetapi, hasil yang didapat dengan menganalisa beberapa data microarray yang dilakukan oleh dua percobaan yang berbeda tetapi dengan tujuan yang sama, dapat menghasilkan hasil yang sangat berbeda. Salah satu alasannya adalah terbatasnya sampel dan terlalu banyaknya profil ekspresi gen. Sehingga diperlukan metode testing statistik untuk memastikan bahwa data microarray tersebut memiliki tingkat signifikansi yang cukup, dan dipastikan bahwa perbedaan tersebut memang karena eksperimen, bukan karena kerusakan alat atau kesalahan prosedur eksperimen.



Gambar 2.4: Perbandingan Ekspresi gen yang relevan dan informatif dibandingkan dengan gen yang tidak relevan(Babu, 2004)

2.2 Pemrosesan Data Microarray

Data yang dihasilkan dari alat microarray ini berupa citra yang perlu diproses lebih lanjut. Sebelum data ekspresi gen dapat dianalisa lebih lanjut, perlu dilakukan pemrosesan awal yang berupa (i) perbaikan background, (ii) normalisasi data dan kemudian (iii) penyaringan data.

1. **Perbaikan Background**

   Perbiakan background ini ditujukan untuk menghilangkan titik-titik noise

yang tidak berasal dari proses hibridisasi. Metode untuk perbaikan background ini banyak diajukan dalam penelitian(Fakoor et al., 2013).

2. **Normalisasi**
   Tujuan dari normalisasi adalah untuk mengatur bias yang dihasilkan oleh variasi proses percobaan microarray. Metode normalisasi data microarray ada banyak, dan pada penelitian ini akan digunakan normalisasi standar untuk data microarray.

3. **Penyaringan data** Tidak semua data yang didapat dari percobaan microarray bagus, kadangkala terjadi kesalahan alat dan noise yang diakibatkan oleh alat, oleh karena itu perlu disaring, mana data yang disebabkan oleh proses biologi, dan mana yang disebabkan oleh noise alat.

4. **Missing Value Imputation**
   Tidak semua data ekspresi gen dapat kita dapatkan, dikarenakan rumitnya percobaan microarray, kadangkala data tidak kita dapatkan, oleh sebab itu diperlukan metode untuk melakukan pendekatan statistic dalam memberikan perkiraan isi data dalam titik data yang hilang tersebut.

5. **Seleksi Fitur**
   Setelah proses diatas, diperlukan teknik untuk menseleksi fitur pada data microarray. Ada banyak metode yang sudah diusulkan oleh para peneliti. Seperti pada table 1 dibawah. Dan pada titik inilah penelitian ini dijalankan. Diharapkan penelitian ini menghasilkan metode reduksi dimensi untuk data microarray.

2.3 Ekstraksi Fitur dan Seleksi Fitur Pada Penelitian Sebelumnya

Pada tabel dibawah ditunjukkan perbandingan penelitian-penelitian ekstraksi fitur dengan menggunakan berbagai macam metode.

Tabel 2.1: Perbandingan Metode Seleksi fitur pada dataset microarray

Pengarang	Judul Paper	Metode	Dataset
C. Aliferis et al. 2003	Machine learning models for classication of lung cancer and selection of genomic markers using array gene expression data.	Reduksi fitur secara rekursif dan melakukan filter secara asosiasi univariate	Lung Cancer Microarray
Ramaswamy, S. et al. 2001	Multiclass cancer diagnosis using tumor gene expression signatures.	Pengurangan fitur secara rekursif dengan mengguanakan SVM	Various Microarray
Wang et al., 2005	Gene-expression proles to predict distant metastasis of lymph-node-negative primary breast cancer.	Mengkombinasikan seleksi fitur yang berbasis korelasi dengan pendekatan assosiasi.	Various Microarray
Sharma et. Al, 2012	Combining multiple approaches for gene microarray classification.	Mengkombinasikan banyak pendekatan ekstraksi fitur	Various Microarray

2.4 Deep Learning

Sebelum tahun 2006, melakukan training dalam arsitektur *deep learning* selalu gagal. Percobaan untuk melakukan training dengan *feedforward neural network* memiliki hasil yang lebih buruk dibandingkan dengan arsitektur yang dangkal, yaitu arsitektur dengan layer 1 atau maksimum 2 layer.
Akan tetapi tiga paper yang terbit pada 2006 secara revolusioner telah merubah hal terssebut. Sehingga setelah tahun 2006 penelitian tentang *deep learning* menjadi lebih intensif sampai sekarang dengan segala variasi arsitekturnya. Salah satu variasi arsitektur *deep learning* yang dipakai dalam thesis ini adalah *arsitektur Deep Believe Network (DBN)*. Ketiga paper tersebut adalah:

1. Hinton, G. E., Osindero, S. and Teh, Y., A fast learning algorithm for deep belief nets Neural Computation 18:1527-1554, 2006 (Hinton et al., 2006)

2. Yoshua Bengio, Pascal Lamblin, Dan Popovici and Hugo Larochelle, Greedy Layer-Wise Training of Deep Networks, in J. Platt et al. (Eds), Advances in Neural Information Processing Systems 19 (NIPS 2006), pp. 153-160, MIT Press, 2007 (**?**).

3. MarcAurelio Ranzato, Christopher Poultney, Sumit Chopra and Yann LeCun Efficient Learning of Sparse Representations with an Energy-Based Model, in J. Platt et al. (Eds), Advances in Neural Information Processing Systems (NIPS 2006), MIT Press, 2007(**?**).

Learning secara *unsupervised* menggunakan *pretraining* secara tiap layer yang disebut dengan *greedy layer-wise training*, yaitu training dilakukan satu layer pada tiap satu waktu. Training ini dilakukan secara berjenjang pada layer selanjutnya. Kemudian dilakukan *supervised training* untuk melakukan *tuning parameter*, yang dimulai dari parameter hasil pretraining yang dilakukan sebelumnya.
DBN menggunakan RBM sebagai bagian terkecil dari layernya, yang menggunakan learning secara unsupervised yang merepresentasikan tiap layer. Sejak 2006, banyak sekali paper-paper yang mulai melakukan eksplorasi tentang deep learning ini, sehingga sejak saat itu deep learning merupakan salah satu teknik *machine learning* yang paling populer, bahkan sampai saat ini (Tutorial, 2014).

2.5 Energy-Based Model (EBM)

EBM mengaitkan sebuah energi skalar pada setiap konfigurasi variable yang diinginkan. Proses learning bertujuan untuk memodifikasi fungsi energi sehingga bentuknya memiliki sifat yang diinginkan. Sebagai contoh, misalnya diinginkan sebuah bentuk konfigurasi yang memiliki energi yang rendah, maka model probabilistik dari EBM didifinisikan sebagi distribusi probabilitas melalui fungsi energi sebagi berikut(Tutorial, 2014)

$$p(x) = \frac{e^{-E(x)}}{Z}. \tag{2.1}$$

Z adalah faktor normalisasi yang disebut sebagai fungsi partisi untuk menganalogikan dengan sistem fisika.

$$Z = \sum_{x} e^{-E(x)} \tag{2.2}$$

EBM bisa dilatih dengan cara melakukan (stochastic) gradient descent pada negative log-likelihood (NLL)-nya secara empiris pada data training. Adapun untuk logistic regression akan didifinisikan terlebih dahulu log-likelihood $\mathcal{L}(\theta, \mathcal{D})$ dan

fungsi loss-nya sebagai NLL $\ell(\theta, \mathcal{D})$ sebagai berikut:

$$\begin{aligned} \mathcal{L}(\theta, \mathcal{D}) &= \frac{1}{N} \sum_{x^{(i)} \in \mathcal{D}} \log\ p(x^{(i)}) \\ \ell(\theta, \mathcal{D}) &= -\mathcal{L}(\theta, \mathcal{D}) \end{aligned} \tag{2.3}$$

Menggunakan stochastic gradient $-\frac{\partial \log p(x^{(i)})}{\partial \theta}$, dimana θ adalah parameter dari modelnya(**?**).

2.5.1 EBM dengan Hidden Units

Pada banyak kasus, sampel x biasanya tidak terobservasi secara penuh, atau akan ditambahkan varibel yang tidak teobservasi secara langsung yang disebut dengan hidden unit, dimana hal ini berguna untuk meningkatkan ekspresivitas dari model. Sehingga dikenalkan bagian yang terobservasi disini dilambangkan dengan x, dan sebuah bagian yang tersembunyi h. Sehingga bisa ditulis sebagai:

$$P(x) = \sum_h P(x,h) = \sum_h \frac{e^{-E(x,h)}}{Z}. \tag{2.4}$$

Pada kasus ini, untuk melakukan pemetaan rumus yang mirip dengan rumus 2.4 , akan dikenalkan notasi (yang merupakan inspirasi dari fisika) yaitu free energy $\mathcal{F}(x)$, yang didifinisikan sebagai berikut:

$$\mathcal{F}(x) = -\log \sum_h e^{-E(x,h)} \tag{2.5}$$

Sehingga bisa diturunkan sebagai :

$$P(x) = \frac{e^{-\mathcal{F}(x)}}{Z} \text{ dengan } Z = \sum_x e^{-\mathcal{F}(x)}.$$

Data dari gradien NLL kemudian memiliki bentuk yang menarik yaitu:

$$-\frac{\partial \log p(x)}{\partial \theta} = \frac{\partial \mathcal{F}(x)}{\partial \theta} - \sum_{\tilde{x}} p(\tilde{x})\ \frac{\partial \mathcal{F}(\tilde{x})}{\partial \theta}. \tag{2.6}$$

Gradien diatas memiliki dua istilah, dimana hal tersebut mereferensikan pada fase positif dan fase negatif. Istilah positif dan negatif ini tidak merujuk pada tanda (positif/negatif) persamaan, akan tetapi merefleksikan efek pada kepadatan probabilitas yang didefinisikan oleh model. Istilah pertama, menambah probabilitas data training (dengan cara mengurangi free energy yg berhubungan), sedangkan istilah

kedua mengurangi probabilitias sampel yang digenerasi oleh model (Tutorial, 2014).

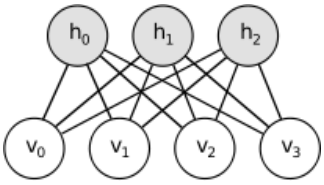
Biasanya sulit untuk menentukan gradien secara analitis, oleh karena berhubungan dengan komputasi dari $E_P[\frac{\partial \mathcal{F}(x)}{\partial \theta}]$. Dikarenakan hal ini merupakan ekspektasi semua kemungkinan konfigurasi input x (pada distribusi P yang dibentuk oleh model).
Oleh karena itu, langkah pertama agar bisa dikomputasi secara analitis maka dilakukan estimasi ekspektasi menggunakan jumlah yang pasti dari sampel pada model. Sampel digunakan untuk mengestimasi gradien dari fase negatif yang direferensikan sebagai partikel negatif, dimana disimbolkan sebagai $\mathcal{N}$. Kemudian, gradien bisa ditulis sebagai (Tutorial, 2014) :

$$-\frac{\partial \log p(x)}{\partial \theta} \approx \frac{\partial \mathcal{F}(x)}{\partial \theta} - \frac{1}{|\mathcal{N}|} \sum_{\tilde{x} \in \mathcal{N}} \frac{\partial \mathcal{F}(\tilde{x})}{\partial \theta}. \tag{2.7}$$

Dimana secara ideal, elemen seperti $\tilde{x}$ dari $\mathcal{N}$ disampel menurut P (sebagai contoh adalah menggunakan teknik sampling Monte-Carlo). Dengan rumus diatas, secara praktis hampir bisa melakukan algoritma stochastic, hanya saja partikel negatif $\mathcal{N}$ belum bisa diekstraksi. Oleh karena itu, pada literatur dengan metode Markov Chain Monte Carlo, sangat bagus digunakan pada model Restricted Boltzmann Machine (RBM) yang merupakan bentuk spesifik dari model EBM (Tutorial, 2014).

2.6 Restricted Boltzmann Machine

Boltzmann Machines (BMs) adalah bentuk khusus dari log-linear Markov Random Field (MRF), dengan kata lain, dimana fungsi energi adalah linear pada parameter bebasnya. Agar membuat BM cukup bisa merepresentasikan distribusi yang kompleks(dengan kata lain, berangkat dari setting parameter yang terbatas kepada non paramter), diasumsikan bahwa beberapa variabel tidak terobserbasi sehingga disebut hidden. Dengan memiliki variabel hidden, bisa dilakukan peningkatan kapasitas model dari BM. RBM, selanjutnya membuat BM yang terbatas pada variabel tanpa koneksi visibel-visibel dan hidden-hidden. Seperti pada gambar 2.5 (Hinton and Salakhutdinov, 2006)



Gambar 2.5: Grafik yang Menggambarkan RBM

Fungsi energi $E(v,h)$ pada RBM didefinisikan sebagai persamaan 2.8.

$$E(v,h) = -b'v - c'h - h'Wv \tag{2.8}$$

Dimana W merepresentasikan bobot yang terkoneksi antara unit hidden dan visible dan b, c adalah bias dari visible dan hidden secara berurutan.

Hal ini bisa diterjemahkan dalam bentuk persamaan energi bebas $\mathcal{F}(v)$ seperti dibawah:

$$\mathcal{F}(v) = -b'v - \sum_i \log \sum_{h_i} e^{h_i(c_i + W_i v)}.$$

Dikarenakan struktur RBM yang spesifik, visibel dan hidden adalah independen secara bersyarat antara satu dengan lainnya. Dengan menggunakan sifat tersebut, maka dapat dituliskan :

$$p(h|v) = \prod_i p(h_i|v)$$

$$p(v|h) = \prod_j p(v_j|h).$$

2.6.1 RBMs yang Menggunakan Unit Biner

Kasus umum jika menggunakan unit biner (dimana v_j dan $h_i \in \{0,1\}$), yang didapat dari persamaan (6) dan (2), versi probabilistik dari fungsi aktivasi neuron adalah sebagai berikut(Hinton and Salakhutdinov, 2006):

$$P(h_i = 1|v) = sigm(c_i + W_i v) \tag{2.9}$$

$$P(v_j = 1|h) = sigm(b_j + W'_j h) \tag{2.10}$$

Selanjutnya, energi bebas dari RBM dengan unit biner, disederhanakan menjadi

persamaan:

$$\mathcal{F}(v) = -b'v - \sum_i \log(1 + e^{(c_i + W_i v)}). \tag{2.11}$$

2.6.2 Update Persamaan dengan Unit Biner

Menghubungkan persamaan (5) dengan (9), didapatkan gradien log-likelihood untuk RBM dengan unit biner sebagai berikut:

$$\begin{aligned} -\frac{\partial \log p(v)}{\partial W_{ij}} &= E_v[p(h_i|v) \cdot v_j] - v_j^{(i)} \cdot sigm(W_i \cdot v^{(i)} + c_i) \\ -\frac{\partial \log p(v)}{\partial c_i} &= E_v[p(h_i|v)] - sigm(W_i \cdot v^{(i)}) \\ -\frac{\partial \log p(v)}{\partial b_j} &= E_v[p(v_j|h)] - v_j^{(i)} \end{aligned} \tag{2.12}$$

2.7 Sampling pada RBM

Sampel dari $p(x)$ bisa didapat dengan menjalankan Markov chain sampai konvergen dengan menggunakan gibbs samping sebagai operator transisi.

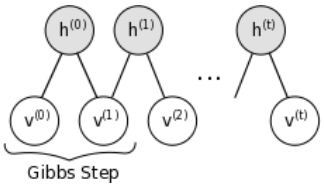
Gibbs sampling dari join variable random sebanyak N dari $S = (S_1, ..., S_N)$ merupakan urutan sebanyak N sampling dari sub-steps dalam bentuk $S_i \sim p(S_i|S_{-i})$ dimana S_{-i} berisi $N-1$ variabel random lain didalam S tetapi diluar S_i.

Untuk RBM, S berisi himpunan dari visible dan hidden unitnya. Akan tetapi, dikarenakan unit ini dipenden secara kondisional, maka salah satunya bisa dilakukan gibbs sampling. Pada setting disini, unit visible disampel secara simultan given nilai fix dari hidden unitnya. Demikian sebaliknya, hidden unitnya disampel secara simultan given unit visibelnya.Sehingga satu langkah Markov chain adalah sebagai berikut:

$$h^{(n+1)} \sim sigm(W'v^{(n)} + c)$$
$$v^{(n+1)} \sim sigm(Wh^{(n+1)} + b),$$

Dimana $h^{(n)}$ menunjik pada himpunan semua hidden unit pada nilai yang ke-n langkah dari Markov chain. Yang artinya adalah sebagai contoh, $h_i^{(n+1)}$ adalah secara random dipilih antara 1 (versus 0) dengan nilai probabilitas $sigm(W_i' v^{(n)} + c_i)$, demikian juga, $v_j^{(n+1)}$ adalah dipilih secara random antara 1 (versus 0) dengan probabilitas $sigm(W_{\cdot j} h^{(n+1)} + b_j)$.

Hal ini seperti digambarkan pada gambar 2.6



Gambar 2.6: Gibbs Sampling

Oleh karena $t \to \infty$, maka sampel $(v^{(t)}, h^{(t)})$ bisa dipastikan akan akurat dalam mensampel $p(v,h)$.

Secara teori, tiap parameter diupdate pada proses learning dibutuhkan satu rantai tersebut untuk konvergen. Akan tetapi hal ini sangat mahal komputasinya. Sehingga banyak diajukan algoritma untuk melatih RBM agar sampel $p(v,h)$ efisien, disaat proses learningnya.

2.8 Contrastive Divergence (CD-k)

Contrastive Divergence(CD) menggunakan trik untuk mempercepat proses sampling:

Contrastive Divergence uses two tricks to speed up the sampling process: Dikarenakan yang diinginkan adalah $p(v) \approx p_{train}(v)$ (distribusi data yang asli), initialisasi Markov chain dengan contoh data training (dimana, berasal dari distribusi yang mendekati p, pada distribusi final dari p). CD tidak menunggu rantai untuk konvergen. Sampel didapatkan setalah langkah ke-k dari Gibbs sampling. Pada prakteknya, $k = 1$ secara mengejutkan sudah menghasilkan hasil yang baik.

2.9 Persistent CD

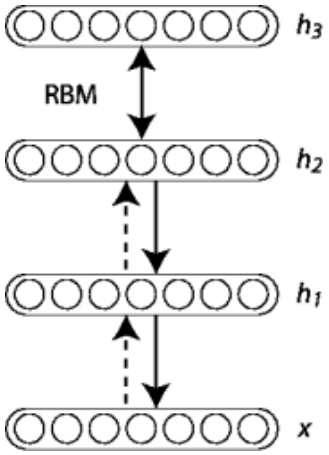
Persistent CD (P-CD) [Tieleman08] menggunakan pendekatan lain untuk mensampling $p(v,h)$. Hal ini bergantung hanya pada Markov chain tunggal, yang memiliki kondisi yang persisten (dimana, tidak melakukan restart chain pada setiap sampel yang terobservasi). Pada setiap umpdate parameter, akan di ekstraksi sampel baru dengan penjalankan chain pada langkah ke-k. Kondisi chain akan dipertahankan pada update selanjutnya.
Intuisinya adalah jika update parameternya cukup kecil dibaningkan dengan rate campuran dari Markov Chain, maka hal ini bisa mengejar perubahan modelnya.

2.10 Deep Believe Network

Hinton menunjukkan bahwa RBM bisa dijajar dan dilatih secara greedy untuk membentuk sebuah jaringan yang dinamakan dengan Deep Belief Network (DBN). DBN adalah model grafis dimana bisa melakukan learning untuk mengekstraksi representasi hirarki yang mendalam (deep) dari data training. Hal ini memodelkan distribusi gabungan antara vektor x sebagai observer dan ℓ layer hidden h^k sebagai berikut:

$$P(x,h^1,\ldots,h^\ell) = \left(\prod_{k=0}^{\ell-2} P(h^k|h^{k+1})\right) P(h^{\ell-1},h^\ell) \tag{2.13}$$

Dimana $x = h^0, P(h^{k-1}|h^k)$ adalah distribusi kondisional untuk unit visible dikondisikan pada unit hidden pada level k dan $P(h^{\ell-1},h^\ell)$ adalah distribusi gabungan visible-hidden pada level teratas dar RBM. Seperti diilustrasikan pada gambar 2.7.



Gambar 2.7: Arsitektur Deep Believe Network (DBN) yang merupakan gabungan dari RBM

Prinsip dari greedy layer-wise unsupervised training bisa di aplikasikan pada DBN dengan RBM sebagai bagian pada tiap layernya [hinton] [bengio]. Pada prinsipnya prosesnya adalah sebagai berikut:

1. Latih layer pertama ssebagai RBM yang memodelkan input $x = h^{(0)}$ sebagai visible layernya.

2. Gunakan layer pertama untuk mendapatkan representasi input yang digunakan sebagai data untuk layer kedua. Ada dua solusi yang sama. Representasi ini bisa dipilih sebagai rata-rata dari aktivasi $p(h^{(1)} = 1|h^{(0)})$ atau sampel dari $p(h^{(1)}|h^{(0)})$.

3. Train layer kedua sebagai RBM dengan mengambil data transformasi (sampel

atau rata-rata aktivasi) sebagai training (untuk layer visible dari RBM tersebut).

4. Iterasikan (2 dan 3) untuk semua layer yang diinginkan, setiap waktu dengan mempropagasikan keatas antara sampel atau nilai rata-ratanya.

5. Fine-tune semua parameter dari arsitektur dengan log-likelihood DBN atau dengan kriteria secara supervised setelah menambahkan layer supervised untuk memprediksikan kelas, sebagai contoh misalnya layer logistic regression.

Pada kasus ini, akan difokuskan pada fine-tuning dengan melakukan gradien descent menggunakan klassifier logistic regression dimana digunakan untuk mengklasifikasikan input x berdasar pada output dari hidden layer $h^{(l)}$ dari DBN. Fine-tune kemudian dilakukan melalui gradien descent dari NLL fungsi costnya. Dikarenakan gradien secara supervised adalah hanya non-null untuk bobot dan bias pada hidden layer pada tiap-tiap layer, maka prosedur ini serupa dengan menerapkan initialisasi parameter dari arsitektur MLP yang deep dengan bobot dan bias dari hidden layer yang didapat pada proses training unsupervised diatas.

2.11 Alasan Melakukan Training Secara Greedy Layer-Wise

Algoritma training deep learning secara greedy layer-wise terbukti bisa bekerja dengan baik, sebagai contoh 2 layer DBN dengan hidden layer $h^{(1)}$ dan $h^{(2)}$ dengan parameter bobot berurutan adalah $W^{(1)}$ dan $W^{(2)}$, (Hinton and Salakhutdinov, 2006) maka $\log p(x)$ bisa ditulis sebagai:

$$\begin{aligned}\log p(x) =& KL(Q(h^{(1)}|x)||p(h^{(1)}|x)) + H_{Q(h^{(1)}|x)} + \\ & \sum_h Q(h^{(1)}|x)(\log p(h^{(1)}) + \log p(x|h^{(1)})).\end{aligned} \tag{2.14}$$

$KL(Q(h^{(1)}|x)||p(h^{(1)}|x))$ merepresentasikan KL divergence antara posterior $Q(h^{(1)}|x)$ dari RBM pertama jika hal ini sendirian, dan probabilitas $p(h^{(1)}|x)$ untuk layer sayng sama tapi didifinisikan oleh keseluruhan DBN (sebagai contoh, perhitungan prior $p(h^{(1)},h^{(2)})$ didefinisikan sebagai top-level RBM). $H_{Q(h^{(1)}|x)}$ adalah entropy dari distribusi $Q(h^{(1)}|x)$.
Hal ini bisa ditunjukkan bahwa jika diinitialisasi kedua layer hidden sehingga $W^{(2)} = {W^{(1)}}^T, Q(h^{(1)}|x) = p(h^{(1)}|x)$ dan KL divergence nya adalah null. Maka jika di lakukan learning pada level awal RBM dan kemudian parameter $W^{(1)}$ dibuat tetap, kemudian dilakukan optimasi pada persamaan 2.14 terhadap $W^{(2)}$ bisa

meningkatkan likelihood dari $p(x)$. Jika diisolasi hanya pada $W^{(2)}$ sehinggi didapatkan:

$$\sum_h Q(h^{(1)}|x)p(h^{(1)})$$

Melakukan optimasi persamaan ini dengan memperhatikan jumlah $W^{(2)}$ training pada tingkat RBM selanjutnya, menggunakan output dari $Q(h^{(1)}|x)$ sebagai distribusi training untuk RBM yang pertama.

2.12 Logistic Regression

Logistic Regression adalah salah satu klassifier yang paling dasar pembentuk dari MLP. Penjelasannya akan dimulai dari bentuk model dasarnya serta notasi matematisnya.

2.12.1 Model Logistic Regression

Logistic regression adalah klasifier yang linear dan probabilistik. Diparameterkan dengan matrik bobot W dan vektor bias b. Proses klasifikasinya adalah dengan cara memproyeksikan vektor input kedalam himpunan *hyperplane*, dimana berkorespondensi pada kelasnya. Jarak dari input ke *hyperplane* merefleksikan probabilatas dari input adalah berkorespondensi dari anggota kelasnya.
Secara matematis, probabilitas vektor input x adalah anggota dari kelas i, isi dari variabel *stochastic* Y, bisa ditulis sebagai berikut:

$$\begin{aligned} P(Y=i|x,W,b) &= softmax_i(Wx+b) \\ &= \frac{e^{W_i x+b_i}}{\sum_j e^{W_j x+b_j}} \end{aligned} \tag{2.15}$$

Prediksi dari model berupa y_{pred} adalah kelas dimana probabilitasnya maksimal, secara spesifik ditulis sebagai:

$$y_{pred} = \text{argmax}_i P(Y=i|x,W,b) \tag{2.16}$$

2.12.2 Mendefinisikan Lost Function dari Logistic Regression

Melakukang *learning* pamameter model dengan cara meminimalisasi *Lost Function*. Pada kasus *logistic regression* yang multi-kelas, sangat umum digunakan minimisasi *negative log likelihood (NLL)* yang ekivalen dengan memaksimalkan

likelihood dari data set $\mathcal{D}$ pada model yang diparameterkan oleh θ. Definisi dari likelihood $\mathcal{L}$ dan loss ℓ maka:

$$\mathcal{L}(\theta = \{W, b\}, \mathcal{D}) = \sum_{i=0}^{|\mathcal{D}|} \log(P(Y = y^{(i)}|x^{(i)}, W, b)) \\ \ell(\theta = \{W, b\}, \mathcal{D}) = -\mathcal{L}(\theta = \{W, b\}, \mathcal{D}) \tag{2.17}$$

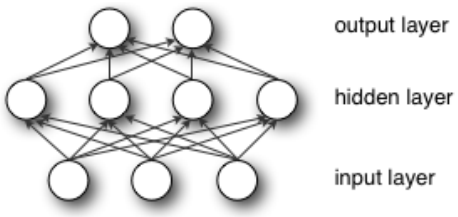
Untuk meminimisasi, digunakan *stochastic gradient descen with minibatches (MSGD)* (Tutorial, 2014).

2.13 Multi Layer Perceptron

Arsitektur selanjutnya yang akan dibahas adalah *Multi Layer Perceptron (MLP)* Arsitektur MLP ini bisa dilihat sebagai klasifier *Logistic Regression* dimana input pada awalnya ditransformasikan menggunakan transformasi non linear Φ. Transformasi ini memproyeksikan data input kepada *space* dimana hal ini bisa terseparasi secara linear. Layer tengah ini direferensikan sebagai *hidden layer*. Satu hidden layer sebenarnya sudah cukup untuk membuat MLP sebagai aproksimator universal. Akan tetapi, ada banyak keuntungan untuk menggunakan hidden unit yang lebih dari satu layer, hal inilah yang digunakan sebagai konsep dasar dari deep learning. Algoritma untuk melakukan *training* dari MLP yang paling sering dipakai adalah algoritma *back-propagation* (Tutorial, 2014).

2.13.1 Model MLP

MLP atau sering disebut juga dengan Artificial Neural Network (ANN) adalah Perceptron yang dibentuk menjadi sebuah jaringan. MLP dengan layer tunggal bisa direpresentasikan secara grafis seperti pada Gambar 2.8 berikut.



Gambar 2.8: Arsitektur Layer Tunggal MLP

Secara formal, hidden layer tunggal dari MLP adalah sebuah fungsi $f : R^D \rightarrow R^L$, dimana D aadlah ukuran dari vektor input x dan L adalah ukuran dari output vektor

$f(x)$ sehingga dengan menggunakan notasi matriks sebagai berikut :

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{(1)}x))), \tag{2.18}$$

Dengan vektor bias $b^{(1)}, b^{(2)}$; dan matrik bobot $W^{(1)}, W^{(2)}$ dan fungsi aktivasinya adalah G dan s. Sedangkan vektor $h(x) = \Phi(x) = s(b^{(1)} + W^{(1)}x)$ merupakan *hidden layer*. Setiap kolom $W^{(1)}_{.i}$ merepresentasikan bobot dari unit input yang ke-*i* dari *hidden unit*. Pilihan fungsi aktifasinya bisa menggunakan tanh, atau fungsi sigmoid.

$$\begin{aligned} tanh(a) &= \frac{(e^a - e^{-a})}{(e^a + e^{-a})} \\ sigm(a) &= \frac{1}{(1 + e^{-a})} \end{aligned} \tag{2.19}$$

Kedua fungsi aktivasi yaitu tanh dan sigmoid adalah fungsi skalar ke skalar akan tetapi bisa diekstensikan menjadi vektor atau tensor yang diaplikasikan secara *element wise*.

Vektor output didapatkan dengan: $o(x) = G(b^{(2)} + W^{(2)}h(x))$. Probabilitas dari keanggotaan kelas didapat dari memilih G sebagai fungsi *softmax* (untuk kasus klasifikasi multi-kelas).

Untuk melakukan *training* MLP dilakukan *learning* parameter dari model menggunakan *Stochastic Gradien Descent* dengan dibagi menjadi bagian kecil-kecil atau disebut dengan *minibatch*. Himpunan parameter pembelajarkan ditulis sebagai himpungan $\theta = \{W^{(2)}, b^{(2)}, W^{(1)}, b^{(1)}\}$. Mendapatkan gradien $\partial\ell/\partial\theta$ didapatkan dengan menerapkan algoritma *backpropagation* (Tutorial, 2014)

BAB 3
METODOLOGI PENELITIAN

Penelitian ini dibagi menjadi empat tahap: (1) Mendapatkan data microarray dan pengolahan awal; (2) Perancangan algoritma; (3) Melakukan eksperimen untuk mendapatkan *hyperparameter* yang optimal. Kemudian dilanjutkan dengan testing dan evaluasi. Gambaran umum dari penelitian ini seperti pada Gambar 3.1

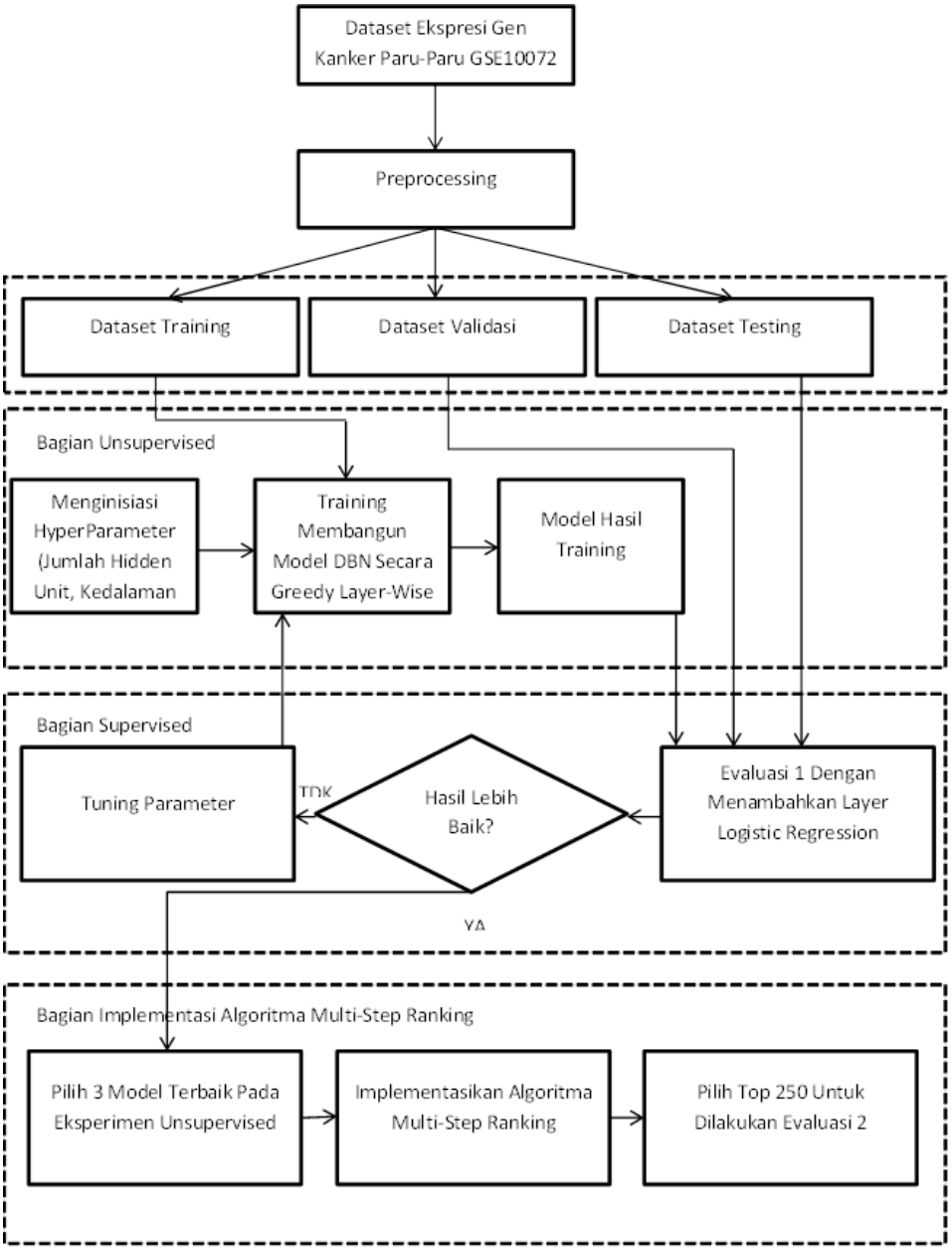
3.1 Gambaran Umum Penelitian

Secara garis besar, penelitian ini dibagi menjadi beberapa tahapan. Yang pertama adalah tahapan persiapan yaitu mendapatkan data *microarray* kemudian mengolahnya menjadi data yang siap untuk dilakukan proses seleksi fitur dan tahapan pelatihan *deep learning*. Yaitu dengan membagi 80% data untuk training, 15% data untuk validasi dan 5% data untuk testing.

Bagian kedua adalah membangun model DBN dengan teknik *unsupervised learning*. Untuk mendapatkan model terbaik secara *greedy* pada tiap-tiap layer RBM-nya. Dimana dilakuan tuning *hyperparameter* (jumlah kedalaman layer, jumlah *hidden unit* pada tiap layernya) digunakan untuk mendapatkan struktur *hyperparameter* yang cocok dengan ciri khas dari data *microarray*. Oleh karena itu diperlukan banyak percobaan untuk mendapatkan hasil yang bagus.

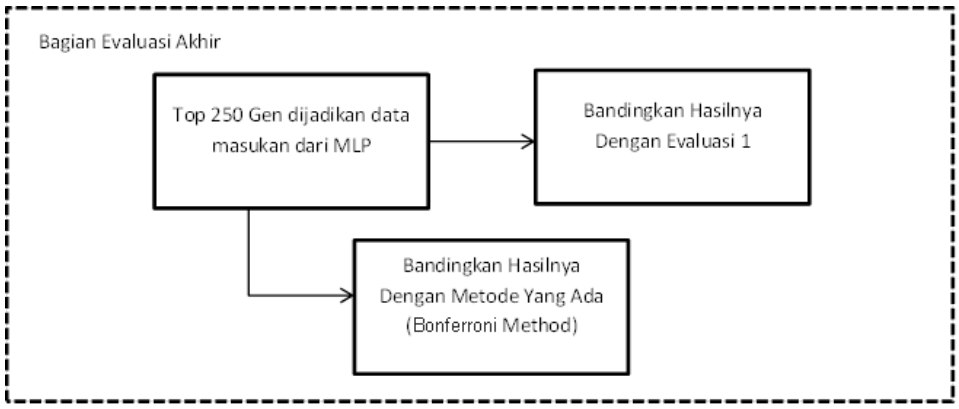
Bagian ketiga, adalah *supervised learning*, dimana merupakan evaluasi sementara dari tahap yang kedua. Dibuat layer output berupa *logistic regression*, yang digunakan untuk menguji sementara hasil dari proses *pretraining* untuk mengklasifikasikan pasien kanker dan pasien normal menggunakan dataset validasi dan dataset testing.

Bagian keempat merupakan bagian yang terpenting karena dimana ide thesis ini dibuat. Yaitu melakukan perankingan gen untuk mencari gen yang paling informatif yang didapatkan dari model pada percobaan sebelumnya. Dimana algoritma seleksi fitur untuk multi-step ranking dijalankan agar didapatkan *biomarker*.



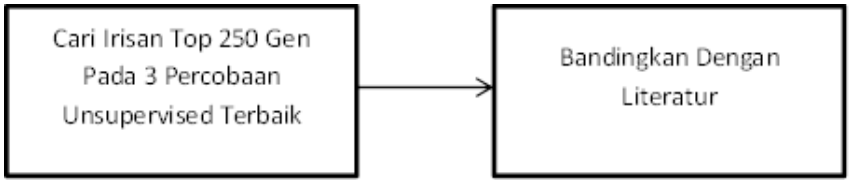
Gambar 3.1: Overview Penelitian

Tahapan terakhir adalah tahap evaluasi akhir, yaitu akan dilakukan dua kali evaluasi, yang pertama evaluasi dengan cara membandingkan evaluasi 1 (*logistic regression* sebelum dilakukan seleksi fitur) dengan evaluasi 2 (MLP setelah dilakukan seleksi fitur). Hasil dari kedua proses ini dibandingkan apakah terjadi perbaikan performa klasifikasinya.



Gambar 3.2: Overview Metode Evaluasi

Untuk evaluasi selanjutnya yaitu dilakukan konfirmasi, dimana hasil dari perankingan gen tersebut dibandingkan dengan penelitian tentang biomarker sebelumnya. Apakah gen biomarker yang ditemukan pada penelitian ini memiliki signifikansi dibandingkan dengan teknik sebelumnya.



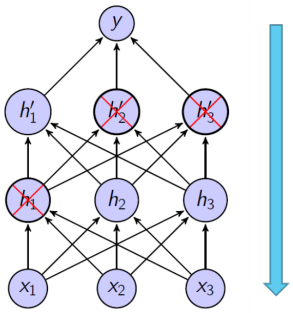
Gambar 3.3: Metode Untuk Mengkonfirmasi Biomarker

3.2 Desain Metode Perangkingan Bobot Secara Multi Step Untuk Mendapatkan Gen Biomarker

Pada penelitian ini, akan dibangun sebuah teknik pencarian *Biomarker* dengan metode seleksi fitur gen. Metode ini menerapkan perankingan gen secara *multi step* terhadap model yang didapatkan pada proses *training* yang dilakukan secara *unsupervised*. Arsitektur untuk mendapatkan modelnya adalah digunakan arsitektur *Deep Belief Network (DBN)* yang merupakan bagian dari metode *deep learning*. Metode perankingan yang digunakan adalah modifikasi dari algoritma seleksi fitur untuk *logistic regression* yang dilakukan oleh Shevade and Keerthi (2003). Akan tetapi metode ini memiliki masalah dalam mengeliminasi fitur jika diterapkan secara langsung pada model DBN, dikarenakan parameter bobot (W) dan bias (b) ditempatkan disetiap fitur dan model ini hanya memiliki satu layer dibandingkan dengan DBN yang memiliki banyak layer.

Pada DBN, *hidden unit* yang paling sering aktif adalah *hidden unit* yang lebih

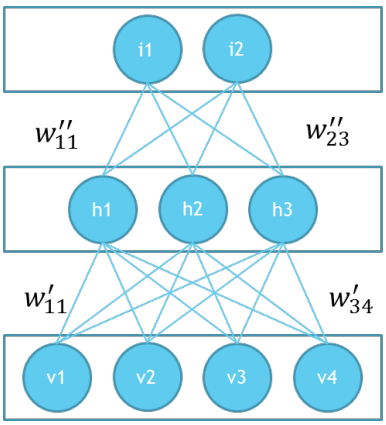
penting dibandingkan dengan unit yang jarang aktif, oleh karena itu *hidden unit* ini memiliki parameter bobot yang lebih besar dibandingkan dengan *hidden unit* yang jarang aktif pada saat proses *training* dilakukan. Pemilihan fitur dilakukan dengan meranking unit-unit yang memiliki bobot tertinggi dimulai dari *layer output* mundur secara multi-step menuju *layer input* untuk mendapatkan fitur gen yang paling berpengaruh terhadap model. Kemudian dilakukan eliminasi bobot pada *hidden unit* per layernya secara *multi step*. Selanjutnya akan dipilih sebanyak *top-n* gen dari hasil perankingan ini untuk dievaluasi apakah *Biomarker* yang ditemukan tersebut informatif atau tidak. Seperti digambarkan pada bagan Gambar 3.4



Gambar 3.4: Hidden unit yang paling sering aktif adalah neuron yang paling penting. Sedangkan yang Kurang Penting Dihapus dengan arah mundur Secara Multi-step (Duh, 2014)

3.2.1 Perhitungan Seleksi Fitur dengan Multi-Step Ranking

Contoh dibawah adalah contoh penyederhanaan dari proses multi-step ranking yang diajukan. Pada prakteknya, *visible unit* dan *hidden unit* memiliki jumlah yang besar. Sebagai contoh, pada kasus data kanker paru-paru yang diteliti ini memiliki fitur 22 ribu gen yang di ukur secara simultan dalam satu percobaan.



Step 1 : Perhitungannya mulai dari layer output (i), mundur sampai layer visible (v) :

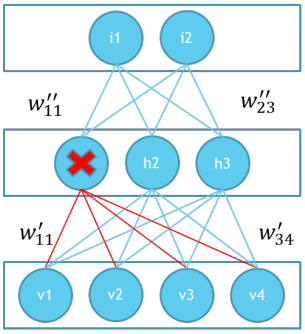
$$i_1 = w''_{11} h_1 + w''_{12} h_2 + w''_{13} h_3$$
$$i_2 = w''_{21} h_1 + w''_{22} h_2 + w''_{23} h_3$$

Ranking :

$$h_1 = w''_{11} + w''_{21}$$
$$h_2 = w''_{12} + w''_{22}$$
$$h_3 = w''_{13} + w''_{23}$$

Kemudian di ranking, misalnya didapatkan h2 dan h3 berturut-turut ranking 1 dan 2, maka h1 dieliminasi, dan bobot h1 dieliminasi untuk proses selanjutnya.

Gambar 3.5: Contoh Perhitungan tahap pertama dimulai dari top hidden unit



Eliminasi Bobot Dengan Ranking Terendah :

$$h_1 = w'_{11}v_1 + w'_{12}v_2 + w'_{13}v_3 + w'_{14}v_4$$
$$h_2 = w'_{21}v_1 + w'_{22}v_2 + w'_{23}v_3 + w'_{24}v_4$$
$$h_3 = w'_{31}v_1 + w'_{32}v_2 + w'_{33}v_3 + w'_{34}v_4$$

Lakukan sampai mencapai visible layer terakhir :

$$v_1 = w'_{11} + w'_{21} + w'_{31}$$
$$v_2 = w'_{12} + w'_{22} + w'_{32}$$
$$v_3 = w'_{13} + w'_{23} + w'_{33}$$
$$v_4 = w'_{14} + w'_{24} + w'_{34}$$

Ranking v untuk mendapatkan biomarker

Gambar 3.6: Contoh Perhitungan tahap pertama dimulai dari top hidden unit

Perhitungan diatas secara iteratif dilakukan mulai dari layer output mundur sampai layer input.

3.3 Implementasi Metode Perangkingan Bobot Secara Multi Step Untuk Mendapatkan Gen Biomarker

Implementasi multi-step ranking dengan menggunakan python:
Listing 3.4 : Implementasi Multi-Step Ranking di python

```
# perkalian matrix rank weight
import numpy as np

def awal(w):
    return np.ones((w.shape[1],), dtype=np.float)

def jumlah_bobot(w, top_ke_n):
    # kalikan w dengan matrix 1
    return w.dot(top_ke_n)

def rank_hasil_jumlah(sum_w):
    # urutkan sum_w dan beri index
    """

    : rtype sum_w : numpy.array
    """
    swi = sum_w.shape[0]
    hsl = np.arange(swi)
    c = np.concatenate((hsl,sum_w))
    c = c.reshape(2,swi)
    c = c.T
    z = c[c[:,1].argsort()[::-1]] # urutkan descending berdasarkan bobot (indeks mengikuti)
    return z

def set_top_n(idx_sum_w, top_n = 2):
    # set = 0 semua yang bukan top n
    # kembalikan ke urutan semula
    z = idx_sum_w.copy()
    z[top_n:,1] = 0.
```

```
    z[0:top_n,1]= 1.
    # print 'z adalah'
    # print z
    d = z[z[:,0].argsort()[::]]
    # print 'd adalah'
    # print d
    return d

# set_rank : melakukan setting 1 untuk top n dan
def extract_top_n(n):
    return n[:,1]

def set_index_dengan_gen(bobot_akhir):
    # index gen dengan urutan perankingannya
    pass

def plot_diagram(a, b):
    # plot himpunan a dan b dan anggota keduanya
    pass


if __name__ == '__main__':
    # w1 adalah bobot untuk testing
    w1 = np.array([[0, 1, 2, 3, 4],
                   [5, 6, 7, 8, 9],
                   [10, 11, 12, 13, 14]])
    a = awal(w1)
    x = jumlah_bobot(w1,a)   # x = perhitungan bobot berdasarkan h ( 10, 35, 60)
    y = rank_hasil_jumlah(x) # (diberi index dan diranking)
    z = set_top_n(y,1)
    print y
    # print x.shape
    # print y  # matrix penjumlahan bobot diranking sebelum diambil top N
    # print z  # matrix penjumlahan bobot setelah diranking dan diset 0 untuk yg bukan top N
    # print extract_top_n(z)
```

Contoh implementasi multistep rank pada model yang disimpan pada file:

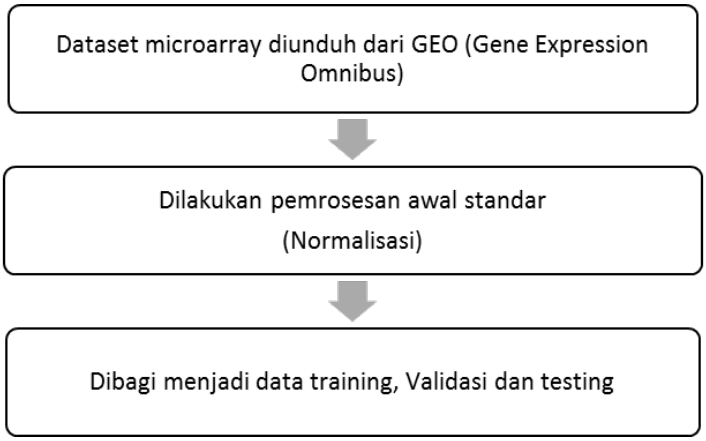
Listing 3.5: Implementasi Multistep rank Pada Model

```
import multistep_rank as mtr
import theano.tensor as T
import numpy as np
from ekstrak_csv import Ekstraktor


# buat function :
# hsl_ranking = multisteprank(model, [100,100,100]):

ekstraktor = Ekstraktor()

model =    ekstraktor.load_data("./dataset/model1000e_10k_5k_1k_500.pkl.gz")
print 'Jumlah_layer_:_%i' % (model.n_layers)

Wlayer3 = model.rbm_layers[3].W
Wlayer2 = model.rbm_layers[2].W
Wlayer1 = model.rbm_layers[1].W
Wlayer0 = model.rbm_layers[0].W
# Wlayer1.shape.eval()

y3 = Wlayer3.get_value(True)
x3 = T.fmatrix()
x3 = y3.copy()

# ranking ujung
awal3 = mtr.awal(x3)
jml_bobot3 = mtr.jumlah_bobot(x3, awal3)
ranking_jml_bobot3 = mtr.rank_hasil_jumlah(jml_bobot3)
top_n3 = mtr.set_top_n(ranking_jml_bobot3,70)

# print "layer 3"
# print 'hasil perankingan top 50: '
# print ranking_jml_bobot3[:50]
# print 'set top n dengan 1 : '
# print top_n3.astype(int)
```

```
y2 = Wlayer2.get_value(True)
x2 = y2.copy()
awal2 = mtr.extract_top_n(top_n3)
jml_bobot2 = mtr.jumlah_bobot(x2, awal2)
ranking_jml_bobot2 = mtr.rank_hasil_jumlah(jml_bobot2)
top_n2 = mtr.set_top_n(ranking_jml_bobot2,700)

# print "layer 2"
# print 'hasil perankingan top 50: '
# print ranking_jml_bobot2[:50]
# print 'set top n dengan 1 : '
# print top_n2.astype(int)

y1 = Wlayer1.get_value(True)
x1 = y1.copy()
awal1 = mtr.extract_top_n(top_n2)
jml_bobot1 = mtr.jumlah_bobot(x1, awal1)
ranking_jml_bobot1 = mtr.rank_hasil_jumlah(jml_bobot1)
top_n1 = mtr.set_top_n(ranking_jml_bobot1,1500)

# print "layer 1"
# print 'hasil perankingan top 50: '
# print ranking_jml_bobot1[:50]
# print 'set top n dengan 1 : '
# print top_n1.astype(int)

y0 = Wlayer0.get_value(True)
x0 = y0.copy()
awal0 = mtr.extract_top_n(top_n1)
jml_bobot0 = mtr.jumlah_bobot(x0, awal0)
ranking_jml_bobot0 = mtr.rank_hasil_jumlah(jml_bobot0)
top_n0 = mtr.set_top_n(ranking_jml_bobot0,70)

print "layer visible"
print 'hasil perankingan top 250 layer visible 10k 5k 1k 500: '
print ranking_jml_bobot0[:250,0].astype(int)
```

3.4 Pengumpulan Data dan Pengolahan Awal

Data microarray tersedia secara bebas di geo [http://www.ncbi.nlm.nih.gov/geo/], dan dapat diunduh, untuk digunakan sebagai data penelitian. Kemudian dilakukan normalisasi standar yang sering di pakai pada data microarray, proses normalisasi ada banyak metode, dan akan digunakan satu metode standar untuk pengolahan awal microarray agar mendapatkan data konsisten dan dapat dibandingkan. Proses pengolahan awal dan normalisasi digunakan tools standar dan tersedia bebas yaitu R-Bioconductor.



Gambar 3.7: Proses Pengumpulan data dan Pengolahan Awal

3.4.1 Implementasi dengan R-Bioconductor

Implementasi preprocessing dengan menggunakan R-Bioconductor

3.5 Data Profil Gen Percobaan Microarray dan Biomarker

Definisi *Biomarker* adalah sesuatu penanda yang bisa digunakan sebagai indikator suatu penyakit dari pasien. [http://www.biomarker.co.uk/whatisabiomarkers.html] Sebagai contoh, untuk mendiagnosa kanker paru-paru, hanya dibutuhkan 26 ekspresi gen saja. Gen yang paling informatif ini disebut dengan Biomarker (Bing, 2006). Pada profil gen GSE10072 yang merupakan kanker paru-paru, menurut (Belinsky, 2004) ada 26 gen yang paling berpengaruh dari 22.283 gen yang diteliti secara bersamaan, seperti ditunjukkan pada Gambar 3.8 yang merupakan contoh dari *biomarker* kanker paru-paru.

Probe ID	Gene	Chromosomal	Current/Never† N = 30		Former/Never N = 23		Tumor/Non-Tumor N = 36	
	Symbol	Location	Fold-change	p-value	Fold-change	p-value	Fold-change	p-value
204641_at	NEK2*	1q32.2–q41	3.45	0.0001	2.84	0.0036	3.14	<0.0001
204822_at	TTK*	6q13–q21	3.27	<0.0001	2.08	0.0123	2.22	<0.0001
218009_s_at	PRC1*	15q26.1	2.99	0.0007	2.61	0.0109	2.60	<0.0001
207828_s_at	CENPF*	1q32–q41	2.88	<0.0001	2.28	0.0034	2.77	<0.0001
202095_s_at	BIRC5*	17q25	2.72	0.0002	2.10	0.0145	2.55	<0.0001
203362_s_at	MAD2L1	4q27	2.67	0.0003	1.93	0.0309	2.74	<0.0001
219918_s_at	ASPM	1q31	2.59	0.0008	2.12	0.0218	2.87	<0.0001
210559_s_at	CDC2	10q21.1	2.54	0.0009	2.02	0.0298	2.37	<0.0001
201897_s_at	CKS1B	1q21.2	2.36	0.0002	1.89	0.0152	2.47	<0.0001
204170_s_at	CKS2	9q22	2.36	0.0006	2.02	0.0148	1.69	0.0015
222077_s_at	RACGAP1*	12q13.12	2.35	0.0003	1.91	0.0178	2.13	<0.0001
203214_x_at	CDC2	10q21.1	2.29	0.0006	1.98	0.0150	2.12	<0.0001
219306_at	KIF15*	3p21.31	2.22	0.0002	2.00	0.0047	1.90	0·0001
209642_at	BUB1*	2q14	2.17	0.0009	1.68	0.0507	2.02	0.0001
210052_s_at	TPX2*	20q11.2	2.06	0.0006	1.87	0.0100	2.07	<0.0001
203418_at	CCNA2	4q25–q31	1.99	<0.0001	1.85	0.0012	1.82	<0.0001
212020_s_at	MKI67	10q25-qter	1.95	<0.0001	1.71	0.0016	1.41	0.0006
201088_at	KPNA2	17q23.1–q23.3	1.82	<0.0001	1.53	0.0079	2.34	<0.0001
211519_s_at	KIF2C*	1p34.1	1.78	0.0004	1.67	0.0062	1.51	0.0002
218252_at	CKAP2	13q14	1.75	0.0008	1.52	0.0292	1.47	0.0001
204887_s_at	PLK4	4q27–q28	1.74	0.0001	1.55	0.0066	1.48	<0.0001
211080_s_at	NEK2*	1q32.2–q41	1.57	0.0001	1.50	0.0019	1.36	0.0002
214894_x_at	MACF1	1p32–p31	0.65	0.0003	0.64	0.0016	0.52	<0.0001
208634_s_at	MACF1	1p32–p31	0.60	0.0001	0.58	0.0004	0.42	<.0.0001
202284_s_at	CDKN1A	6p21.2	0.54	0.0003	0.70	0.0668	0.65	0.0082
208893_s_at	DUSP6	12q22–q23	0.34	0.0003	0.32	0.0012	0.84	0.3102

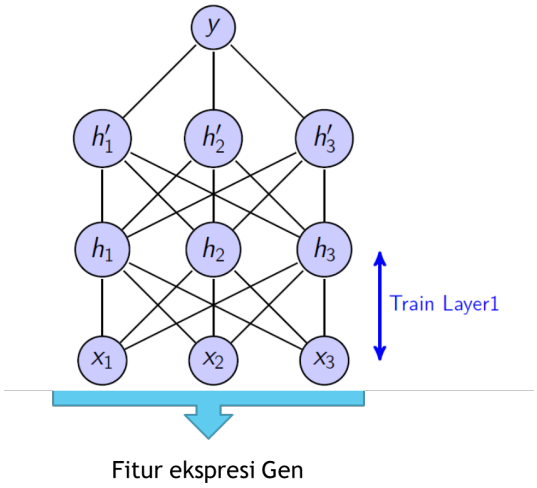
†Probe selection restricted to estimates with p<0.001 and fold-change >1.5 or <0.6667, and within the most inclusive category of genes with p≤0.001 in the GoMiner analysis (GO ID 7049, Appendix S2D).
*Genes involved in the mitotic spindle formation. The double line separates up-regulated and down-regulated probes.
doi:10.1371/journal.pone.0001651.t002

Gambar 3.8: Contoh 26 Gen Biomarker Kanker Paru-paru GSE10072

3.6 Perancangan Metodologi Penelitian

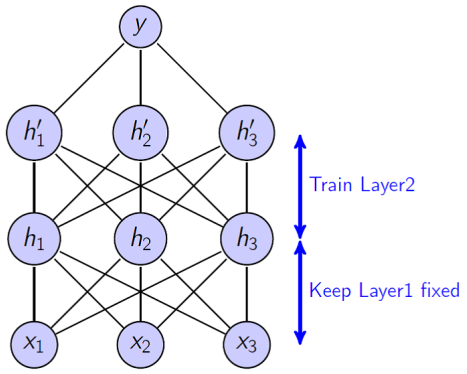
3.6.1 Tahapan Unsupervised

Tahap *unsupervised* adalah tahapan dimana model DBN ditraining secara *unsupervised* dengan data training pada tiap-tiap layernya secara *greedy*, artinya, proses pelatihan dilakukan secara berjenjang mulai dari layer visibel dengan hidden layer 0 dan kemudian layer ini bobotnya dibuat tetap dan digunakan sebagai input pada leyer berikutnya. Tiap layernya dihitung cost untuk kemudian diminimisasi errornya. Konsep ini disebut *greedy layer-wise training* yaitu setiap layer di traning secara independen dan satu-satu mulai dari layer input yang merupakan data ekspresi gen yang sudah disesuaikan dan dinormalisasi sampai layer output. Seperti pada Gambar 3.9



Gambar 3.9: Greedy layer-wise training pada layer visible dan hidden pertama

Setelah layer pertama selesai di training, layer pertama dibuat *fixed* dan dipakai sebagai inputan visible dari layer selanjutnya. Demikian selanjutnya sampai layer terakhir yaitu layer output. Seperti pada Gambar 3.10



Gambar 3.10: Greedy layer-wise training pada selanjutnya, yaitu dengan membuat layer sebelumnya Fixed

Pada tahapan training secara unsupervised ini dihitung cost function antara error konstruksi dibandingkan dengan error rekonstruksinya. Dalam RBM yaitu error konstruksi atau disebut error fase positif dibandingkan dengan error rekonstruksi atau error fase negatif.
log-likelihood $\mathcal{L}(\theta, \mathcal{D})$ dan fungsi loss-nya sebagai NLL $\ell(\theta, \mathcal{D})$ sebagai berikut:

$$\begin{aligned} \mathcal{L}(\theta, \mathcal{D}) &= \frac{1}{N} \sum_{x^{(i)} \in \mathcal{D}} \log\ p(x^{(i)}) \\ \ell(\theta, \mathcal{D}) &= -\mathcal{L}(\theta, \mathcal{D}) \end{aligned} \tag{3.1}$$

Menggunakan stochastic gradient $-\frac{\partial \log p(x^{(i)})}{\partial \theta}$, dimana θ adalah parameter dari

modelnya.
Loss function merupakan negative log-likelihood dari log-likelihood model. Data dari gradien NLL kemudian memiliki bentuk yaitu:

$$-\frac{\partial \log p(x)}{\partial \theta} = \frac{\partial \mathcal{F}(x)}{\partial \theta} - \sum_{\tilde{x}} p(\tilde{x}) \frac{\partial \mathcal{F}(\tilde{x})}{\partial \theta}. \tag{3.2}$$

3.6.2 Tahapan Supervised

Pada saat *training* secara *unsupervised* dilakukan, diukur *cost* yang menunjukkan perbedaan antara konstruksi dan rekonstruksi pada tiap layernyanya. Akan tetapi, hal ini hanya untuk mengetahui *cost* tiap-tiap layer RBM-nya, bukan seberapa baik model dalam melakukan klasifikasi. Oleh karena itu diperlukan satu layer output yang yang berupa *logistic regression* untuk mengetahui seberapa baik model dalam membedakan kelas kanker dan bukan kanker.

3.6.2.1 Implementasi Logistic Regression pada Layer Output

Logistic regression adalah klasifier linear yang memiliki matriks bobot W dan vektor bias b. Klasifikasi merupakan proyeksi titik data pada sebuah himpunan *hyperplane* yang jaraknya digunakan sebagai penentu probabilitas keanggotaan kelasnya. Secara matematis bisa dituliskan sebagai:

$$\begin{aligned} P(Y = i|x, W, b) &= softmax_i(Wx + b) \\ &= \frac{e^{W_i x + b_i}}{\sum_j e^{W_j x + b_j}} \end{aligned} \tag{3.3}$$

Output dari model akan memprediskikan dengan menghitung *argmax* dari vektor dimana elemen ke i adalah $P(Y = i|x)$.

$$y_{pred} = argmax_i P(Y = i|x, W, b) \tag{3.4}$$

Implementasinya menggunakan optimisasi stochastic gradient descent. Untuk implementasi lengkapnya ada di lampiran.

3.6.3 Tahapan Tuning Parameter

Parameter yang akan dilakukan *tuning* disini adalah: jumlah hidden units, jumlah banyaknya layer hidden dan banyaknya epoch. Tuning parameter dilakukan agar bisa didapatkan hasil yang optimum dari percobaan yang dilakukan. Tahap ini

adalah tahap yang paling krusial untuk mendapatkan hasil yang diinginkan. Dikarenakan uniknya data microarray, maka dilakukan *trial and error* dari parameter-parameternya.
Proses tuning parameter ini memerlukan waktu yang lama karena setiap percobaan memiliki parameter yang diubah-ubah untuk menyesuaikan hasil yang diinginkan. Dikarenakan sifat dari microarray yang berbeda dengan citra yang sudah banyak dilakukan oleh peneliti, tuning parameter untuk data *microarray* pada arsitektur deep learning jarang dilakukan oleh peneliti, sehingga proses tuning dilakukan setiap selesai dilakukan percobaan yang memerlukan waktu antara 2 hari sampai 5 hari, tergantung dari epoch dan jumlah layer dan hidden unitnya.
Proses training pada arsitektur *deep learning* juga memerlukan kekuatan komputasi komputer yang kuat dan memory yang relatif lebih besar untuk mendapatkan model yang optimal.

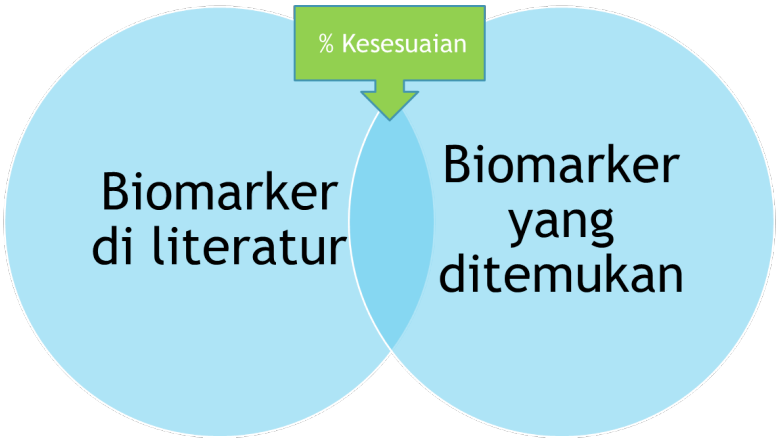
3.7 Melakukan Testing Arsitektur DBN

Hasil dari unsupervised learning yang dilakukan oleh DBN, akan diuji dahulu dengan dengan data testing, apakah error rekonstruksinya lebih baik. Setelah dilakukan perankingan *biomarker*, diperlukan pengujian apakah apakah seleksi fitur tersebut menggambarkan hasil yang diinginkan, dengan membandingkan biomarker yang dihasilkan dengan literature.

3.8 Evaluasi Hasil Perangkingan Dengan Klasifikasi Secara Supervised Menggunakan MLP

Proses evaluasi dilakukan dua kali, pertama, saat menggunakan data asli dan tidak dilakukan seleksi fitur, yang kedua setelah dilakukan seleksi fitur. Hal ini dilakukan untuk mengetahui apakah seleksi fitur tersebut bisa memperbaiki hasil klasifikasi secara signifikan dibandingkan tanpa dilakukan seleksi fitur.
Evaluasi hasil hasil perankingan secara *supervised* diperlukan untuk mengetahui apakah hasil perankingan tersebut memperbaiki hasil klasifikasi pasien kanker dan sehat hanya dengan menggunakan gen-gen yang dipilih berdasarkan ranking yang didapatkan.

3.9 Perbandingan Hasil Perangkingan Dengan Literatur



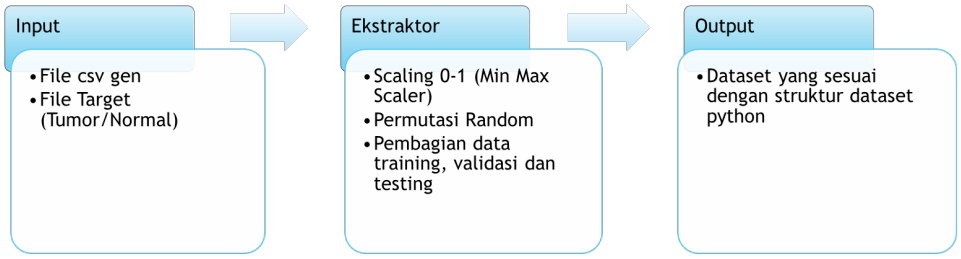
Gambar 3.11: Persen Kesesuaian Antara Biomarker yang Ditemukan dibandingkan dengan Biomarker di Literatur

Hasil perankingan pada percobaan tersebut selanjutnya diteliti apakah gen hasil perankingan tersebut adalah gen yang memiliki signifikansi terhadap penyakit yang diinginkan. Dalam kasus ini yaitu penyakit kanker paru-paru. Berikut adalah contoh 26 gen biomarker pada percobaan GSE10072 yang disitasi dari paper.

3.10 Modul-modul Pendukung

3.10.1 Kelas Ekstraktor

Untuk melakukan pengolahan pengolahan awal, didevelop sebuat kelas yang bernama kelas Ekstraktor. Kelas ini berfungsi untuk mengekstrak file csv dari data gen, menjadi file yang memiliki struktur data yang sesuai dengan library dbn.py di python. Hal ini dilakukan agar datanya memiliki struktur yang sesuai dengan dbn yaitu dilakukan normalisasi data profil gen yang berbentuk ekspresi gen menjadi rentang antara 0 sampai 1.



Gambar 3.12: Kelas Ekstraktor, Untuk melakukan Ekstraksi data Gen

3.10.2 Implementasi Kelas Ekstraktor di Python

Listing 4.1: Ekstraksi dataset untuk disesuaikan dengan struktur data modul dbn.py

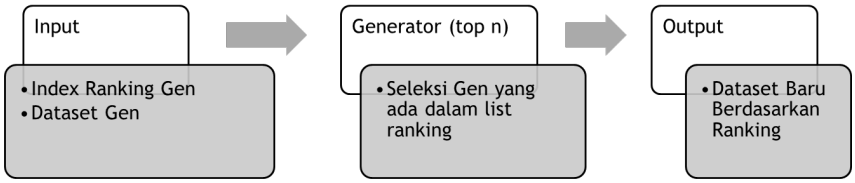
```
from sklearn import preprocessing
from sklearn import utils
import numpy as np
import gzip, cPickle
from utilitas import top_n_dataset

class Salah(Exception):
    pass

class Ekstraktor:
    nama_file = str
    data = np.empty
    target_file = str
    y = np.empty
    jumlah_data = int
    def norm_dataset(self,nama_file):
        self.nama_file = nama_file + ".csv"
        self.data = np.genfromtxt(self.nama_file, dtype=float, delimiter=",")
        min_max_scaler = preprocessing.normalize(self.data)
        #min_max_scaler = preprocessing.scale(self.data)
        #min_max_scaler = preprocessing.minmax_scale(self.data)
        np.savetxt(nama_file + "_norm.csv", min_max_scaler, delimiter=",")

    def generate_dataset(self, nama_file, target_file,train,valid,test, suffle = True):
        self.nama_file = nama_file + ".csv"
        self.target_file = target_file + ".csv"
        self.data = np.genfromtxt(self.nama_file, dtype=float, delimiter=',')
        self.y = np.genfromtxt(self.target_file, dtype=float, delimiter=',')
        self.data = self.data.transpose()
        self.jumlah_data = self.ambil_jumlah_dataset(self.data)
        jml_train,jml_valid,jml_test =self.ambil_train_valid_test(self.jumlah_data,train,valid,test)
        if suffle:
            self.data, self.y = utils.shuffle(self.data,self.y,random_state = 5)
        train_set_x = self.data[0:jml_train]
        valid_set_x = self.data[jml_train+1:jml_train+1+jml_valid]
        test_set_x = self.data[jml_train+1+jml_valid+1:jml_train+1+jml_valid+1+jml_test]
        train_set_y = self.y.transpose()[2][0:jml_train]
        valid_set_y = self.y.transpose()[2][jml_train+1:jml_train+1+jml_valid]
        test_set_y = self.y.transpose()[2][jml_train+1+jml_valid+1:jml_train+1+jml_valid+1+jml_test]
        train_set = train_set_x, train_set_y
        valid_set = valid_set_x, valid_set_y
        test_set = test_set_x, test_set_y
        dataset = [train_set, valid_set, test_set]
```

```
        self.simpan_data(self.nama_file + '_dataset.pkl.gz',dataset)
        return dataset

    def ambil_jumlah_dataset(self,data):
        return data.shape[0]

    def ambil_train_valid_test(self,jml_dataset,train,valid,test):
        # ambil train valid test dalam %
        if int(round(train+valid+test)) != 100 :
            raise Salah("train+valid+test harus = 100%")
        jml_train_set = int(round(float(jml_dataset)*(float(train)/100.)))
        jml_valid_set = int(round(float(jml_dataset)*(float(valid)/100.)))
        jml_test_set  = int(round(float(jml_dataset)*(float(test)/100.)))
        return jml_train_set,jml_valid_set,jml_test_set

    def simpan_data(self, n_file, data_simpan):
        f = gzip.open(n_file, 'wb')
        cPickle.dump(data_simpan, f, protocol=2)
        f.close()
        return data_simpan

    def load_data(self, data):
        # model_hasil = load cpickel
        f = gzip.open(data, 'rb')
        model_hasil = cPickle.load(f)
        return model_hasil

class Generator:
    ekstraktor = Ekstraktor()
    # data_rank adalah array dari ranking data
    def top_n_dataset(self, data_rank,dataset, namafile):
        data_hasil = top_n_dataset(data_rank, dataset)
        np.savetxt(namafile + ".csv", data_hasil, delimiter=",")
        return data_hasil

if __name__ == '__main__':
    ekstraktor = Ekstraktor()
    generator = Generator()
    array_rank= np.array([2, 3])
    train = 80.5
    valid = 14.5
    test = 5
    ekstraktor.norm_dataset("./dataset/iris_dataset")
    dataset_iris = np.genfromtxt("./dataset/iris_dataset_norm.csv", dtype=float, delimiter=",")
    generator.top_n_dataset(array_rank, dataset_iris,"./dataset/iris_dataset_rank")
    dataset_iris = ekstraktor.generate_dataset("./dataset/iris_dataset_rank",
                                        "./dataset/iris_target", train, valid, test, True)

    print dataset_iris
    # ekstraktor.norm_dataset("./dataset/GSE10072_dataset")
```

Kelas ekstraktor ini melakukan adaptasi data yang tadinya memiliki struktur yang tidak kompatibel dengan library Theano yang di python, menjadi kompatibel dan memiliki struktur data yang disesuaikan. Kemudian, dilakukan juga permutasi random agar datanya memiliki sebaran yang normal untuk kemudian dilakukan pembagian data yang terdiri dari sekian persen data training, validasi dan testing.

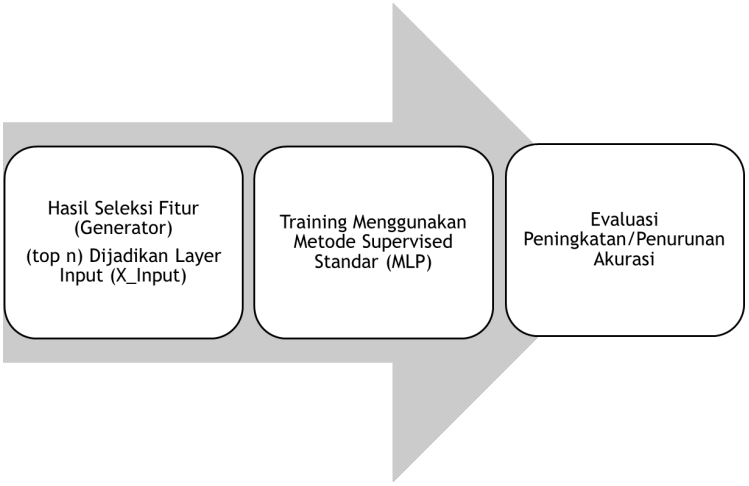
3.10.3 Kelas Generator

Kelas Generator ini adalah modul yang dibuat agar bisa secara otomatis memilih gen-gen yang dianggap penting pada sebuah array yang berisi index dari gen yang ada pada dataset.



Gambar 3.13: Diagram Kelas Generator yang digunakan untuk menggenerasi data gen berdasarkan rankingnya

3.10.4 Hasil Evaluasi Dengan Multi Layer Perceptron



Gambar 3.14: Diagram Proses Menggenerasi Data Untuk Dijadikan Dataset Training

BAB 4
PEMBAHASAN

Pada bab ini akan dibahas tentang hasil penelitian dari metodologi yang ada pada bab tiga, dan masalah-masalah yang dihadapi pada saat implementasinya dan pembahasannya.

4.1 Overview Metodologi

Model yang dihasilkan dari *unsupervised learning* yang dilakukan oleh DBN menggunakan data training, harus diuji dahulu dengan dengan data validasi dan data testing, yaitu dengan cara memberikan satu layer output menggunakan *logistic regression* hal ini untuk mengetahui apakah klasifikasinya lebih baik atau sebaliknya. Hasil ini berpengaruh pada proses tuning parameter (jumlah layer dan jumlah hidden unitnya) untuk didapatkan *cost* yang paling optimal pada saat pre-training. Setelah dilakukan perankingan secara multi-step dari hasil percobaan yang terbaik, diperlukan pengujian apakah apakah seleksi fitur tersebut mendapatkan hasil klasifikasi yang lebih baik dengan menggunakan fitur yang telah diseleksi saja. Dengan cara membandingkan *biomarker* yang ditemukan oleh algoritma multi-step rangking dibandingkan dengan algoritma yang ada di literatur yaitu metode bonferroni untuk melakukan test statistik pada data gen tersebut (Hochberg, 1988).

4.2 Hasil Percobaan DBN Dengan Setting Hyperparameter yang Berbeda

Untuk mendapatkan hasil yang optimal dibutuhkan banyak percobaan dan setting parameter yang berbeda-beda, mulai dari jumlah layer, jumlah hidden unit tiap layernya, learning rate, banyaknya gibbs step dan ukuran batch-nya. Oleh karena itu, dibawah adalah rekapitulasi percobaan dengan hasil terbaik dari sekian percobaan, dipilih lima percobaan yang paling baik hasilnya untuk emudian dianalisa lebih jauh. Percobaan dibawah memiliki setting parameter seperti pada daftar berikut:

Tabel 4.1: Setting Parameter Awal

No.	Item	Keterangan
1	Dataset	Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival (Landi et al., 2008)
2	Total Data	107 Pasien
3	Kanker	58 Pasien
4	Normal	49 Pasien
5	Training	69 Pasien
6	Validasi	14 Pasien
7	Testing	20 Pasien
8	Epoch	1000 dan 2000
9	Learning Rate	0.01
10	Fitur Gen	22.283 Gen

Setelah dilakukan eksperimen secara *unsupervised* diperoleh *cost* terbaik pada Percobaan dan hasilnya ada di tabel 4.2 :

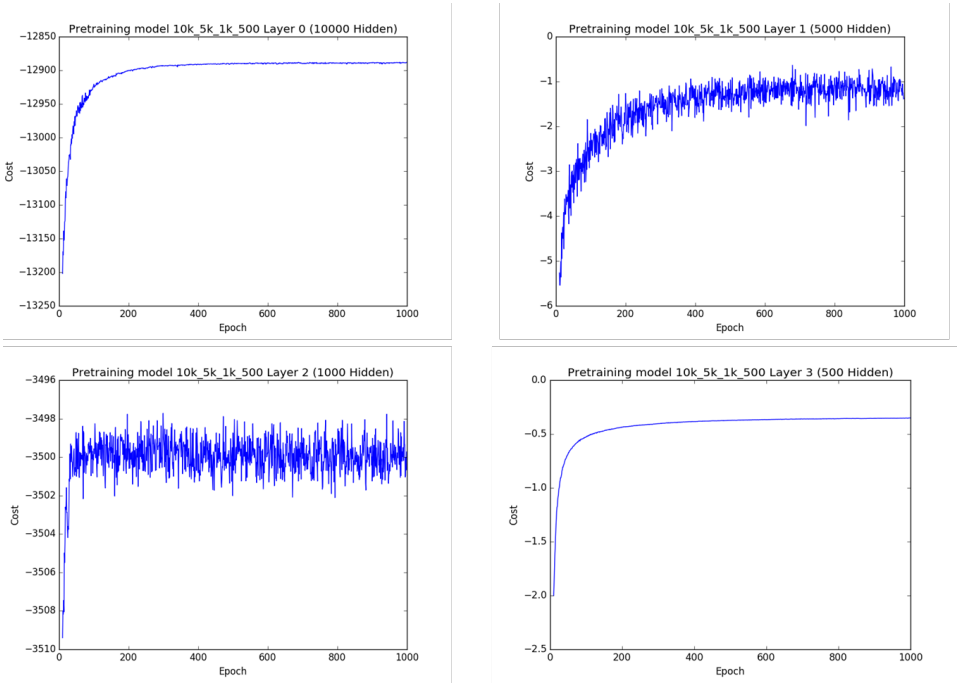
Tabel 4.2: Eksperimen DBN Unsupervised

Eks	Hidden	Epoch	Cost Lyr 0	Cost Lyr 1	Cost Lyr 2	Cost Lyr 3	Waktu (Jam)
1	[10000, 5000, 1000, 500]	1000	-12888.2	-1.37401	-3499.73	-0.351105	65
		2000	-12888.2	-0.828167	-3484.73	-0.150991	132
2	[7000, 10000, 5000, 1000]	1000	-12886.8	-1.36201	-6866.37	-0.163702	63
		2000	-12886.7	-1.57877	-6873.31	-0.0729352	138
3	[3000, 2000, 1000, 100]	1000	-12897.8	-0.862442	-1410.18	-3.244	58
		2000	-12897.0	-0.849616	-1397.09	-3.14657	123
4	[15000, 8000, 2000]	1000	-12934.5	-32.4227	-2756.41	-	68
5	[25000, 17000, 7000]	1000	-12888.1	-12.1715	-5446.34	-	72

Tabel diatas menunjukkan bahwa dengan epoch 1000 dan 2000 costnya tidak menunjukkan perbaikan secara signifikan. Bahkan untuk beberapa kasus, hasil-

nya lebih buruk. Dibawah adalah plot cost untuk percobaan yang dilakukan secara *greedy layer wise*, dari plot tersebut dapat dilihat bahwa cost pada epoch 700-an sudah tidak lagi membaik secara signifikan. Hal ini bisa dikarenakan oleh terbatasnya data training yang dipakai yaitu hanya 69 pasien dikarenakan oleh terbatasnya data yang didapatkan karena mahalnya percobaan microarray itu sendiri.

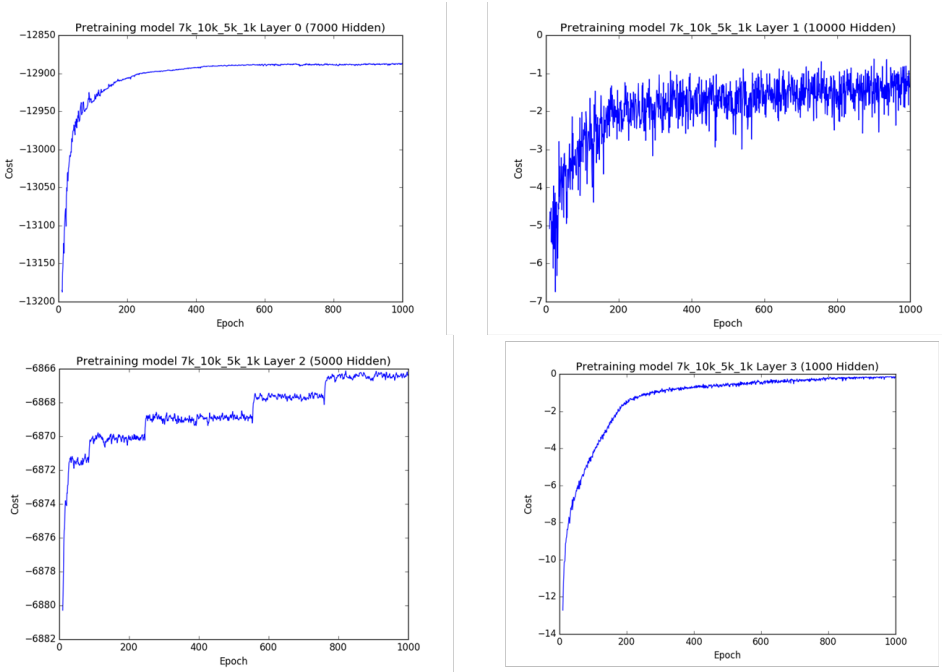
4.2.1 Plot Cost Percobaan 1



Gambar 4.1: Perbandingan Cost Pada Percobaan 1 Sampai 1000 Epoch Pada Tiap Layernya

Pada Gambar 4.1 merupakan perbandingan cost dari layer 0 sampai 3 (4 layer) dengan konfigurasi hidden [10000, 5000,1000,500] disitu bisa dilihat bahwa setelah epoch 500 tidak terjadi perbaikan cost yang signifikan. Juga cost pada layer 2 dan 3 memiliki ritme yang tidak stabil.

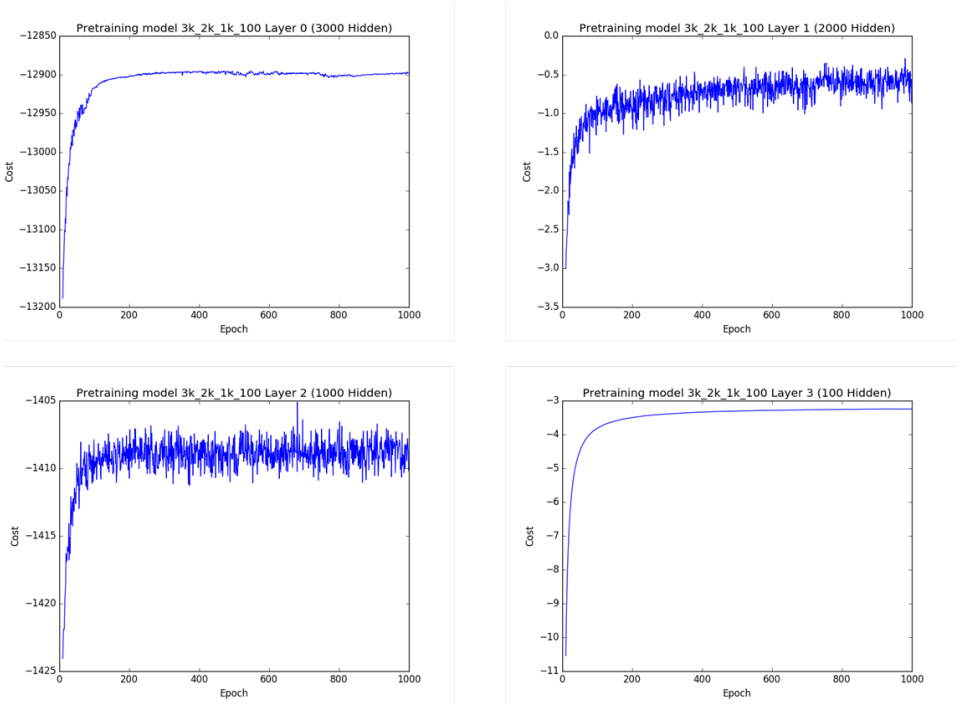
4.2.2 Plot Cost Percobaan 2



Gambar 4.2: Perbandingan Cost Pada Percobaan 2 Sampai 1000 Epoch Pada Tiap Layernya

Pada Gambar 4.2 merupakan perbandingan cost dari layer 0 sampai 3 (4 layer) disitu bisa dilihat bahwa setelah epoch 500 tidak terjadi perbaikan cost yang signifikan. Juga cost pada layer 2 dan 3 memiliki ritme yang juga tidak stabil.

4.2.3 Plot Cost Percobaan 3



Gambar 4.3: Perbandingan Cost Pada Percobaan 3 Sampai 1000 Epoch Pada Tiap Layernya

Pada Gambar 4.3 merupakan perbandingan cost dari layer 0 sampai 3 (4 layer) disitu bisa dilihat bahwa setelah epoch 500 tidak terjadi perbaikan cost yang signifikan. Juga cost pada layer 2 dan 3 memiliki ritme yang tidak stabil.

Berarti dari ketiga percobaan tersebut, secara garis besar, epoch lebih dari 700-an tidak mempengaruhi perbaikan error rekonstruksinya. Hal ini bisa disebabkan karena kurangnya data training.

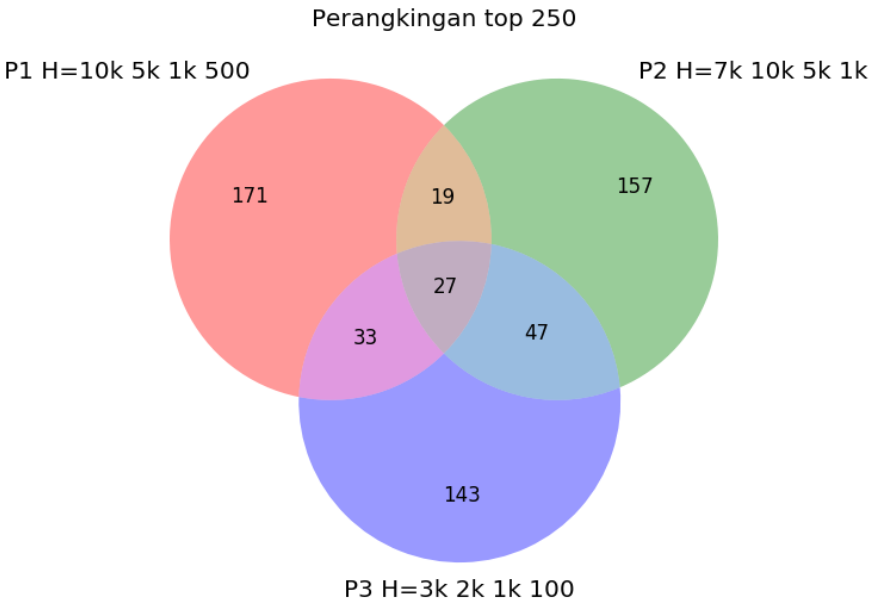
4.3 Hasil Penerapan Multi Step Ranking Bobot

Percobaan training DBN secara *unsupervised* yang dilakukan dengan setting pada tabel 4.2 diatas dipilih tiga percobaan terbaik untuk dilakukan algoritma multi-step ranking.

4.3.1 Diagram Venn Perpotongan Percobaan 1, 2 dan 3

Pada saat dilakukan multi-step ranking pada percobaan 1, 2 dan 3. Dibuat perankingan top 250 gen yang paling berpengaruh terhadap model-nya masing-masing. Kemudian, dibuat sebuah diagram untuk mendapatkan perpotongan 250 gen tersebut pada tiap-tiap percobaan. Hal ini digunakan untuk mengetahui gen-gen mana

yang selalu muncul di percobaan 1,2,3 atau muncul di dua percobaan dan hanya muncul di satu percobaan. Maka didapatkan diagram venn seperti pada Gambar 4.4



Gambar 4.4: Perbandingan Perankingan Top 250 pada tiga percobaan yang paling baik, ada 27 gen yang selalu muncul pada ketiga percobaan tersebut

Pada diagram venn diatas, ditunjukkan bahwa ada 27 gen yang selalu muncul pada percobaan 1, 2, 3. Hal ini menunjukkan bahwa gen tersebut adalah gen yang diindikasikan lebih informatif dibandingkan dengan gen yang lainnya. Ke 27 gen tersebut ada pada tabel 4.3 penemuan 27 gen yang selalu muncul pada tiga percobaan terbaik tersebut bisa diindikasikan sebagai *biomarker*. Yaitu gen yang bisa mencirikan seseorang terkena kanker paru-paru atau tidak.

Tabel 4.3: Index dan Kode Gen yang Diindikasikan sebagai *Biomarker*

Index	Kode Gen
7303	207783_x_at
1418	201891_s_at
9666	210183_x_at
15890	216520_s_at
24	200004_at
21919	38691_s_at
11298	211911_x_at
13741	214363_s_at
46	200026_at
307	200780_x_at
12727	213347_x_at
13246	213867_x_at
4418	204892_x_at
6084	206559_x_at
13765	214387_x_at
328	200801_x_at
201	200674_s_at
21860	37004_at
101	200081_s_at
232	200705_s_at
11370	211984_at
879	201352_at
11120	211720_x_at
20968	221607_x_at
115	200095_x_at
1019	201492_s_at
511	200984_s_at

Ke-27 gen pada tabel tersebut merupakan gen yang diindikasikan memiliki pengaruh yang signifikan pada percobaan. Akan tetapi hal ini perlu dilakukan konfirmasi lebih lanjut untuk memastikan bahwa gen tersebut memang berpengaruh secara signifikan terhadap penyakit kanker paru-paru. Ada dua tahapan konfirmasi yang pertama tahap konfirmasi dengan memastikan bahwa hasil klasifikasi dengan hanya menggunakan top 250 gen tersebut bisa mengklasifikasikan pasien

sehat dan pasien kanker. Tahap yang kedua adalah dengan cara konfirmasi melalui literatur tentang biomarker kanker paru-paru yang sudah ditemukan pada penelitian sebelumnya.

4.4 Bagian Supervised Learning Dengan Multi Layers Perceptron (MLP)

Setelah dilakukan perbandingan gen biomarker yang ditemukan pada proses perankingan diatas, top 250 gen tersebut dibuat menjadi data input untuk kasus klasifikasi. Untuk di evaluasi apakah hasil klasifikasinya lebih baik dibandingkan dengan tanpa seleksi fitur.
Tabel 4.4 merupakan perbandingan error antara logistic regression yang ditempatkan pada layer akhir DBN, tanpa dilakukan seleksi fitur. Dibandingkan dengan MLP yang memiliki 1 layer hidden dan 250 hidden unit. Untuk dilakukan training ulang dan dibandingkan dengan hasil yang diperoleh dari logistic regression.

Tabel 4.4: Perbandingan Error Antara Dengan dan Tanpa Seleksi Fitur

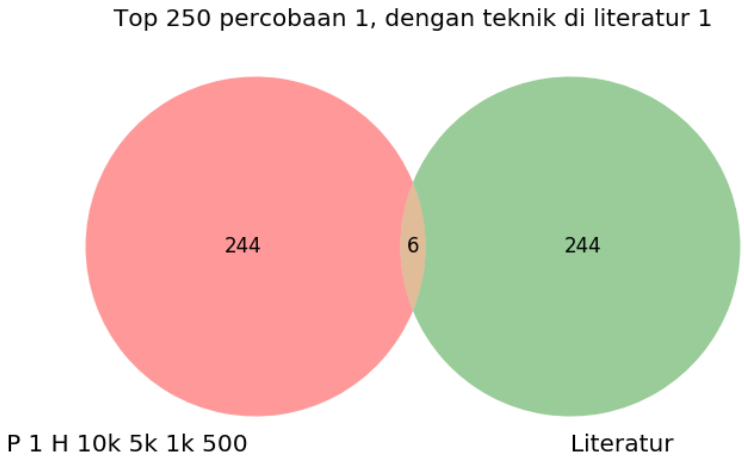
	Tanpa Seleksi Fitur(LogReg)		Dengan Seleksi Fitur(MLP)	
Percobaan	Validation Error	Test Error	Validation Error	Test Error
1	50%	66%	5.55%	0%
2	50%	30%	0%	8.33%
3	50%	30%	0%	16%

Dari tabel 4.4 dapat disimpulkan bahwa terjadi perbaikan signifikan antara validation dan test error dibandingkan tanpa dilakukan seleksi fitur. Akan tetapi hal ini masih belum menunjukkan apakah seleksi fitur gen tersebut merupakan *biomarker*. Oleh karena itu diperlukan evaluasi lebih lanjut yaitu dengan evaluasi literatur untuk memastikan bahwa gen yang ditemukan memang informatif untuk kasus kanker paru-paru.

4.5 Hasil Evaluasi Dengan Literatur Pertama Bonferroni Method(Hochberg, 1988)

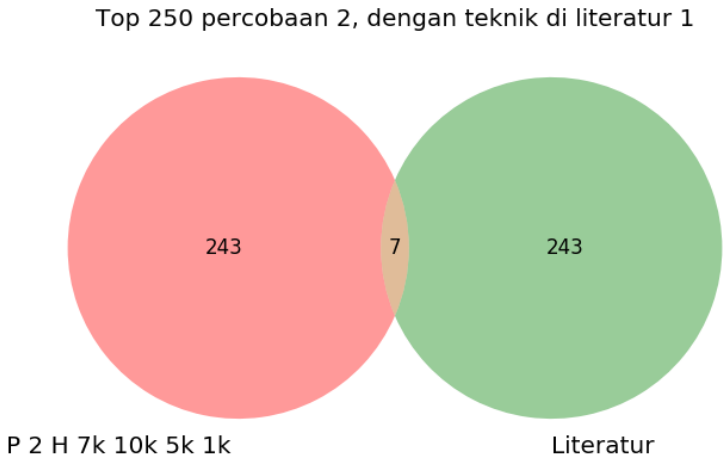
Metode Bonferroni adalah metode multipel testing di statistik yang paling umum digunakan untuk dataset dari percobaan *microarray*. Metode ini adalah metode yang dipakai oleh Landi et al. (2008) dalam menganalisa dataset GSE10072 yang merupakan hasil eksperimen kanker paru-paru (Landi et al., 2008) Dengan melakukan

test statistik menggunakan metode bonferroni dipilih 250 gen yang paling signifikan dari hasil test statistik tersebut dibandingkan dengan gen yang dipilih dari metode multi-step ranking, didapatkan hasil sebagai berikut.



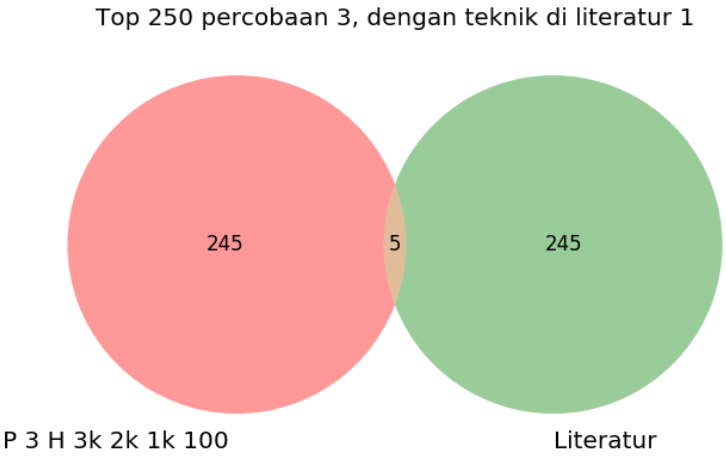
Gambar 4.5: Hasil top 250 Gen dibandingkan dengan Metode bonferroni

Pada percobaan 1, dihasilkan perpotongan 6 gen. Walaupun kelihatan kecil tetapi perpotongan 6 gen dari 22 ribu-an gen menjadi sangat signifikan untuk diteliti lebih lanjut gen-gen tersebut sebagai kandidat *Biomarker*



Gambar 4.6: Hasil top 250 Gen dibandingkan dengan Metode bonferroni

Percobaan 2 dibandingkan dengan metode bonferroni juga memiliki perpotongan yang tidak besar yaitu 7 gen saja.



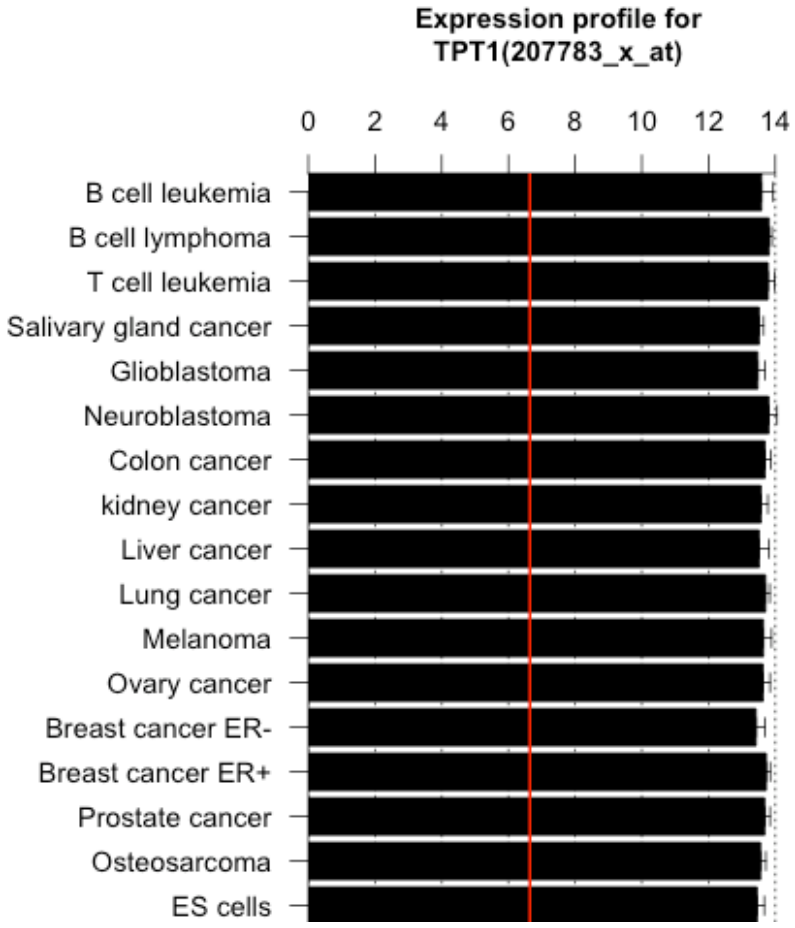
Gambar 4.7: Hasil top 250 Gen dibandingkan dengan Metode bonferroni

Percobaan 3 dibandingkan dengan metode bonferroni memiliki perpotongan ke-

sesuaaina 5 gen.

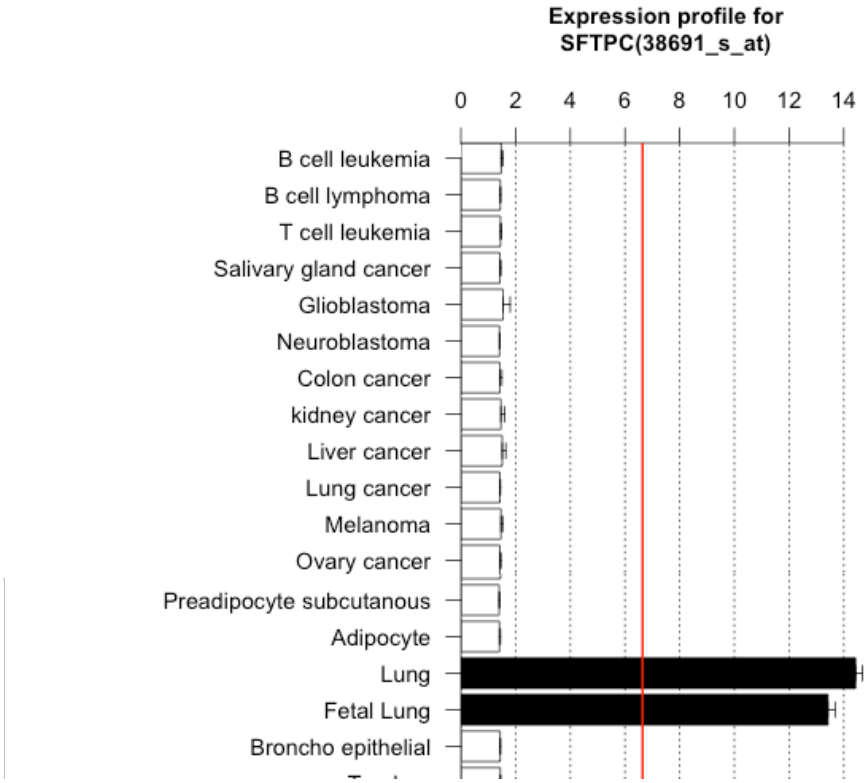
4.6 Hasil Konfirmasi Dengan Literatur Kedua Harvard Cancer Center (https://ccib.mgh.harvard.edu/xavier)

Sebanyak 27 gen yang ditemukan untuk irisan tiga percobaan terbaik, akan dilakuan review literatur lebih jauh. Menurut situs harvard cancer center, gen-gen tertentu bisa menununjukkan tingkat signifikansi gen tersebut terhadap sebuah penyakit kanker. Sebagai contoh gen yang berada pada ranking 1 pada 27 gen tersebut memiliki signifikasni yang tinggi terhadap kanker paru-paru dibandingkan dengan gen yang dipilih secara acak.



Gambar 4.8: Profil Ekspresi Gen TPT1 yang merupakan ranking pertama

Dari gambar bisa dilihat bahwa signifikansi gen TPT1 yang merupakan gen dengan ranking pertama memiliki signifikansi terhadap penyakit kanker paru-paru (lung cancer). Sumber profil gen didapat dari



Gambar 4.9: Profil Ekspresi Gen TPT1 yang merupakan ranking pertama

Pada dua contoh profil yang ditemukan yaitu gen TPT1 dan gen SFTPC bisa disimpulkan bahwa walaupun ekspresi gen tersebut ditemukan pada kanker paru-paru, tetapi tidak unik dan juga ditemukan di kanker-kanker yang lain misalnya leukemia, lymphoma dan sebagainya. Hal ini terjadi karena data yang dipakai adalah data kanker paru-paru saja. Sehingga menemukan biomarker yang unik pada kanker paru-paru saja.

4.7 Kendala-Kendala yang Dialami Selama Melakukan Percobaan

Pada saat melakukan percobaan dengan menggunakan arsitektur *deep learning* kendala yang paling utama adalah lamanya waktu training dan penggunaan resource memory yang sangat besar. Dengan menggunakan komputer core i5 dengan memory vga 2 GB, dan RAM 4 GB diperlukan waktu rata-rata 3-5 hari. Seperti pada tabel 4.5. Dikarenakan oleh kendala ini maka untuk melakukan percobaan dengan arsitektur yang lebih besar, misalnya dilakukan penambahan layer (lebih dari 4 layer) dan penambahan hidden unit, menjadi terbatas. Juga masalah pada terbatasnya dataset untuk training yang hanya 107 sampel pasien, hal ini disebabkan oleh

mahalnya percobaan *microarray* yang dilakukan sehingga sulit untuk mendapatkan data yang lebih besar lagi.

Tabel 4.5: tabel ukuran model dan waktu running

Percobaan	Konfigurasi Hidden (h0, h1, h2, h3)	Ukuran Model	Running (Jam) (1000e, 2000e)
1	10000, 5000,1000, 500	1 GB	65, 132
2	7000,10000,5000,1000	1 GB	63, 138
3	3000,2000,1000,100	275 MB	58, 123
4	15000,8000,2000	Out of Memory	-
5	25000, 17000, 7000	Out of Memory	-

Pada tabel diatas, bisa dihilhat bahwa hidden yang melebihi 15000 sudah menghabiskan RAM komputer yang hanya berukuran 4 GB. Oleh karena itu, percobaan yang seharusnya bisa memperdalam layer dan memperbesar hidden unit tidak memungkinkan untuk dilakukan.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Metodologi pencarian biomarker secara *unsupervised* dengan menggunakan teknik Deep Believe Network (DBN) memiliki kelebihan yaitu tidak diperlukannya pengetahuan kita tentang gen secara lengkap, dan bisa dilakukan generalisasi pada penyakit-penyakit lainnya.

Algoritma perankingan gen secara multi-step pada jaringan DBN yang merupakan modifikasi dari metode sebelumnya yang hanya bisa dilakukan pada teknik *logistic regression* sekarang bisa dilakukan untuk network DBN yang di training secara unsupervised murni.

Evaluasi yang dilakukan secara bertahap yaitu mulai dari dibandingkannya metode unsupervised dengan masalah klasifikasi supervised dengan MLP menunjukkan peningkatan hasil klasifikasi yang signifikan. Dan biomarker yang ditemukan, dibandingkan dengan literatur yaitu metode bonferroni menunjukkan bahwa gen yang ditemukan memiliki signifikansi yang tinggi.

5.2 Saran

Karena keterbatasan waktu penelitian dan mesin yang digunakan, maka ada banyak hal yang bisa dilakukan untuk penelitian selanjutnya, yaitu melakukan generalisasi, apakah metode ini cocok juga dilakukan untuk penyakit-penyakit lainnya.

Karena metode ini menggunakan arsitektur deep learning yang memiliki jaringan yang dalam, apakah dengan melakukan pada network DBN yang lebih dalam bisa meningkatkan keakuratan pendeteksian biomarker. Dikarenakan terbatasnya komputer, makan hal ini belum memungkinkan untuk dilakukan.

DAFTAR REFERENSI

M Mwanadan Babu. Introduction to microarray data analysis. *Computational genomics: Theory and application*, pages 225–249, 2004.

Supriyo Bandyopadhyay, Saurav Mallik, and Amit Mukhopadhyay. A survey and comparative study of statistical tests for identifying differential expression from microarray data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 11(1):95–115, 2014.

Steven A Belinsky. Gene-promoter hypermethylation as a biomarker in lung cancer. *Nature Reviews Cancer*, 4(9):707–717, 2004.

Kevin Duh. Deep learning & neural networks lecture. 2014.

Mourad Elloumi and Albert Y Zomaya. *Algorithms in computational molecular biology: techniques, approaches and applications*, volume 21. John Wiley & Sons, 2011.

Rasool Fakoor, Faisal Ladhak, Azade Nazi, and Manfred Huber. Using deep learning to enhance cancer diagnosis and classification. *roceedings of the International Conference on Machine Learning.*, 2013.

Mikael Häggström. Diagram of the pathways of human steroidogenesis. *Medicine*, 1:1, 2014.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

Maria Teresa Landi, Tatiana Dracheva, Melissa Rotunno, Jonine D Figueroa, Huaitian Liu, Abhijit Dasgupta, Felecia E Mann, Junya Fukuoka, Megan Hames, Andrew W Bergen, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PloS one*, 3(2):e1651, 2008.

Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

Deep Learning Tutorial. Lisa lab. *University of Montreal*, 2014.

Youngmi Yoon, Jongchan Lee, and Sanghyun Park. Building a classifier for integrated microarray datasets through two-stage approach. In *BioInformatics and BioEngineering, 2006. BIBE 2006. Sixth IEEE Symposium on*, pages 94–102. IEEE, 2006.

LAMPIRAN

LAMPIRAN 1

@todo
Yang dilampirkan : 1. Kode program 2. Log Pekerjaan 3. Dataset GSE10072 4. Review Literatur 27 Gen biomarker.