Deteksi *Polycystic Ovarian Syndrome* (PCOS) Menggunakan Klasifikasi Microarray Data dengan Algoritma *Artificial Neural Network* (ANN) *Backpropagation* dan *Principal Component Analysis*

Tiara Laksmi Basuki[1], Jondri, M.Si.[2], Untari Novia Wisesty, S.T.,M.T.[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]Divisi Digital Service PT Telekomunikasi Indonesia
[1]tiaralb@students.telkomuniversity.ac.id, [2]jondri@telkomuniversity.ac.id, [3]untarinw@telkomuniversity.ac.id

Abstrak

PCOS (*polycystic ovary syndrome*) atau sindrom ovarium polikistik merupakan kondisi terganggunya fungsi ovarium pada wanita yang berada di usia subur. Kondisi ini menyebabkan hormon wanita yang menderita PCOS menjadi tidak seimbang karena hal-hal yang tidak diketahui. Penelitian ini bertujuan untuk membuat sistem klasifikasi menggunakan data yang berbentuk *microarray* karena berguna untuk menganalisis beribu-ribu sampel pada waktu bersamaan yang dapat membantu analisis dan diagnosis terhadap penyakit PCOS. Sistem klasifikasi akan terdiri dari tiga tahapan, yaitu *pre-processing* data dengan normalisasi, ekstraksi fitur dengan menggunakan Principal Component Analysis, dan klasifikasi menggunakan metode *Artificial Neural Network* yaitu Backpropagation, dan didapatkan hasil akurasi sebesar 50% - 100%.

Kata kunci:*PCOS, microarray, Principal Component Analysis, Artificial Neural Network*

Abstract

PCOS (polycystic ovary syndrome) or polycystic ovary syndrome is a condition of impaired ovarian function in women in childbearing age. This condition causes female hormones that suffer from PCOS to become unbalanced because of things that are not known. This study aims to make a classification system using data in the form of a microarray because it is useful to analyze thousands of samples at the same time which can help analysis and diagnosis of PCOS disease. The classification system will consist of three stages, namely pre-processing data with normalization, feature extraction using Principal Component Analysis, and classification using the Artificial Neural Network method, namely Backpropagation, and the results are 50%-100%.

Keywords:PCOS, microarray, Principal Component Analysis, Artificial Neural Network

1. Pendahuluan

Latar Belakang

Bioinformatika adalah ilmu yang mempelajari penerapan teknik komputasional untuk mengelola dan menganalisis informasi biologis. Bidang ini mencakup penerapan metode-metode matematika, statistika, dan informatika untuk memecahkan masalah-masalah biologis. Salah satunya analisis ekspresi gen, yang dapat dilakukan dengan identifikasi dan penyelidikan terhadap data microarray untuk mengetahui gambaran satu gen[4] Hasil analisis ekspresi gen dapat memprediksi suatu penyakit yang kemungkinan dapat diderita seseorang.

Salah satu penyakit berbahaya yang kurang kesadaran untuk mencegahnya adalah PCOS (*polycystic ovary syndrome*), atau disebut dengan sindrom ovarium polikistik, merupakan kelainan metabolik yang menyerang kaum perempuan di usia produktif (masa pubertas hingga pra menapouse). Kelainan tersebut menyebabkan adanya resistensi insulin, sehingga perempuan yang terkena PCOS akan mengalami gangguan ovulasi (pematangan sel telur) akibat kadar hormon yang tidak seimbang (estrogen dominan). Akibatnya akan terjadi gangguan menstruasi juga kesulitan untuk hamil apabila sudah aktif seksual.

PCOS disebut juga sebagai *silent killer* [1] karena jika tidak ditangani sesegera mungkin, resistensi insulin yang sudah terjadi bisa semakin parah sehingga bisa berkembang menjadi penyakit diabetes millitus, jantung koroner, kanker ovarium, tumor, kista, dan penyakit lainnya.

Penyebab sindrom ini antara lain pola hidup tidak sehat dan kelebihan insulin, akan tetapi terdapat bukti adanya kelainan genetik diwariskan oleh ibu atau ayah, atau mungkin keduanya. Gen tersebut bertanggung jawab atas terjadinya resistensi insulin dan hiperandrogenisme pada wanita dengan sindrom ovarium polikistik

Berdasarkan penelitian sebelumnya, klasifikasi PCOS dengan dataset biasa dan menggunakan algortima ANN, SVM, *classification tree*, dan *Naïve Bayes* hasil akurasinya [1] berturut-turut sebesar 83.70%, 76.45%, 75.25% dan 82.75%.

Pada tugas akhir ini *Artificial Neural Network Backpropagation* akan digunakan sebagai metode klasifikasi penyakit PCOS dengan menggunakan data *microarray*. Dan menggunakan *Principal Component Analysis* sebagai *Feature Selection* Pemilihan ini berdasarkan penelitian sebelumnya [1] dengan akurasi mencapai 83.70%, sehingga dapat diharapkan dengan menggunakan data berbentuk *microarray* dapat memperbesar tingkat akurasi.

Tujuan

Dalam penelitian Tugas Akhir ini bertujuan untuk mengimplementasikan algoritma Principal Component Analysis sebagai seleksi fitur serta menggunakan Artificial Neural Network dengan algoritma pembelajaran Backpropogation sebagai klasifikasi pada data ekspresi gen. Lalu agar dapat mengukur peformansi dan menganalisa akurasi dari penggunakan Principal Component Analysis sebagai seleksi fitur terhadap klasifikasi Artificial Neural Network.

2. Kajian Pustaka

2.1. *Polycystic ovarian syndrome (PCOS)*

Gangguan hormon sindrom ovarium polikistik atau Polycystic ovarian syndrome (PCOS) merupakan kondisi yang paling banyak menyebabkan wanita sulit hamil. Polycystic ovarian syndrome (PCOS) merupakan gangguan hormonal yang umum di kalangan wanita usia reproduksi. Wanita dengan PCOS memiliki banyak kista kecil yang terletak di sepanjang tepi luar dari masing-masing ovarium (indung telur). Hal inimenyebabkan tidak adanya ovulasi, sehingga menyulitkan wanita untuk mendapatkan keturunan. Tanda-tanda awal PCOS adalah masa ovulasi atau subur yang tidak beraturan, meningkatnya kadar hormon pria (androgen) dalam tubuh wanita, dan munculnya banyak kista (kantong berisi cairan) pada ovarium [8] Dalam beberapa dekade terakhir data microarray banyak di gunakan dalam diagnosis penyakit untuk meningkatkan akurasi diagnosis penyakit khususnya PCOS dibandingkan dengan data tradisional[4]. Microarray dapat digunakan untuk melihat tingkat gene expression dalam suatu sampel sel tertentu untuk menganalisis ribuan gen secara bersamaan.

2.2. Microarray

Microarray Expression merupakan suatu percobaan yang memungkinkan untuk menghitung tingkat ekspresi dari ribuan gen secara bersamaan. Percobaan ini memonitoring kondisi yang berbeda dari setiap gen yang bergantian mengevaluasi setiap gen dalam suatu jenis jaringan. Terutama terhadap PCOS. Klasifikasi untuk mendeteksi PCOS dapat dilakukan dengan mengelompokkan data ke dalam kelas yang sudah ditetapkan.

2.3. *Artificial Neural Network*

Artificial Neural Networks ini dikenal juga dengan Jaringan Saraf Tiruan (JST), adalah sebuah sistem komputasi yang mana arsitektur dan operasinya merupakan adopsi dari sistem kerja saraf pada otak manusia. salah satu permsalahannya yaitu klasifikasi. *Backpropagation* merupakan salah satu bagian dari *Neural Network*. *Backpropagation* merupakan metode pelatihan terawasi (supervised learning), dalam artian mempunyai target yang akan dicari[8] ciri dari *Backpropagation* adalah meminimalkan error pada output yang dihasilkan oleh jaringan. dalam metode backpropagation, biasanya digunakan jaringan multilayer.

Perhitungan maju :

a. Inisialisasi bobot dan bias secara acak
b. Menghitung nilai keluaran dari hidden layer:

$$V1 = X(W1 + B1) \quad (1)$$

Dengan X adalah input training data yang sudah reduksi W1 adalah bobot hidden dan B1 adalah bias hidden.

c. Menghitung nilai aktivasi setiap unit hidden sebagai output unit hidden dengan rumus :

$$A1 = \frac{1}{1+ e^{-V1}} \quad (2)$$

d. Setelah di dapatkan A1 hitung keluaran ouput layer :

$$V2 = W2(A1 + B2) \quad (3)$$

Dengan W2 adalah bobot output dan B2 adalah bias output

Lalu hitung Fungsi Aktivasi unit output sebagai output jaringan

$$A2 = \frac{1}{1+ e^{-V2}} \quad (4)$$

e. Langkah selanjutnya yaitu menghitung nilai error sebelum ke tahap perhitungan mundur, untuk mengetahui jumlah error

$$E = T - A2 \quad (5)$$

Dengan T adalah kelas target dari setiap inputan , dan hitung Mean Squared Error dengan rumus sebagai berikut

$$MSE = \frac{\sum E^2}{N} \quad (6)$$

Perhitungan mundur :

a. Perbaikan bobot dan bias :

$$D2 = (1 - A2^2)E \quad (7)$$
$$D1 = (1 - A2^2)(W2 * D2)$$
$$dW1 = lr * D1 * P$$
$$dB1 = lr * D1$$
$$dW2 = lr * D2 * A1$$
$$dB2 = lr * D2$$

b. Perbaikan bobot jaringan :

$$W1 = W'1 + dW1 \quad (8)$$
$$B1 = B'1 + dB1$$
$$W2 = W'2 + dW2$$
$$B1 = B'1 + dB1$$

2.4. Reduksi Dimensi

Teknik DNA *microarray* menunjukan dampak besar dalam menentukan gen-gen informatif penyebab PCOS. Namun, masih ditemui kelemahan utama dalam *microarray*, yaitu masalah dimensionalitas. Karena, data *microarray* memiliki jumlah fitur data yang jauh lebih banyak dibandingkan dengan jumlah baris data itu sendiri. Sering ditemukan microarray berdimensi tinggi sehingga perlu untuk di reduksi sehingga akan beban komputasi akan menjadi tidak stabil, oleh karena itu di butuhkan reduksi dimensi pada data *microarray* sebelum proses klasifikasi. Bertujuan untuk menghemat waktu komputasi serta menghindari *overvitting* pada classifier, Overfitting terjadi karena informasi yang diberikan terlalu spesifik untuk model yang dibuat atau tidak seimbang untuk keseluruhan range yang mungkin terjadi. Sehingga rasio kesalahan sangat kecil ketika dibandingkan dengan training set tersebut. Ini menjadi bahaya karena sebenarnya tidak realistis untuk memprediksi data yang belum diketahui. [3]

Metode reduksi dimensi data bekerja dengan cara tertentu untuk menangkap karakteristik data dengan memetakan set data dari dimensi semula ke dimensi lain yang relatif rendah. Pemetaan ini menghasilkan prinsipal componen yang kemudian dapat diambil komponen atau fitur dari dimensi baru yang mempunyai pengaruh yang besar pada set data dan membuang data yang tidak berpengaruh. salah satu metode yang sudah digunakan secara luas adalah *Principal Component Analysis.*

2.4.1 Principal Component Analysis

Principal component analysis atau yang sering disingkat sebagai *PCA* ini merupakan metode yang dapat mereduksi dimensi-dimensi pada suatu kumpulan besar data, sehingga ukuran dimensinya akan mengecil. Metode fitur reduksi dari *PCA* dapat dijelaskan sebagai berikut :

a. Menghitung standarisasi data dengan Z-score :

$$Z = \frac{Xi - \mu}{\sigma} \quad (9)$$

Dimana Xi merupakan data ke-i, μ adalah rata-rata dan σ adalah standar deviasi.

b. Mencari kovariansi dari data :

$$C = \sum_{k=1}^{n} (X_k - \mu)(X_k - \mu)^T \quad (10)$$

c. Menghitung nilai *eigen* dan vektor *eigen*

$$CU_n = \lambda_n U_n \quad (11)$$

d. Pilih beberapa vektor *eigen* dengan nilai *eigen* tertinggi

e. Lakukan tranformasi data menggunakan vektor *eigen* yang telah dipilih

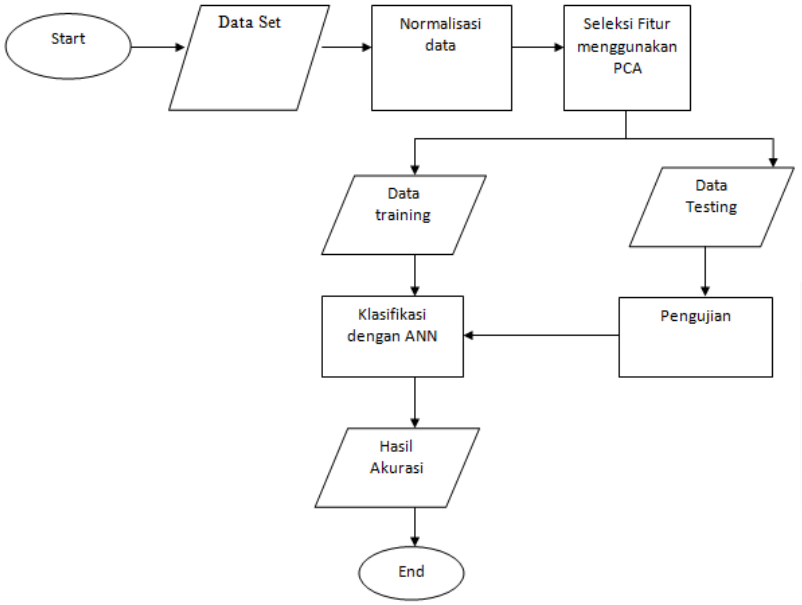
2.5 Akurasi

Akurasi deteksi PCOS dihitung dengan menggunakan rumus :

$$\text{Akurasi (\%)} = \frac{Jumlah\ data\ yang\ dikenali\ secara\ benar}{Jumlah\ data\ yang\ diujikan}\ x\ 100\%.$$

3. Perancangan Sistem

Rancangan sistem yang dibangun berupa tahapan penelitian yang terstruktur yang bertujuan untuk menghasilkan prediksi dengan tingkat akurasi yang baik pada PCOS. Sistem yang dibangun dengan mengimplementasikan metode PCA sebagai seleksi fitur dan *Artificial Neural Network* sebagai klasifikasi ekspresi gen. Berikut merupakan *flowchart* pengerjaan Tugas Akhir ini :



Gambar 3.1 Gambaran umum system

Sistem yang akan dibuat terbagi menjadi 3 tahap utama, yaitu *preprocessing* dengan normalisasi data, ekstrasi ciri dan klasifikasi. Pertama, data yang diambil dari [9] diolah terlebih dahulu, setelah dilakukan *preprocessing* maka dilakukan seleksi fitur dengan menggunakan PCA agar dan yang terakhir proses klasifikasi menggunakan metode Artificial Neural Network dengan memperhitungkan tingkat akurasi berdasarkan hasil dari klasifikasi tersebut.

3.1.Data Set

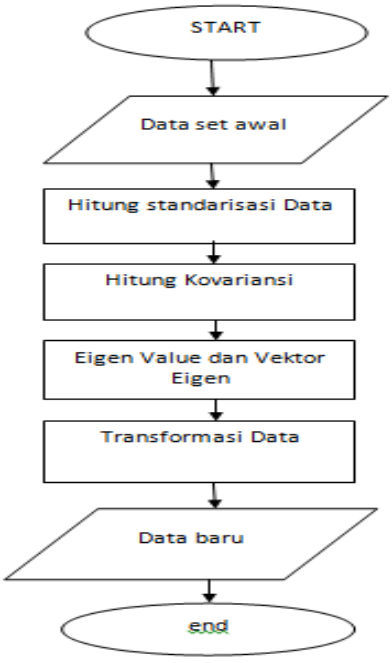
Data set yang digunakan pada tugas akhir ini adalah data ekspresi gen berupa data PCOS yang diperoleh dari *National Center for Biotechnology Information Repository* [9]. Data ekspresi gen yang akan digunakan berupa data penyakit PCOS yang memiliki 28 *record* dan 16000 atribut.

Tabel 3.1 Tabel Dataset PCOS

No.	Atribut 1	Atribut 2	Atribut 3	Atribut 4	Atribut 5	...	Atribut 16000	Kelas
1	0.426693	0.079051	-0.03711	0.102249	0.031652	...	0.102643	1
2	0.047176	0.101923	0.081573	1.48231	0.036219	...	0.053397	1
3	0.061133	0.273007	0.151132	-1.08528	0.033979	...	-0.03329	1
4	0.417477	0.197721	0.455904	0.665998	0.034796	...	0.032564	1
5	-0.04499	0.211543	0.362855	0.987918	0.047671	...	0.132632	1
...	...	...	...	...	...	...	...	...
28	0.280311	0.125371	0	-0.28384	0.030195	...	0.420758	2

3.2 *Feature Selection* PCA

Berikut merupakan alur dari pengerjaan seleksi fitur yang dilakukan pada Tugas akhir ini. Dari dataset awal pertama akan diseleksi fitur dengan algoritma *Principal Component Analysis* untuk mereduksi dimensi atribut dan kemudian menjadi data baru, lalu data baru akan dimasukan ke dalam klasifikasi.

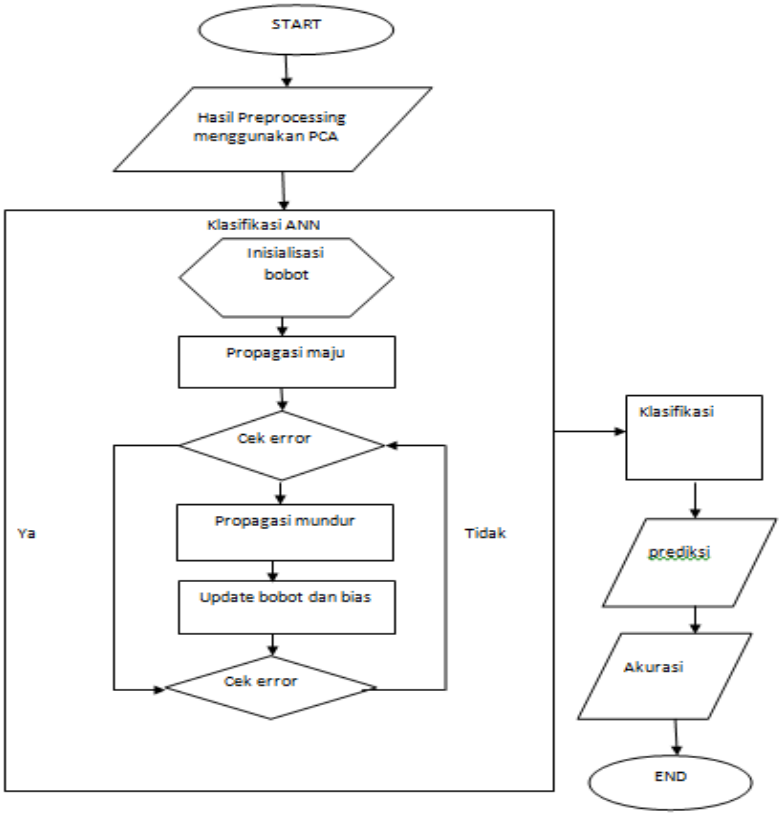


Gambar 3.2 Flowchart Seleksi Fitur

PCA digunakan untuk mereduksi dimensi tanpa mengurangi karakteristik data secara signifikan. Metode ini mengubah sebagian besar variabel asli yang berkolerasi menjadi satu himpunan variabel baru yang lebihkecil dan saling bebas. Keluaran dari PCA menghasilkan sekumpulan dimensi baru yang dinamakan Principal Component (PC).

3.3.Klasifikasi ANN

Setelah mengurangi kompleksitas dimensi data *microarray*, hal selanjutnya dilakukan yaitu proses klasifikasi dimana pada tahap ini akan dilakukan diagnosis seseorang mengidap PCOS atau tidak berdasarkan data microarray.



Gambar 3.3 Flowchart Klasifikasi ANN

Pada klasifikasi ini digunakan arsitektur Multi Layer Perceptron dengan 28 inputan, 7 neuron layer tersembunyi dan 2 buah neuron output.

3.4.Skenario

Pada Tugas akhir ini dilakukan 7 skenario dimana masing-masing skenario ini memiliki perbandingan pembagian data training dan testing, serta parameter *Artifial Neural Network* yang berbeda-beda. Berikut merupakan 7 skenario tersebut:

Tabel 3.2 Tabel Skenario

Skenario	Seleksi Fitur	Pembagian Data		Parameter ANN	
	PCA	Training	Testing	Epoch	Learning rate
1	YA	80	20	300	0.15
2		80	20	600	0.15
3		80	20	600	0.05
4		70	30	300	0.1
5		70	30	600	0.1

6		60	40	300	0.05
7		60	40	600	0.05

4. Evaluasi

4.1 Hasil Pengujian

Tabel 4.1 Hasil akurasi skenario

Skenario	Seleksi Fitur	Pembagian Data		Parameter ANN		Akurasi
	PCA	Training	Testing	Epoch	Learning rate	
1	YA	80	20	300	0.15	80%
2		80	20	600	0.15	60%
3		80	20	600	0.05	100%
4		70	30	300	0.1	62.5%
5		70	30	600	0.1	50%
6		60	40	300	0.05	90.09%
7		60	40	600	0.05	54.54%

4.1.1 Skenario Seleksi Fitur dan Skenario Pembagian Data

Dilakukan tiga pembagian skenario dalam seleksi fitur yang dimana semua skenario menggunakan PCA karena dimensi data yang sangat tinggi tidak memungkinkan, semua skenario menggunakan data yang sudah di reduksi oleh PCA dan hasilnya

a. Pembagian data dengan perbandingan 80% data training dan 20% data testing didapatkan akurasi paling besar 100%.
b. Pembagian data dengan perbandingan 70% data training dan 30% data testing didapatkan akurasi paling besar 62,5%.
c. Pembagian data dengan perbandingan 60% data training dan 40% data testing didapatkan akurasi sebesar 90,09%.

4.1.2 Skenario Parameter ANN

Dilakukan empat skenario pada bagian parameter ANN yaitu :

a. Parameter ANN yang digunakan ialah epoch =300 dan learning rate = 0.15 pada skenario ini didapatkan hasil akurasi 80 %.
b. Parameter ANN yang digunakan ialah epoch =300 dan learning rate = 0.1 pada skenario ini didapatkan hasil akurasi 62.5 %.
c. Parameter ANN yang digunakan ialah epoch =300 dan learning rate = 0.05 pada skenario ini didapatkan hasil akurasi 90.09 %.
d. Parameter ANN yang digunakan ialah epoch =600 dan learning rate = 0.15 pada skenario ini didapatkan hasil akurasi 60 %.
e. Parameter ANN yang digunakan ialah epoch =600 dan learning rate = 0.05 pada skenario ini didapatkan hasil akurasi 54.54% dan 100%.
f. Parameter ANN yang digunakan ialah epoch =600 dan learning rate = 0.1 pada skenario ini didapatkan hasil akurasi 40 %.

4.2. Analisis Hasil Pengujian

Dari hasil pengujian skenario diatas terdapat perbandingan yaitu : skenario 1, 2, dan 3 menggunakan seleksi fitur berupa PCA dan menggunakan 80% data training dan 20% data testing

mendapatkan hasil maksimal akurasi yaitu 100% apabila epoch yang di gunakan besar dalam skenario ini memakai epoch 600 dan learning rate yang di pakai 0.05, skenario 4 dan 5 menggunakan 70% data training dan 30% data testing mendapatkan hasil akurasi 62.5% dan 50% memakai epoch 300 dan 600, untuk skenario 6 dan 7 memakai data training 60% dan data testing

40% mendapatkan hasil akurasi 90.09% jika epoch yang digunakan sebesar 300 dan learning rate 0.05, dan 54.54% jika memakai epoch 600 dan learning rate yang sama yaitu 0.05.

Dari perbandingan ke tujuh skenario yang telah dilakukan hasil akurasi yang terbesar didapatkan dari skenario 3 dan 6 yaitu 90.09% dan juga 100% menggunakan learning rate sebesar 0.05, learning rate sendiri berpengaruh pada hasil akurasi karena semakin besar learning rate maka ketelitian jaringan akan semakin berkurang, tetapi berlaku sebaliknya. Apabila learning rate-nya semakin kecil, maka ketelitian jaringan akan semakin besar, harus disesuaikan dengan epoch yang digunakan.

5. Kesimpulan

Berdasarkan dari hasil pengujian yang telah dilakukan maka didapatkan kesimpulan sebagai berikut :

1. Algoritma Artificial Neural Network sebagai klasifikasi dan Principal Component Analysis sebagai seleksi fitur mampu memprediksi dengan baik pada jenis data PCOS menghasilkan akurasi sebesar 100% .
2. PCA mampu menyeleksi fitur dari data ekspresi gen yang memiliki 16000 atribut sehingga terpilih beberapa atribut terbaik yang akan digunakan dalam klasifikasi.
3. Dengan menggunakan Principal Component Analysis sebagai seleksi fitur pada klasifikasi ANN, nilai bobot dari data training juga data testing, learning rate dan epoch membuat akurasi dalam klasifikasi karena hasil akurasi yang didapatkan lebih tinggi.

Daftar Pustaka

[1] K. Meena1, M. Manimekalai, S. Rethinavalli, 2015, *Correlation Of Artificial Neural Network Classification And Nfrs Attribute Filtering Algorithm for Pcos Data*. *India* : *Department of Computer Applications, Shrimati Indira Gandhi College*

[2] Ermatita, Sri Hartati, Retantyo Wardoyo, Agus Harjoko,2010, Medical Imaging Untuk Analisis Ekspresi Gen Dalam Deteksi Penyakit Tahun, Yogyakarta

[3] Dr. K. Meena,Dr. M. Manimekalai,2015, *Implementing Neural Fuzzy Rough Set and Artificial Neural Network forPredicting PCOS*. India: Department of Computer Application,Shrimati Indira Gandhi College

[4] "*What is Bioinformatics*" [online] Available : http://fatchiyah.lecture.ub.ac.id/teaching-responsibility/bioinformatics/whats-bioinformatics/[Diakses tanggal 19 November 2017]

[5] "Sindrom Ovarium dan Kesuburan" [online] Available: http://www.ledisia.com/program-hamil/waspadai-gejala-sindrom-ovarium-polikistik-yang-dapat-merusak-kesuburan-anda/ [Diakses 26 November 2017]

[6] Pritam Kumar Panda, Riya Rane,Rahul Ravichandran, Shrinkhla Singh, Hetalkumar. 2015,*Genetics of PCOS: A systematic bioinformatics approach to unveil the proteins responsible for PCOS*, India : *School of Biotechnology and Bioinformatics, D. Y. Patil University, CBD Belapur, Navi Mumbai, Maharashtra, India*

[7] Nurfalah, A. Adiwijaya. and Suryani, A., 2016. *Cancer Detection Based On Microarray Data Classification Using Pca And Modified Back Propagation.* Far East Journal of Electronics and Communications, 16(2), p.269.

[8] Suyanto, 2008, "Soft Computing: Membagun Mesin Ber-IQ Tinggi", Bandung, Informatika.

[9] PCOS Dataset Source -ftp://ftp.ncbi.nlm.nih.gov/geo/datasets/GDS4nnn/GDS4987/