**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.

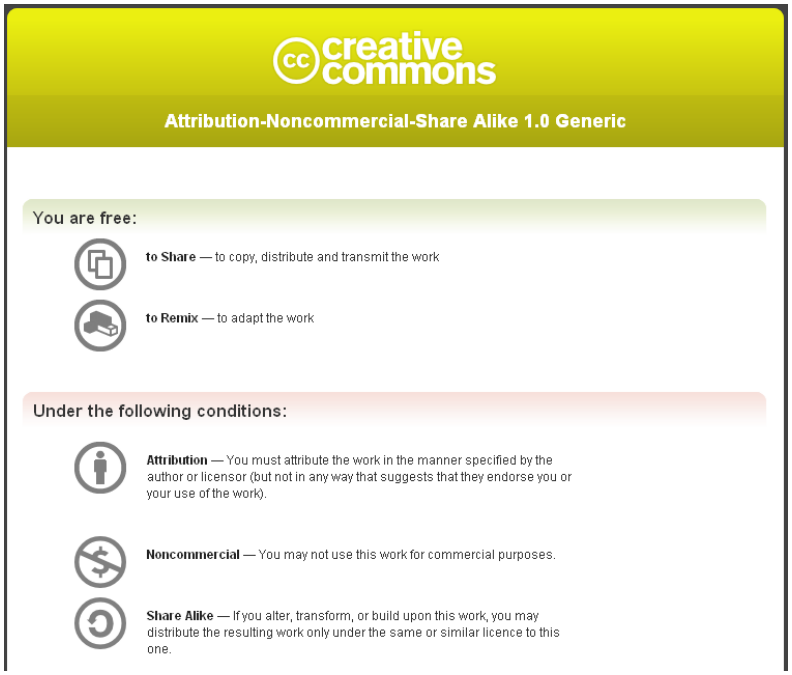
| | | |
|---|---|---|
| Ditetapkan di | : | Depok |
| Tanggal | : | XX Januari 2016 |

# KATA PENGANTAR

Template ini disediakan untuk orang-orang yang berencana menggunakan LaTeX untuk membuat dokumen tugas akhirnya. Mengapa LaTeX? Ada banyak hal mengapa menggunakan LaTeX, diantaranya:

1. pertama
2. kedua
3. ketiga

1. LaTeX membuat kita jadi lebih fokus terhadap isi dokumen, bukan tampilan atau halaman.
2. LaTeX memudahkan dalam penulisan persamaan matematis.
3. Adanya automatis dalam penomoran caption, bab, subbab, subsubbab, referensi, dan rumus.
4. Adanya automatisasi dalam pembuatan daftar isi, daftar gambar, dan daftar tabel.
5. Adanya kemudahan dalam memberikan referensi dalam tulisan dengan menggunakan label. Cara ini dapat meminimalkan kesalahan pemberian referensi.




**Gambar 1:** *Creative Common License 1.0 Generic*

Gambar 1 diambil dari http://creativecommons.org/licenses/by-nc-sa/1.0/deed.en_CA. Jika ingin mengentahui lebih lengkap mengenai *Creative Common License 1.0 Generic*, silahkan buka http://creativecommons.org/licenses/by-nc-sa/1.0/legalcode. Seluruh dokumen yang dibuat dengan menggunakan template ini sepenuhnya menjadi hak milik pembuat dokumen dan bebas didistribusikan sesuai dengan keperluan masing-masing. Lisensi hanya berlaku jika ada orang yang membuat template baru dengan menggunakan template ini sebagai dasarnya.

Dokumen ini dibuat dengan LaTeX juga. Untuk meyakinkan Anda, coba lihat properti dari dokumen ini dan Anda akan menemukan bagian seperti Gambar 2. Dokumen ini dimaksudkan untuk memberikan gambaran kepada Anda seperti apa mudahnya menggunakan LaTeX dan juga memperlihatkan betapa bagus dokumen yang dihasilkan. Seluruh url yang Anda temukan dapat Anda klik. Seluruh referensi yang ada juga dapat diklik. Untuk mengerti template yang disediakan, Anda tetap harus membuka kode LaTeX dan bermain-main dengannya. Penjelasan dalam PDF ini masih bersifat gambaran dan tidak begitu mendetail, dapat dianggap sebagai pengantar singkat. Jika Anda merasa kesulitan dengan template ini, mungkin ada baiknya Anda belajar sedikit dasar-dasar LaTeX.


**Gambar 2:** Dokumen Dibuat dengan PDFLatex

Semoga template ini dapat membantu orang-orang yang ingin mencoba menggunakan LaTeX. Semoga template ini juga tidak berhenti disini dengan ada kontribusi dari para penggunanya. Kami juga ingin berterima kasih kepada Andreas Febrian, Lia Sadita, Fahrurrozi Rahman, Andre Tampubolon, dan Erik Dominikus atas kontribusinya dalam template ini.

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 banyak. Algoritma *clustering* dan algoritma *deep learning* merupakan dua algoritma *unsupervised learning*, yang bisa membantu menganalisa data ekspresi gen. Algoritma pemilihan fitur digunakan untuk mendapatkan fitur gen yang paling penting. Dan kemudian akan digunakan algoritma *clustering* untuk mendapatkan struktur cluster dari data ekspresi gen. 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 LANDASAN TEORI
  Landasan teori dipakainya metodologi yang akan diterapkan dalam eksperimen ini.

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

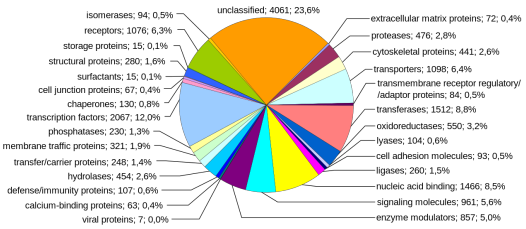
- Bab 4 HASIL PENELITIAN DAN PEMBAHASAN
  Pembahasan hasil dari eksperimen yang sudah dilakukan.

- Bab 5 KESIMPULAN DAN SARAN

# BAB 2
# LANDASAN TEORI

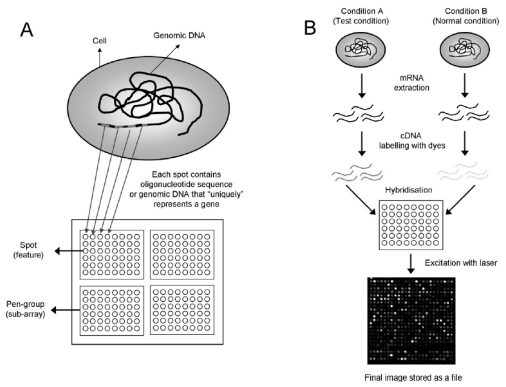
## 2.1 Ekspresi Gen

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.




**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.(Babu, 2004)

| **@todo** |
| --- |
| tambahkan dua gambar yang digambar sendiri |

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.

## 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. [15]

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

| **@todo** |
| --- |
| tulis tabel perbandingan seleksi fitur |

## 2.4 Deep Learning

| **@todo** |
| --- |
| tuliskan secara singkat tentang deep learning |

## 2.5 Restricted Boltzmann Machine

| @todo |
| --- |
| tuliskan secara detail tentang rbm |

## 2.6 Deep Believe Network

| @todo |
| --- |
| tuliskan secara detail tentang dbn |

## 2.7 Multi Layer Perceptron

| @todo |
| --- |
| tuliskan secara detail tentang dbn |

## 2.8 Logistic Regression

| @todo |
| --- |
| tuliskan secara detail tentang dbn |

# BAB 3
# METODOLOGI PENELITIAN

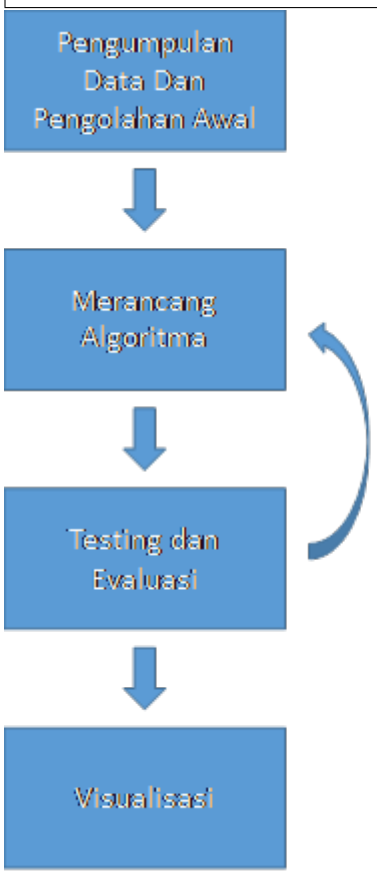
Penelitian ini dibagi menjadi empat tahap: (1) Mendapatkan data microarray dan pengolahan awal; (2) Perancangan algoritma; (3) Melakukan eksperimen untuk mendapatkan *hyperparameter* yang optimal. Kemudian dilanjutkan dengan testing dan evaluasi. Gambaran umum dari penelitian ini seperti pada gambar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

| **@todo** |
| --- |
| untuk sementara copy paste dari laporan |

## 3.1 Gambaran Umum Penelitian

**@todo**

1. include grafik
2. Bagian unsupervised dan supervised
3. RBM
4. Cost Function
5. DBN
6. Proses Multistep ranking
7. copy paste gambar tensorflow
8. list semua rumus yang dibutuhkan
9. tulis bab 2 yang isinya deep learning tutorial




## 3.2 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.

**@todo**

ambil source dari geo2r

## 3.3 Data Profil Gen Percobaan Microarray dan Biomarker

**@todo**

terangkan data profil gen sitasi pada paper :

1. paper tentang multi
2. papernya gse10072
3. contoh tabel ekspresi gen
4. detail kondisi pasien

## 3.4 Perancangan Algoritma

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.

### 3.4.1 Tahap Unsupervised

Tahap unsupervised adalah tahapan dimana model DBN ditraining secara unsupervised dengan data training pada tiap-tiap layernya secara greedy. Tiap layernya dihitung cost untuk kemudian diminimisasi errornya.

### 3.4.2 Tahap Supervised

Pada saat training unsupervised dilakukan untuk merekonstruksi xxx oleh karena itu tidak diketahui xxx

**@todo**

baca dan cari referensi supervised dan unsupervised

### 3.4.3 Tahap Tuning Parameter

## 3.5 Melakukan Testing Arsitektur DBN

**@todo**

buat bagan pengujian 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.

## 3.6 Implementasi Metode Perangkingan Bobot Secara Multi Step Untuk Mendapatkan Gen Biomarker

**@todo**

insert gambar 1 pada presentasi

**@todo**

beri keterangan perhitungan dan algoritmanya

**Algorithm 1:** How to write algorithms

**Data**: this text
**Result**: how to write algorithm with LaTeX2e
initialization;
**while** *not at end of this document* **do**
    read current;
    **if** *understand* **then**
        go to next section;
        current section becomes this one;
    **else**
        go back to the beginning of current section;

**@todo**

beri keterangan code pythonnya

## 3.7 Evaluasi Hasil Perangkingan Dengan Klasifikasi Secara Supervised

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.

## 3.8 Perbandingan Hasil Perangkingan Dengan 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.

# BAB 4
# HASIL PENELITIAN DAN PEMBAHASAN

**@todo**

tambahkan kata-kata pengantar bab 4 disini

## 4.1 Overview Metodologi

**@todo**

buat gambar overview metodologi yang akan dilakukan

## 4.2 Hasil Percobaan RBM Dengan Layer-layer yang Berbeda

**@todo**

percobaan 1 pecobaan 2 percobaan 3

## 4.3 Hasil Penerapan Multi Step Ranking Bobot

**@todo**

top 250 gen pada percobaan 1 2 dan 3, serta diagram venn hasil

## 4.4 Hasil Evaluasi Dengan Multi Layer Perceptron

**@todo**

penekanannya pada improvement klasifikasi

## 4.5 Hasil Evaluasi Dengan Literatur Benferrony Method

| **@todo** |
| --- |
| penekanannya pada bahwa hasil percobaan berpotongan xxx persen |

## 4.6 Hasil Evaluasi Dengan Literatur Harvard Cancer

| **@todo** |
| --- |
| penekanannya bahwa gen yang ditemukan sangat berkorelasi dengan kanker paru2 |

# BAB 5
# KESIMPULAN DAN SARAN

| **@todo** |
| --- |
| Tambahkan kesimpulan dan saran terkait dengan perkerjaan yang dilakukan. |

## 5.1 Kesimpulan

## 5.2 Saran

# 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.

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

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.

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.

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** |
| --- |
| Membuat todolist apa yang akan dikerjakan untuk thesis<br>1. belajar copy paste code<br>3. belajar buat bagan<br>4. belajar pseudo code<br>5. kumpulan dalam sebuah totorial dan link dengan cepat secara offline jika diperlukan<br>7. export ke odf |