Klasifikasi DNA *Microarray* Menggunakan *Principal Component Analysis* (PCA) dan *Artificial Neural Network* (ANN)

Dian Atia Ihsani, Achmad Arifin, dan Muhammand Hilman Fatoni
Departemen Teknik Biomedik, Institut Teknologi Sepuluh Nopember (ITS)
e-mail: *arifin@bme.its.ac.id*

Abstrak*—Kanker merupakan suatu kelompok penyakit tingkat molekular yang ditandai dengan pembelahan sel tidak terkendali yang memiliki potensi untuk menyerang jaringan biologis lainnya, baik dengan pertumbuhan pada jaringan sekitar (invasi) atau dengan perpindahan ke jaringan lain (*metastatic*). Banyaknya jumlah ekspresi gen serta besarnya data yang terkandung pada DNA membuat diagnosis bagi pasien penderita kanker terhambat. Lama waktu yang dibutuhkan untuk diagnosis kanker membuat perawatan yang diberikan tertunda, sehingga sel kanker sering kali telah menginvasi organ lain yang kemudian memicu tingginya tingkat kematian akibat kanker. DNA *microarray* merupakan suatu metode baru dalam dunia teknologi yang membantu proses analisis tingkat ekspresi jutaan gen pada suatu waktu. Melalui ekspresi gen ini, diagnosis penyakit, identifikasi tumor, serta deteksi mutasi dapat dilakukan secara efektif dan efisien. Dalam tugas akhir ini, salah satu kelas dari *Artificial Neural Network* (ANN), yakni *Multilayer Perceptron* (MLP), diaplikasikan sebagai metode klasifikasi. Untuk meningkatkan efisiensi proses klasifikasi, data yang berukuran besar direduksi dimensionalitasnya menggunakan metode *Principal Component Analysis* (PCA). Dua sub-tipe kanker paru-paru, yakni *Adenocarcinoma* (AC) dan *Squamous Cell Carcinoma* (SCC) digunakan sebagai data penguji untuk memvalidasi keberhasilan metode yang diajukan. Hasil klasifikasi dari dataset PCA dengan nilai *variance* yang meningkat menunjukkan nilai akurasi yang meningkat pula dengan nilai maksimal akurasi dari dataset *variance* 100% sebesar 90,02%.

***Kata Kunci—Artificial Neural Network, Cancer Classification,* DNA *Microarray, Principal Component Analysis*.**

I. PENDAHULUAN

KANKER merupakan suatu penyakit genetik, dimana terjadi suatu anomali yang disebabkan oleh kelainan satu atau lebih gen yang dapat mempengaruhi kondisi fenotip suatu individu. Kanker bermula pada abnormalitas gen yang mengatur produksi atau perkembangan (proliferasi) sel yang kemudian menimbulkan pertumbuhan tak terkendali yang bersifat *malignant* atau memiliki kecenderungan untuk memburuk. Pertumbuhan sel ini dipicu adanya mutasi gen[1]. Dalam pendekatan deteksi kanker, terdapat beberapa akar molekular utama yang harus dipahami yakni gen, *messenger*-RNA (mRNA), serta protein yang diproduksi. Ekspresi gen dapat dianalisis melalui proses transkripsi dari DNA menjadi *Ribonucleic Acid* (RNA), atau translasi RNA menjadi protein. Proses transkripsi melibatkan pembuatan copy RNA yang sempurna dari gen menggunakan DNA sebagai 'template'[2].

DNA *microarray* merupakan teknik esensial pada ilmu biologi molekular yang memungkinkan analisa dari tingkat ekspresi jutaan gen dalam satu waktu. DNA *microarray* memanfaatkan susunan dua dimensi dari *probe* oligonukleotida, dimana ratusan hingga ribuan *probe* oligonukleotida dapat direpresentasikan sebagai metode analisis sekuens DNA untuk menguji adanya mutasi genetik[3]. Oligonukleotida yang merupakan berkas polimer nukleotida pendek, yang terdiri atas 5 hingga 20 basa N, akan menunjukkan tingkat ekspresi gen berdasarkan sintesa dari berbagai molekul mRNA.

Proses ekstraksi DNA *microarray* dilakukan melalui kombinasi antara referensi DNA sehat dengan DNA penguji. Ribuan rantai tunggal dari sekuens DNA akan disintesis untuk mendapatkan sekuens komplementer, dimana sampel tersebut kemudian diletakkan pada suatu kontainer kaca untuk kemudian dihibridasi. Melalui DNA *microarray*, observasi mengenai hubungan antara kondisi fisiologis sel dengan pola ekspresi gen untuk studi tumor, progress perkembangan penyakit, respons sel terhadap stimulus, serta identifikasi obat tertarget dapat dilakukan[4]. Melalui DNA microarray, akan dapat diperhatikan gen-gen yang mempengaruhi suatu kondisi sampel, sehingga deteksi tipe kanker dan identifikasi obat yang tepat berdasarkan gen tersebut dapat dilakukan dengan tepat.

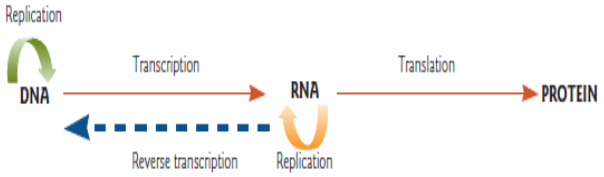
II. OVERVIEW

A. Genome Data

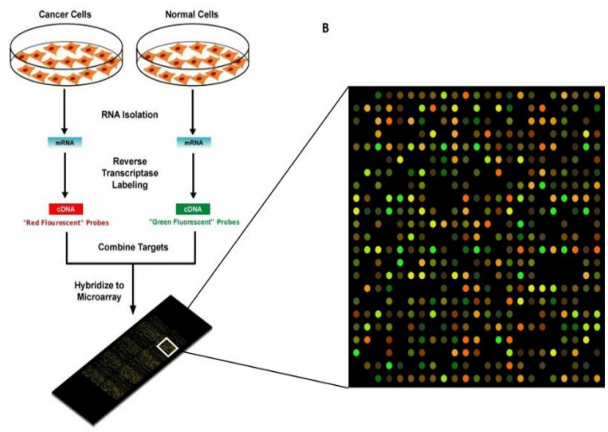
Dalam proses identifikasi molekular, terdapat dua metode utama yang umum digunakan. Pendekatan pertama adalah dengan menggunakan pendekatan DNA sekuens atau rantai DNA. Metode ini mengamati jumlah pasangan basa: *Adenine* (A), *Guanine* (G), *Thymine* (T), dan *Cytosine* (C), pada beberapa rantai DNA pembentuk gen yang diamati. Metode lain yang umum digunakan dalam identifikasi gen adalah melalui ekspresi gen. Akumulasi data ekspresi gen akan mampu menunjukkan kecenderungan dari suatu motif ekspresi, sehingga dapat menjadi suatu karakteristik atau informasi individual suatu data [2], antara lain adalah sebagai berikut

1) Deoxyribonucleic Acid (DNA)

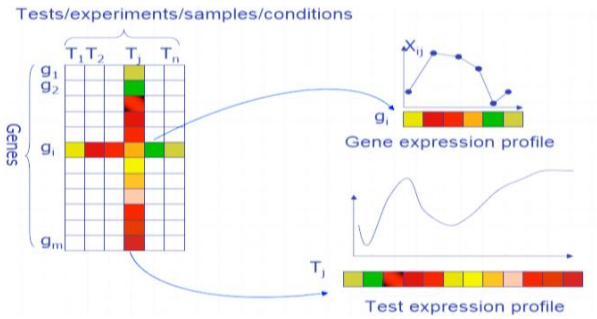
DNA merupakan wadah penyimpanan informasi genetika, dimana informasi ini tersimpan dalam struktur kimiawi yang disebut sebagai *nucleic acid*. *Nucleic acid* merupakan sebuah rantai atau polimer dari *nucleotides*. Tiap *nucleotide* terdiri atas tiga bagian utama, yakni sebuah 5-karbon *sugar*, sebuah kelompok fosfat, dan basa. Selain pada DNA, informasi genetika dapat tersimpan pula pada *Ribonucleic Acid* atau RNA. Perbedaan utama dari DNA dengan RNA adalah struktur polimer dimana DNA merupakan *double helix* atau dua rantai *nucleic* yang berbentuk *spiral*, sedangkan RNA



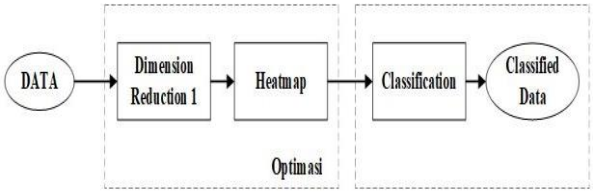
Gambar 1. Proses *Central Dogma*



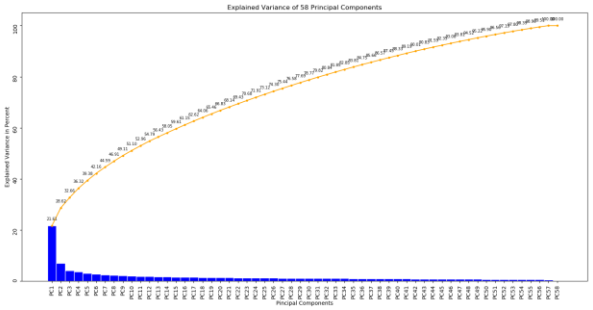
Gambar 2. Diagram eksperimen *microarray*



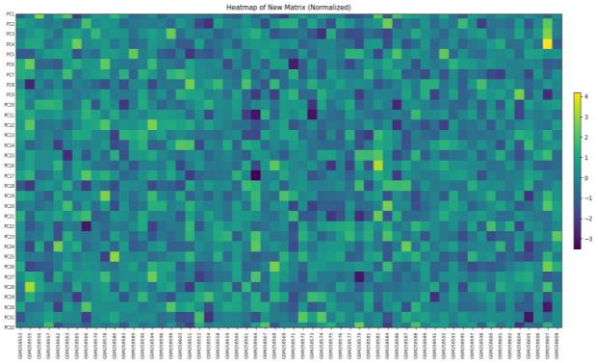
Gambar 3. Skema visualisasi dari kolom dan baris matriks *heatmap*.



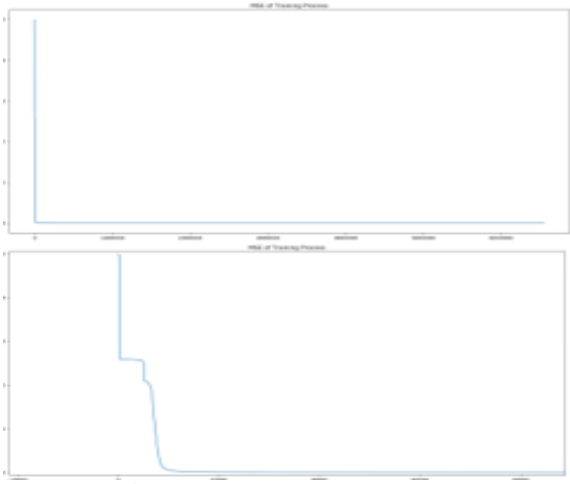
Gambar 4. Blok diagram sistem keseluruhan.



Gambar 5. Grafik nilai PC (k=58).



Gambar 6. Heatmap hasil PCA setelah dinormalisasi (k=32).



Gambar 7. Grafik MSE proses *training dataset* PC 32 (atas) dan *zoom in* grafik(bawah).

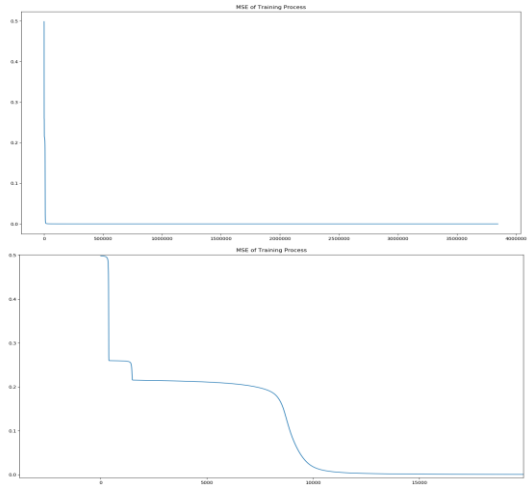
hanya terdiri atas satu *helix* atau rantai. RNA sendiri merupakan struktur polimer yang terbentuk dari untaian DNA yang telah diputus hubungan antar rantai *nucleic*-nya menggunakan *enzyme* DNA *polymerease.* Proses ini terjadi pada tahap transkripsi DNA[5].

2) Ekspresi Gen

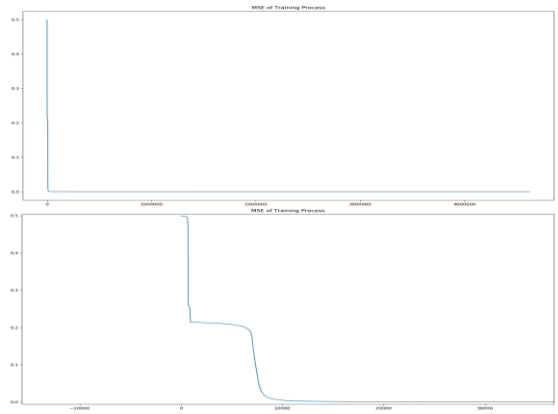
Pada 1930-an, diketahui bahwa gen merupakan unit dasar dari warisan genetika [5]. Pada dasarnya, gen merupakan potongan spesifik *nucleotide* pada DNA atau RNA yang mengandung informasi yang dapat digunakan untuk menghasilkan protein tertentu. Informasi yang terkandung pada gen dapat diekstraksi berdasarkan tingkat ekspresinya. Ekspresi gen merupakan tingkat protein atau RNA yang mengandung informasi genetika dari gen. Deskripsi dari alur informasi material genetika disebut juga sebagai *central dogma*. Sebagai ilustrasi alur *central dogma*, dapat diperhatikan Gambar 1. Proses transkripsi merupakan tahap dimana DNA dikopi untuk menghasilkan rantai RNA, sedangkan proses translasi merupakan proses RNA dikonversi menjadi rantai asam amino.

3) DNA Microarray

Salah satu pemanfaatan dari teknologi *cell sequencing* adalah DNA *microarray*. DNA *microarray* merupakan suatu teknologi yang mampu menghasilkan informasi genetika dengan memanfaatkan *array* berdensitas tinggi dari DNA atau *probe* oligonukleotida. Oligonukleotida merupakan potongan suatu rantai DNA yang umumnya memiliki panjang kurang dari 50 basa. *Array* yang didapat dari DNA atau *probe* oligonukleotida ini kemudian diletakkan pada material semikonduktor yang disebut sebagai *chips*. Umumnya, *array* atau *probe* oligonukliotida yang diletakkan pada *chips* mengandung jumlah *picomole* (10^{-12} moles) dari sekuens DNA terspesifik. *Probe* yang diletakkan pada *chip* dapat merupakan bagian pendek dari gen atau elemen lainnya dari DNA yang digunakan untuk hibridasi cDNA atau cRNA. Dalam proses hibridasi DNA *microarray*, *probe* disebut juga



Gambar 8. Grafik MSE proses *training* dataset PC 42 (atas) dan *zoom in* grafik (bawah).



Gambar 9. Grafik MSE proses *training* dataset PC 58 (atas) dan *zoom in* grafik (bawah).

sebagai *reporters* sedangkan cDNA atau cRNA disebut sebagai *target*. Hibridasi *probe-target* umumnya dideteksi dan dikuantitasi dengan deteksi *fluorophore-*, *silver-*, atau *chemiluminescence-labeled targets* untuk menentukan sekuens *nucleic acid* dari *target*. Pemanfaatan *chip* ini membuat DNA *microarray* disebut juga sebagai *DNAchips* atau *Genechips*. Penggunaan lebih dari satu *probe* pada tahap hibridasi, dimana jarak antar probe berukuran sangat kecil hingga ukuran mikro, kemudian menjadi dasar penamaan DNA *Microarray* [6]. Secara umum, proses hibridasi DNA *microarray* dapat diilustrasikan seperti pada Gambar 2;

4) Microarray Heatmap

DNA *microarray* dapat diorganisir menggunakan matriks *heatmap*. Matriks *heatmap* merupakan matriks dua dimensi, dimana amplitudo data akan direpresentasikan menggunakan spektrum warna. Pada implementasi visualisasi DNA microarray menggunakan *heatmap*, baris akan merepresentasikan ekspresi gen (g_n), sedangkan kolom akan merepresentasikan jumlah sampel atau tes (t_n) yang dilakukan. Dengan menggunakan visualisasi *heatmap*, tingkat ekspresi gen dari masing-masing sampel akan dapat terlihat ketinggiannya. Sebagai contoh, warna hijau dapat diasosiasikan dengan regulasi bawah yang menunjukkan rendahnya tingkat ekspresi suatu gen, sedangkan warna merah merepresentasikan regulasi atas dimana tingkat ekspresi gen tinggi. Penggunaan heatmap dapat membantu analisis awal dalam pembentukan hipotesis serta untuk memberikan informasi mengenai penggunaan sistem yang akan digunakan. Skema Visualisasi dapat dilihat pada Gambar 3.

5) Principal Component Analysis

Principal Component Analysis atau PCA merupakan suatu metode yang mampu membantu reduksi dimensionalitas data yang terdiri atas variabel yang saling berhubungan dalam skala besar, dengan menyimpan kemungkinan variasi sebanyak mungkin dari data set tersebut. PCA akan mentransformasi data menjadi suatu variabel set baru yang disebut sebagai *principal component* (PC). PC merupakan variabel yang didapat melalui dekomposisi *eigen* yang terdiri atas *eigenvalue* dan *eigenvector*. PCA sendiri merupakan metode pengembangan dari teori *Karhunen-Loève Transform* (KLT) yang merupakan suatu metode transformasi *linear* pada metode pengolahan citra. Pada praktiknya, metode PCA dan KLT dapat dikatakan sebagai metode yang sama apabila data memiliki vector dengan nilai rata-rata 0[7]. Perhitungan dari proses PCA dapat diperhatikan pada Persamaan 1.

$$Y = \phi^T X \quad (1)$$

dimana:
X = matriks dimensi N x M dengan $X = [x_1, \ldots, x_m]^T$
ϕ = N x M matriks *orthogonal* berbasis vektor
Y = N x M koefisien *weighting* dari matriks
N = jumlah sampel data
M = jumlah fitur data

6) Artificial Neural Network (ANN)

Artificial Neural Network (ANN) merupakan suatu seri persamaan matematikan yang bersifat *non-linear* dan berinterkoneksi, yang kemudian menyerupai sistem *neuronal* yang kemudian digunakan untuk mengkomputasi suatu variabel output berbasis variabel input terkontrol[8]. *Multilayer Perceptron* (MLP) merupakan salah satu kelas dari ANN yang terdiri atas tiga bagian utama yakni *input layer*, *hidden layer*, serta *output layer*. Data matriks yang telah direduksi dimensionalitasnya menjadi masukkan jaringan untuk kemudian diklasifikasi berdasarkan subtipe kanker yang direpresentasikan. Arsitektur MLP akan menggunakan 2 *hidden layer* dengan jumlah *node* pada *hidden layer* 1 sejumlah 50 dan pada *hidden layer* 2 sejumlah 10 *node*. Metode komputasi *feedforward* dan *backpropagation* akan digunakan sebagai proses komputasi klasifikasi. Komputasi *feedforward* akan menghitung data masukkan hingga mencapai output *layer*. Untuk memperbaiki nilai *weight* dan *connection* antar *layer*, dilakukan komputasi *backpropagation*. *Update weight* dari masing-masing *node* dan *connection* akan dilakukan dengan metode *Adaptive Linear Element* (ADLINE). Fungsi aktivasi yang akan digunakan pada proses perhitungan *backpropagation* adalah *binary sigmoid function*. Pada tiap iterasi *feedforward* dan *backpropagation*, akan terjadi *loss* pada *layer* terakhir yang disebut sebagai *Mean Square Error* (MSE). Nilai MSE ini kemudian menjadi fungsi objektif yang akan diminimalisir hingga tercapai nilai konvergensi tertentu yang sebelumnya telah ditentukan.

Tabel 1.
Hasil klasifikasi menggunakan 32 PC

No	Training & Testing Process									
	1		2		3		4		5	
	D.	O.	D.	O.	D.	O.	D.	O.	D.	O.
1	AC	AC	AC	AC	AC	SCC	AC	AC	SCC	AC
2	AC	AC	AC	AC	AC	AC	SCC	SCC	AC	SCC
3	AC	AC	AC	AC	AC	SCC	AC	AC	AC	SCC
4	AC	SCC	AC	AC	AC	AC	AC	AC	AC	AC
5	AC	AC	AC	AC	AC	SCC	SCC	SCC	AC	AC
6	AC	SCC	AC	AC	AC	AC	AC	AC	AC	AC
7	AC	AC	AC	SCC	AC	AC	AC	AC	SCC	SCC
8	AC	AC	AC	AC	SCC	SCC	AC	SCC	SCC	SCC
9	SCC	SCC	SCC	SCC	SCC	AC	SCC	SCC	AC	AC
10	SCC	SCC	SCC	AC	SCC	SCC	AC	AC	AC	SCC
11	SCC	AC	SCC	AC	SCC	SCC	AC	AC	AC	AC
12	SCC	SCC	SCC	SCC	AC	AC	-	-	-	-
		Accuracy: 75%		*Accuracy*: 75%		*Accuracy*: 66.67%		*Accuracy*: 90.9%		*Accuracy*: 63.63%
										Average accuracy: 74.24%

.

Tabel 2.
Hasil klasifikasi menggunakan 42 PC

No	Training & Testing Process									
	1		2		3		4		5	
	D.	O.	D.	O.	D.	O.	D.	O.	D.	O.
1	AC	AC	AC	AC	AC	SCC	AC	SCC	SCC	SCC
2	AC	AC	AC	AC	AC	AC	SCC	SCC	AC	AC
3	AC	AC	AC	AC	AC	AC	AC	AC	AC	SCC
4	AC	SCC	AC	AC	AC	SCC	AC	AC	AC	AC
5	AC	AC	AC	AC	AC	SCC	SCC	SCC	AC	AC
6	AC	AC	AC	AC	AC	AC	AC	AC	AC	AC
7	AC	AC	AC	SCC	AC	AC	AC	AC	SCC	SCC
8	AC	AC	AC	AC	SCC	SCC	AC	SCC	SCC	SCC
9	SCC	SCC	SCC	SCC	SCC	AC	SCC	SCC	AC	AC
10	SCC	SCC	SCC	AC	SCC	SCC	AC	AC	AC	SCC
11	SCC	AC	SCC	AC	SCC	SCC	AC	AC	AC	AC
12	SCC	SCC	SCC	AC	AC	AC	-	-	-	-
		Accuracy: 83.33%		*Accuracy*: 66.67%		*Accuracy*: 66.67%		*Accuracy*: 81.81%		*Accuracy*: 81.81%
										Average accuracy: 76.058%

Tabel 3.
Hasil klasifikasi menggunakan 58 PC

No	Training & Testing Process									
	1		2		3		4		5	
	D.	O.	D.	O.	D.	O.	D.	O.	D.	O.
1	AC	AC	AC	AC	AC	SCC	AC	AC	SCC	SCC
2	AC	AC	AC	AC	AC	AC	SCC	SCC	AC	AC
3	AC	AC	AC	AC	AC	AC	AC	AC	AC	AC
4	AC	AC	AC	AC	AC	AC	AC	AC	AC	AC
5	AC	AC	AC	AC	AC	AC	SCC	SCC	AC	AC
6	AC	AC	AC	AC	AC	AC	AC	AC	AC	AC
7	AC	AC	AC	SCC	AC	AC	AC	SCC	SCC	SCC
8	AC	AC	AC	AC	SCC	SCC	AC	AC	SCC	SCC
9	SCC	SCC	SCC	SCC	SCC	SCC	SCC	SCC	AC	AC
10	SCC	SCC	SCC	AC	SCC	SCC	AC	AC	AC	AC
11	SCC	AC	SCC	AC	SCC	SCC	AC	AC	AC	AC
12	SCC	SCC	SCC	AC	AC	AC	-	-	-	-
		Accuracy: 91.67%		*Accuracy*: 66.67%		*Accuracy*: 91.67%		*Accuracy*: 100%		*Accuracy*: 100%
										Average accuracy: 90.02%

III. METODOLOGI

A. *Desain Sistem Secara Umum*

Secara umum, alur kerja sistem penelitian dapat diperhatikan pada ilustrasi diagram blok sistem pada Gambar 4. Sistem akan memiliki dua alur utama, yakni proses optimasi dan proses klasifikasi. Pada proses optimasi, PCA akan dilakukan dengan menentukan nilai PC yang akan digunakan sebagai dimensi dari matriks baru. Pemilihan nilai PC dilakukan dengan nilai *variance* 80%, 90% dan 100%. Pemilihan nilai *variance* yang berbeda ini akan digunakan sebagai pembanding variabel untuk proses klasifikasi.

Proses klasifikasi data menggunakan permodelan ANN akan mengadaptasi topologi *multilayer perceptron* (MLP). Arsitektur dari ANN akan memanfaatkan 2 buah *hidden layer* dengan masing-masing *layer* memiliki jumlah *node* 50 dan 30. Metode komputasi *feedforward* dan *backpropagation* akan dilakukan untuk mendapatkan nilai *weight* paling ideal. Pendapatan nilai *weight* yang ideal dilakukan melalui proses training dengan memanfaatkan teori ADLINE sebagai metode untuk mengupdate nilai *weight*. Proses *training* akan terus dilakukan hingga nilai konvergen sebesar 10^{-6} tercapai, dengan pemanfaatan nilai *learning rate* sebesar 0.01 dan

fungsi aktivasi *binary sigmoid* dengan nilai *alpha* sebesar 0.5. Dari 58 jumlah sampel yang terdiri atas subtype kanker AC dan SCC, akan dilakukan pembagian jumlah sampel untuk *training* dan *testing* berdasarkan pembagian *k-fold cross validation* dengan nilai *k* sebesar 5.

B. Dataset

Penelitian ini akan memanfaatkan DNA *microarray* sebagai tipe data *input*, dimana 2 subtipe kanker dari tipe kanker *Non-Small Cell Lung Cancer* (NSCLC) akan digunakan. *Dataset* didapat melalui situs *ncbi.nlm.nih.gov* dengan nomor *dataset* GDS3627 dan nomor Pubmed ID 18486272. *Platform dataset* adalah GPL579: [HG-U133_Plus_2] dengan tipe organisme *Homo Sapiens* dan sampel tipe RNA. Proses hibridasi DNA *microarray* dilakukan dengan metode *in situ oligonucleotide*. Nomor seri dataset adalah GSE 10245 dengan subset GDS3627_1 merupakan penomoran subtipe kanker *Squamous Cell Carcinoma* (SCC) dan subset GDS3627_2 untuk subtipe kanker *Adenocarcinoma* (AC). Sebanyak 58 sampel terdiri atas 40 set sampel subtype kanker AC dan 18 subtipe kanker SCC.

C. Optimasi

Principal Component Analysis (PCA) akan digunakan sebagai metode optimasi data. Proses pencarian nilai PC akan memanfaatkan metode *eigen decomposition* dari matriks *covariance* (58 x 58). Setelah nilai PC ditemukan, akan dilakukan visualisasi grafik dari nilai PC berserta nilai *cumulative variance* dari masing-masing PC tersebut. Adapun untuk nilai PC yang akan digunakan dibataskan dengan nilai *variance* sebesar 80%, 90%, dan 100%.

Sebelum data diproses menggunakan PCA, data dinormalisasi terlebih dahulu untuk menyamakan nilai distribusi dari data. Data yang diambil dari format .csv akan dinormalisasi menggunakan metode *StandardScaler*. Metode *StandardScaler* akan mentrasform data sehingga distribusi data akan memiliki nilai rata-rata 0 dan *standard deviasi* senilai 1. Seluruh proses komputasi PCA menggunakan bahasa pemrograman *Python* akan memanfaatkan *module numpy* dan *pandas* sebagai *library* untuk mengelola data, *module sklearn* untuk proses standardisasi menggunakan *StandardScaler*, module *numpy*.*lilag* untuk dekomposisi *eigen*, serta *module matplotlib* sebagai library visualisasi.

Setelah data ternormalisasi, dilakukan perhitungan matriks baru menggunakan nilai PC maksimal, yakni 58. Melalui tahap ini, dapat dilihat jumlah *variance* data yang direpresentasikan oleh ke-58 PC tersebut. Karna tujuan dari PCA adalah untuk mendapatkan matriks yang paling *compact*, maka ditentukan batas *variance* 80% sebagai *threshold* penentu nilai PC pertama. Sebagai data uji untuk tahap klasifikasi, akan digunakan 3 nilai *variance* yang berbeda yakni 80%, 90%, dan 100%. Penentuan nilai *variance* ini dirasa cukup untuk merepresentasikan *variance* matriks data origin. Setelah didapat nilai dari dekomposisi PC, dilakukan visualisasi grafik untuk menunjukkan persebaran nilai PC dan *cumulative variance*.

D. Klasifikasi

Proses klasifikasi akan memberikan informasi mengenai subtype kanker berdasar tingkat ekspresi gen. Proses klasifikasi akan memanfaatkan metode komputasi *Artificial Neural Network* (ANN) dengan arsitektur *Multilayer Perceptron* (MLP) dengan algoritma *feedforward* and *backpropagation*. Metode ini dipilih karna kemampuan MLP untuk mengklasifikasikan data yang bersifat *non-linearly separable*. Akan dimanfaatkan 2 *hidden layer* dengan masing-masing *layer* terdiri atas 50 dan 30 *node*. Nilai *learning rate* yang digunakan pada proses *training* adalah 0.01. Dengan fungsi aktivasi *binary sigmoid*, akan diaplikasikan besar *alpha* sebesar 0.5 dan batas konvergen sebesar (10^{-6}).

Dengan 58 jumlah sampel yang tercampur dari subtype kanker AC dan SCC, akan digunakan *k-fold cross validation* dengan nilai *k* sebesar 5 untuk proses *training* dan *testing*. Dilakiukan 5 kali proses *training* dan *testing*, dengan 3 kali proses *testing* menggunakan 12 sampel dan 2 kali proses testing menggunakan 11 sampel berdasarkan pembagian dari nilai *k* yang digunakan, dengan data *training* sebanyak 46 untuk data *training* 1, *training* 2, dan *training 3*, dan data *training* sebanyak 47 unutuk *dataset training* 4 dan *training* 5. Jumlah dari *node* pada *input layer* akan bergantung pada jumlah dimensi dari dataset yang digunakan. Grafik Nilai Pc (k=58) dapat dilihat pada Gambar 5.

IV. PENGUJIAN DAN HASIL

A. Pengujian

Tahap pengujian pertama dilakukan dengan melakukan proses optimasi. Reduksi dimensionalitas data menggunakan PCA dilakukan dengan data matriks yang sudah di *transpose* menjadi 58 x 54,675. Selanjutnya dilakukan kalkulasi matriks *covariance* untuk kemudian dilakukan *eigen decomposition*. Didapatkan maksimal nilai PC sebesar 58 dengan nilai *variance* 100%. Setelah ditentukan nilai PC yang akan digunakan, dilakukan visualisasi *heatmap* dari matriks dengan dimensi baru. Matriks baru ini kemudian mejadi nilai masukkan dari proses klasifikasi menggunakan ANN.

Topologi ANN akan memanfaatkan arsitektur dari MLP, dengan jumlah *input layer* sejumlah dengan pembagian proporsi *training* dan *testing dataset* sesuai pembagian *k-fold cross validation*. Digunakan 2 buah *hidden layer* dengan masing-masing *layer* memiliki 50 dan 30 *node*. Fungsi yang digunakan pada proses ANN adalah *binary sigmoid* dengan nilai *alpha* sebesar 0.5.

B. Hasil

Hasil yang diperoleh dari beberapa proses antara lain:

1) Optimasi

Proses optimasi dimulai dengan mentranspos dataset awal. Setelah dilakukan proses transpos, didapatkan matriks berdimensi 58 x 54,678. Melalui data transpos ini, dilakukan normalisasi menggunakan metode StandardScaler dan dilanjutkan dengan menghitung covariance matrix. Diketahui nilai eigenvalue tertinggi adalah 58 dilihat dari dimensi matriks setelah ditranspos. Komputasi kemudian dilakukan dengan mengurutkan hasil eigen decomposition berdasarkan nilai eigenpairs dari nilai tertinggi ke rendah. Setelah didapatkan urutan nilai eigenvalue, selanjutnya dilakukan perhitungan nilai cumulative variance. Setelah 58PC didapatkan, dilakukan analisis terhadap nilai cumulative variance dari PC. Karna threshold yang diharapkan dari variance yang representative data adalah 80%, 90%, dan

100%. Akan digunakan PC-32 sebagai representasi 80% variance, PC-42 sebagai representasi 90% variance, dan PC-58 sebagai representasi 100% variance.

2) Klasifikasi

Proses komputasi menggunakan 32 PC dilakukan dengan 5 pairs dataset training dan testing. Masing-masing proses training dan testing memakan waktu rata-rata 3 jam untuk mencapai MSE. Heatmap hasil PCA setelah dinormalisasi (k=32) dapat dilihat pada Gambar 6. Pada tabel, D merepresentasikan desired output dan O merepresentasikan hasil komputasi ANN. Grafik perubahan nilai MSE menggunakan dataset variance 80% dapat diperhatikan pada Gambar 7 dan hasil klasifikasi pada Tabel 1. Selanjutnya proses klasifikasi dengan persentasi variance sebesar 90% dan 100% dilakukan. Digunakan PC 42 sehingga besar dimensi data yang menjadi data masukkan proses klasifikasi adalah (58 x 42) untuk variance 90% dan 58 PC untuk variance 100%. Hasil klasifikasi data dengan 42 PC dapat diperhatikan pada Tabel 2 dengan grafik MSE pada Gambar 8. Hasil klasifikasi menggunakan 100% variance atau 58 PC dapat diperhatikan pada Tabel 3 dengan grafik MSE dari masing-masing dataset pada Gambar 9. Kenaikan nilai akurasi sesuai dengan meningkatnya nilai variance yang digunakkan sesuai dengan teori PCA bahwa semakin besar nilai variance yang digunakan, semakin besar pula kemiripan yang dimiliki oleh matriks baru namun dimensi akan semakin besar pula.

V. KESIMPULAN/RINGKASAN

DNA *microarray* merupakan teknik pada ilmu biologi molekular yang memungkinkan analisis dari tingkat ekspresi jutaan gen dalam satu waktu. Pemanfaatan DNA *microarray* sebagai proses deteksi atau analisis penyakit kanker merupakan pendekatan yang efisien karna besarnya informasi yang dapat diperoleh dari *microarray*. Peningkatan efisiensi analisis terhadap *microarray* akan mampu membantu memberikan informasi untuk pemberian diagnosis dan/atau penanggulangan atas penyakit kanker.

Besarnya informasi yang terkandung pada DNA *microarray* membuat dibutuhkan adanya proses *pre-processing* sehingga *dataset* dapat diproses sesuai dengan metode dan tujuan yang diharapkan. Penentuan metode yang tepat akan memaksimalkan informasi yang terkadung pada DNA *microarray*. *Principal Component Analysis* (PCA) dapat diaplikasikan sebagai metode reduksi dimensi *microarray* tanpa menghilangkan fitur-fitur utama data. Pada penelitian ini, proses klasifikasi menggunakan proses reduksi dimensi PCA memiliki nilai akurasi semakin meningkat dengan penggunaan nilai *variance* yang meningkat pula. Didapatkan hasil akurasi klasifikasi maksimal pada nilai 90.02% dengan menggunakan *variance* 100% atau nilai PC 58.

DAFTAR PUSTAKA

[1] R. W. R. M.D, *Cancer Biology Volume 2*. 2007.

[2] J. F. Holland, *Cancer Medicine*, 5th ed. Hamilton, Ontario: B.C. Decker Inc, 2000.

[3] J. Boultwood and C. Fidler, *Molecular Analysis of Cancer*. Totowa, New Jersey: Humana Press Inc., 2002.

[4] V. Trevino, F. Falciani, and H. A. Barrera-Saldana, "DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research," *Mol. Med.*, vol. 13, no. 9, pp. 527–541, 2007, doi: 10.2119/2006.

[5] L. A. Allison, *Fundamental Molecular Biology*. 2007.

[6] S. Knudsen, "Guide to Analysis of DNA Microarray Data," *Curr. Top. Med. Chem.*, vol. 4, no. 13, pp. 1355–1368, 2005, doi: 10.2174/1568026043387773.

[7] J. J. Gerbrands, "On the Relationships between SVD, KLT, and PCA," *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, vol. 3, no. 4, pp. 375–381, 1980, doi: 10.1109/iembs.1996.652703.

[8] S. Shanmuganathan and S. Samarasinghe, *Artificial Neural Network Modelling*.