

MULTI LAYER PERCEPTRON DAN PRINCIPAL COMPONENT ANALYSIS UNTUK DIAGNOSA KANKER PAYUDARA

Made Satria Wibawa^{1,a}, I Md. Dendi Maysanjaya^{2,b}

¹Sistem Komputer, STMIK STIKOM Bali, Bali, Indonesia - 80226

²Pendidikan Teknik Informatika, Universitas Pendidikan Ganesha, Bali, Indonesia - 81116
e-mail: satria.wibawa@stikom-bali.ac.id^a, dendi.ms@undiksha.ac.id^b

Abstrak

CAD (*Computer Aided Diagnosis*) merupakan teknik diagnosa berbantuan komputer untuk meningkatkan akurasi hasil diagnosa dari suatu penyakit. CAD telah banyak digunakan untuk diagnosa dari berbagai penyakit, khususnya penyakit kanker payudara. *Multi layer perceptron* (MLP) sebagai salah metode dari jaringan saraf tiruan telah banyak digunakan untuk klasifikasi kanker payudara. Penelitian ini bertujuan untuk mencari kombinasi parameter paling optimal untuk mendiagnosa kanker payudara. Kombinasi parameter tersebut juga diujikan dengan metode reduksi fitur *Principal Component Analysis* (PCA). Hasil penelitian menunjukkan bahwa parameter paling optimal adalah fungsi optimisasi RELU serta TANH dengan fitur optimisasi adam dengan tingkat akurasi 0.974.

Kata kunci: CAD, kanker payudara, MLP, PCA, RELU, TANH, adam

Abstract

CAD (*Computer Aided Diagnosis*) is a computer-aided diagnosis techniques used to improve the accuracy diagnosis disease result. CAD has been widely used for diagnosis of various diseases, especially cancer of the breast. *Multi layer perceptron* (MLP) as the method of artificial neural network has been widely used for classification of breast cancer. This research aims to find the most optimal combinations of parameters to diagnose breast cancer. The combination of these parameters were also examined by the method of reduction of *Principal Component feature Analysis* (PCA). The results showed that the most optimal parameter is a function WITHIN a RELU optimization and optimization feature adam 0973 accuracy.

Keywords : CAD, breast cancer, MLP, PCA, RELU, TANH, adam

PENDAHULUAN

Jaringan syaraf tiruan (*artificial neural network*) telah menjadi salah satu metode yang sering digunakan pada proses klasifikasi. Banyaknya penelitian di bidang klasifikasi menggunakan jaringan syaraf tiruan menjanjikan metode ini dapat dipilih sebagai alternatif untuk mengganti metode klasifikasi konvensional(1)(2).

Jaringan syaraf tiruan memiliki beberapa kelebihan dibandingkan metode klasifikasi konvensional lainnya. Yang pertama, jaringan syaraf tiruan merupakan metode yang mengadaptasikan dirinya dengan data, sehingga tidak diperlukan adanya spesifikasi bentuk fungsional dan distribusional model yang eksplisit. Kedua,

jaringan syaraf tiruan fungsi universal untuk memperkirakan hubungan antar keanggotaan kelas dengan atribut dari objek. Ketiga, jaringan syaraf tiruan umumnya lebih reliabel terhadap *noise* dalam data. Yang terakhir, jaringan syaraf tiruan mampu menghitung nilai probabilitas posterior, yang merupakan dasar untuk membangun aturan klasifikasi dan analisis statistika. Salah satu metode dalam jaringan syaraf tiruan yang luas digunakan adalah *multilayer perceptron* (MLP). MLP telah banyak diterapkan untuk berbagai bidang, salah satunya adalah bidang kesehatan, terutama diagnosa penyakit kanker(3).

Kanker merupakan salah satu penyebab utama morbiditas dan kematian di seluruh dunia dengan angka mencapai 14 juta lebih kasus. Kanker merupakan penyebab kematian nomor dua di seluruh dunia dan pada tahun 2015 angka kematian karena kanker mencapai 8,8 juta lebih. Secara global, penyebab 1 dari 6 kematian adalah penyakit kanker (4). Beberapa jenis kanker yang paling umum ditemui adalah kanker paru-paru, hati (liver), perut dan payudara. Kanker payudara merupakan penyakit kanker dengan persentase kasus baru tertinggi, yaitu 43,3% dan persentase kematian akibat kanker payudara sebesar 12,9%(5).

Menurut *World Health Organization* (WHO), deteksi dini dapat meningkatkan peluang penanganan yang sukses dalam kasus penyakit kanker. Diagnosa kanker payudara dapat dilakukan dengan cara biopsi atau metode screening. Metode screening meliputi pemeriksaan fisik, MRI, dan mamografi. Jika hasil screening inkonklusif, maka biopsi melalui analisa mikroskopis seperti *Fine Needle Aspiration* (FNA) dapat digunakan untuk menentukan hasil diagnosa. Biopsi merupakan cara

METODE

Metode memberikan penjelasan tentang langkah-langkah, data, lokasi penelitian, metode evaluasi yang digunakan serta penjelasan terstruktur tentang algoritma atau metode dari penelitian yang dibahas.

1. Data

Klasifikasi yang dilakukan adalah klasifikasi biner dengan kelas tumor jinak (*benign*) dan tumor ganas (*malignant*). Jumlah data yang digunakan adalah sebanyak 569 data dengan 357 data kelas *benign* dan 212 kelas *malignant*(8). Fitur merupakan properti pengukuran terhadap suatu proses/objek yang sedang diamati. Berdasarkan kumpulan fitur tersebut, metode *machine learning* dapat melakukan klasifikasi. Fitur biasanya disusun oleh ahli di domain dimana *machine learning* diterapkan. Pada penelitian ini, fitur yang digunakan adalah beberapa parameter

yang paling definitif dalam diagnosa kanker payudara diantara metode diagnosis lainnya(6).

Untuk membantu dalam diagnosa penyakit kanker khususnya kanker payudara, *Computer-Aided Diagnosis* (CAD) dapat digunakan. CAD telah secara luas digunakan untuk mendeteksi dan diagnosa dalam kanker payudara. Oleh karena itu, peningkatan kinerja dari CAD dalam bentuk akurasi, sensitivitas dan spesifisitas telah menjadi salah satu bidang riset yang paling utama dan penting. Kelebihan utama dalam penerapan penggunaan komputer untuk membantu diagnosis medis terletak pada pengolahan data mining yang mampu mengekstrak pola tertentu dari data. Diagnosis dengan bantuan komputer jika dilakukan dengan penanganan yang tepat memiliki potensi untuk menangkap keahlian interpretasi dari seorang pakar. Sehingga dapat meningkatkan akurasi diagnosis dan tingkat keyakinan seorang pakar(7).

Penelitian ini bertujuan untuk melakukan diagnosa kanker payudara menggunakan CAD dengan metode MLP.

pengukuran citra jaringan sel tumor yang diambil dari proses FNA.

Penjelasan mengenai fitur dapat dilihat pada Tabel I. Terdapat 10 pengukuran tekstur dan kontur yang diambil dari citra FNA. Sepuluh fitur tersebut adalah *radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension*. Dari 10 utama tersebut diambil lagi 3 ciri statistiknya, yaitu rerata, standar deviasi dan nilai terbesar, sehingga total fitur berjumlah 30 buah.

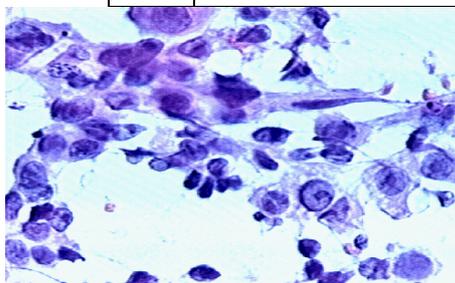
Jumlah fitur yang dimiliki oleh dataset tersebut cukup banyak. Mengingat semakin banyaknya jumlah fitur, maka jumlah neuron pada *neural network* akan semakin banyak. Hal ini akan menyebabkan penghitungan bobot yang semakin banyak dan membutuhkan waktu yang lama untuk mencapai konvergensi. Karena itu diperlukan suatu teknik untuk mengurangi/mereduksi fitur agar waktu

komputasi pada saat training menjadi lebih sedikit. Selain itu, fitur yang banyak juga

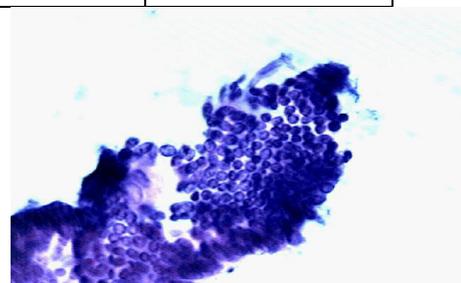
belum tentu dapat memberikan kontribusi terhadap klasifikasi yang maksimal.

Tabel 1. Fitur Data

NO	FITUR	KETERANGAN	TIPE DATA
1	Radius	Jarak rata-rata dari titik pusat citra ke perimeter	Real
2	Texture	Standar deviasi dari nilai keabuan pada citra	Real
3	Perimeter	Panjang dari garis tepi pada sel	Real
4	Area	Luas dari penampang sel	Real
5	Smoothness	Variasi lokal pada panjang radius	Real
6	Compactness	$\frac{perimeter^2}{area} - 1.0$	Real
7	Concavity	Banyaknya bagian concave pada kontur sel	Real
8	Concave points	Jumlah bagian yang concave pada kontur sel	Real
9	Symmetry	Tingkat kesamaan belahan sisi kanan dan kiri sel	Real
10	Fractal dimension	Rasio kompleksitas dari sebuah pola	Real



(a) Tumor Ganas



(b) Tumor Jinak

Gambar 1. Contoh Citra Tumor Payudara FNA

Contoh citra tumor payudara ganas dapat dilihat pada Gambar 1a, sedangkan untuk tumor payudara jinak dapat dilihat pada Gambar 1b.

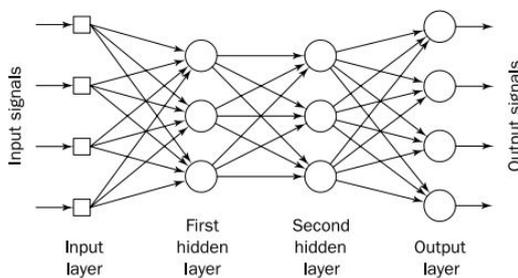
2. PCA

Principal Component Analysis (PCA) merupakan salah satu teknik untuk mereduksi dimensi dari suatu data. Tidak seperti metode seleksi fitur yang mengurangi jumlah fitur dengan cara menghilangkan fitur yang dianggap tidak penting tanpa membentuk fitur baru, PCA mengurangi dimensi data dengan cara 'mengkombinasikan' intisari dari atribut dengan membentuk alternatif subset fitur yang lebih kecil.

Jadi, pada metode PCA terbentuk fitur yang baru. Jumlah principal component dapat berkisar 1-n, dimana n adalah jumlah fitur. Namun, tentu saja jumlah principal component sebaiknya paling banyak sebesar n-1, karena tujuan awal adalah untuk mengurangi waktu komputasi (9)(10)(11).

3. MLP

Multilayer Perceptron atau yang biasa disingkat MLP adalah salah satu jenis dari *feed-forward neural network* dengan satu atau lebih *hidden layer*. Pada umumnya MLP terdiri dari layer input yang merupakan kumpulan neuron untuk memasukkan data; minimal satu *hidden layer* sebagai neuron komputasi dan layer output sebagai neuron penampung hasil komputasi.

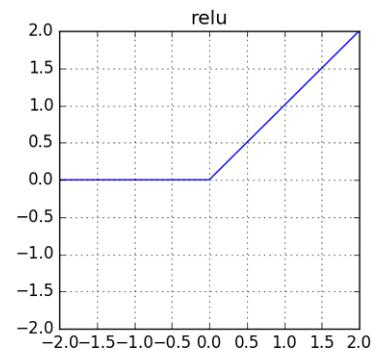


Gambar 2 Arsitektur MLP Standar

Dalam MLP, terdapat dua parameter penting yaitu fungsi aktivasi dan fungsi optimisasi. Fungsi aktivasi menentukan output pada sebuah node dari beberapa elemen input. Sedangkan fungsi optimisasi berfungsi untuk menentukan bobot paling tepat sesuai dengan input dan output.

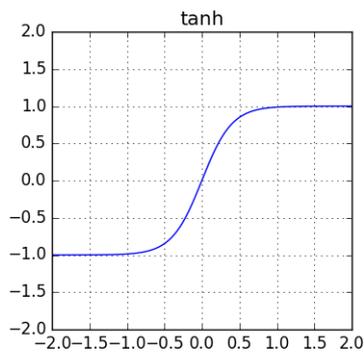
Pada penelitian ini, terdapat dua buah fungsi aktivasi dan tiga buah fungsi optimisasi yang diujikan. Kedua fungsi aktivasi itu adalah *Rectified Linear Unit* (RELU) dan fungsi aktivasi *Hyperbolic Tangent Function* (TANH). Sedangkan untuk ketiga fungsi optimisasi yang digunakan adalah fungsi *Limited-memory Broyden-Fletcher-Goldfarb-Shanno* (L-BFGS), *Stochastic Gradient Descent* (SGD) dan Adam.

Fungsi aktivasi *Rectified Linear Unit* (RELU) merupakan fungsi aktivasi yang memiliki perhitungan yang sederhana. Proses *forward* dan *backward* dengan RELU menggunakan kondisi *if*. Tidak ada operasi eksponensial, perkalian atau pembagian pada RELU. Jika elemen bernilai negatif, maka nilainya di-set menjadi 0. Cara perhitungan yang sederhana seperti itu membuat memiliki waktu komputasi yang relatif sedikit pada saat proses *training* dan *testing*.



Gambar 3 Grafik Fungsi ReLU

Fungsi aktivasi *Hyperbolic Tangent Function* (TANH) umumnya memiliki performa yang lebih baik dibandingkan fungsi aktivasi sigmoid dan logistic. Hal ini dikarenakan fungsi TANH dapat mencapai konvergensi lebih cepat dibandingkan fungsi sigmoid dan logistic. Performa fungsi TANH hampir sama dengan performa fungsi RELU.



Gambar 4 Grafik Fungsi TANH

Sama seperti metode optimasi lainnya, *Stochastic Gradient Descent* bertujuan untuk mencari global minima dari *cost function*. Fungsi optimisasi *Limited-memory Broyden-Fletcher-Goldfarb-Shanno* atau disingkat L-BFGS termasuk ke dalam metode quasi-Newton yang menggunakan algoritma BFGS. Disebut limited karena menggunakan komputer memori yang terbatas. Karena metode Newton sulit dan memerlukan waktu komputasi yang lama maka dikembangkanlah L-BFGS yang memiliki keunggulan dimana L-BFGS memerlukan memori yang lebih sedikit.

HASIL DAN PEMBAHASAN

Nilasi akurasi dari semua skema klasifikasi MLP dengan seleksi fitur PCA dapat dilihat secara lengkap pada Tabel II dan pergerakan nilai terhadap perubahan jumlah dimensi PCA dapat dilihat pada Gambar 5.

Nilai akurasi tertinggi yang dicapai per skema klasifikasi diberi tanda kuning, sedangkan nilai akurasi terendah diberi tanda merah. Pada metode ReLU & SGD, akurasi tertinggi yang dicapai adalah 0.854, sedangkan nilai akurasi terendah adalah 0.744. Akurasi tertinggi pada metode ini

dicapai pada saat penggunaan 2 dimensi PCA, sedangkan akurasi terendah diperoleh saat menggunakan 7 dimensi PCA.

ReLU dan L-BFGS mencapai akurasi tertinggi saat menggunakan 10 dimensi PCA, nilai akurasi yang diperoleh adalah 0.967. Akurasi terendah diperoleh saat menggunakan 1 dimensi PCA, yaitu dengan nilai 0.932.

Akurasi tertinggi yang dicapai dengan metode ReLU & Adam adalah 0.974 dengan penggunaan 8 dimensi PCA. Sama dengan ReLU & L-BFGS, nilai akurasi terendah diperoleh saat menggunakan 1 dimensi PCA. Nilai akurasi terendah yang diperoleh adalah 0.932.

Metode TANH & L-BFGS menghasilkan akurasi dengan rentang diatas 0.800, namun tidak mencapai 0.9. Akurasi tertinggi yang dicapai adalah 0.886 dengan penggunaan 5 dimensi PCA. Sedangkan untuk akurasi terendah diperoleh saat penggunaan 20 dimensi PCA, dengan nilai sebesar 0.830.

TANH dan L-BGS rata-rata memberikan nilai akurasi di atas 0.9. Akurasi tertinggi dicapai saat penggunaan 14 dimensi PCA, yaitu dengan akurasi sebesar 0.968. Akurasi terendah diperoleh saat penggunaan 1 dimensi, yaitu dengan nilai sebesar 0.933.

Metode terakhir, yaitu TANH & Adam, menghasilkan akurasi tertinggi sebesar 0.974. Akurasi tertinggi tersebut dicapai saat penggunaan 8 dimensi PCA. Akurasi terendah yang dicapai sangat jauh berbeda dibandingkan akurasi tertinggi. Akurasi terendah diperoleh saat penggunaan 23 dimensi PCA, dengan nilai sebesar 0.765.

Tabel II Nilai Akurasi pada Setiap Skema Klasifikasi

JUMLAH DIMENSI PCA	ReLU & SGD	ReLU & L-BFGS	ReLU & Adam	TANH & SGD	TANH & L-BFGS	TANH & Adam
1	0.809	0.932	0.932	0.860	0.933	0.932
2	0.854	0.942	0.946	0.874	0.947	0.944
3	0.821	0.945	0.954	0.863	0.939	0.960

JUMLAH DIMENSI PCA	ReLU & SGD	ReLU & L-BFGS	ReLU & Adam	TANH & SGD	TANH & L-BFGS	TANH & Adam
4	0.844	0.954	0.965	0.874	0.937	0.965
5	0.846	0.963	0.967	0.886	0.956	0.965
6	0.831	0.944	0.965	0.874	0.944	0.965
7	0.744	0.965	0.972	0.835	0.967	0.968
8	0.830	0.958	0.974	0.877	0.963	0.974
9	0.837	0.952	0.974	0.854	0.956	0.974
10	0.775	0.967	0.974	0.858	0.965	0.968
11	0.809	0.954	0.968	0.863	0.967	0.968
12	0.789	0.967	0.972	0.847	0.961	0.970
13	0.830	0.960	0.967	0.872	0.960	0.968
14	0.833	0.961	0.967	0.865	0.968	0.968
15	0.830	0.951	0.970	0.860	0.958	0.967
16	0.810	0.963	0.970	0.875	0.961	0.965
17	0.842	0.956	0.968	0.847	0.963	0.970
18	0.816	0.960	0.974	0.854	0.963	0.974
19	0.844	0.965	0.970	0.868	0.960	0.968
20	0.768	0.967	0.958	0.830	0.963	0.961
21	0.795	0.956	0.967	0.849	0.961	0.965
22	0.810	0.958	0.965	0.858	0.958	0.961
23	0.802	0.953	0.963	0.858	0.961	0.765
24	0.788	0.960	0.965	0.858	0.963	0.965
25	0.782	0.961	0.968	0.844	0.961	0.967
26	0.781	0.958	0.967	0.842	0.963	0.967
27	0.828	0.956	0.963	0.860	0.958	0.963
28	0.831	0.947	0.967	0.868	0.956	0.965
29	0.758	0.954	0.963	0.847	0.961	0.963

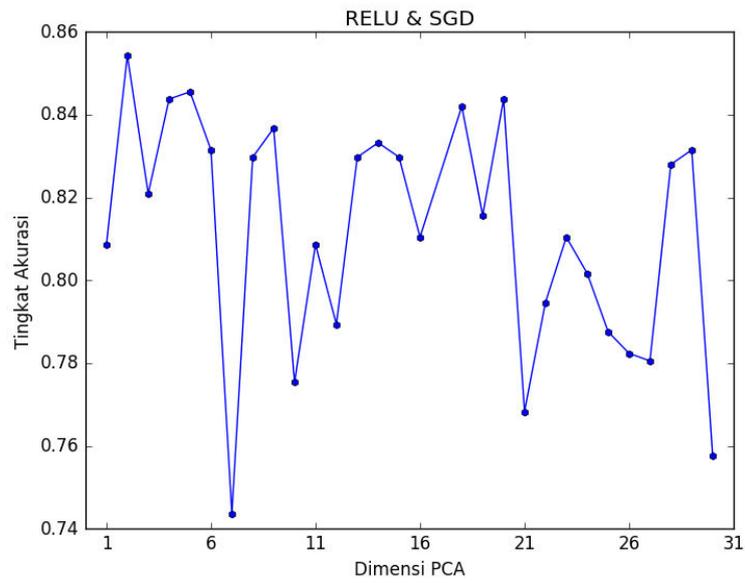
Seperti yang terlihat pada Tabel II, nilai akurasi tertinggi diperoleh pada metode ReLU & Adam. Nilai tertinggi tersebut sebesar 0.974 dan dicapai saat menggunakan 8 dimensi PCA. Akurasi terendah di antara semua metode diperoleh pada penggunaan metode ReLU & SGD. Nilai terendah yang diperoleh adalah 0.744 pada penggunaan 7 dimensi PCA.

Jika dilihat dari pola nilai akurasi yang terbentuk, fungsi optimasi SGD memberikan hasil yang relatif lebih rendah dibandingkan L-BFGS dan Adam pada kedua fungsi aktivasi ReLU dan TANH.

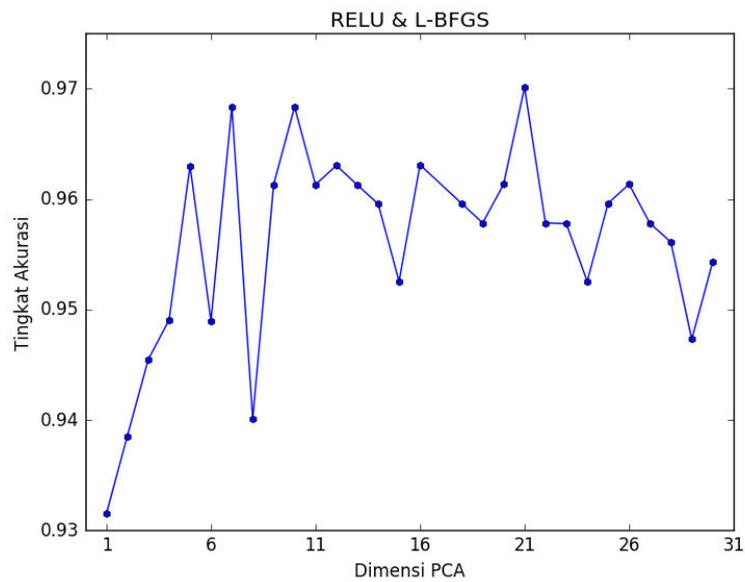
Gambar 5 menjelaskan pengaruh jumlah dimensi terhadap nilai akurasi

keenam metode klasifikasi. Skema ReLU & SGD, ReLU & L-BFGS, ReLU & Adam, TANH & SGD, TANH & L-BFGS dan TANH & Adam secara berturut-turut digambarkan pada Gambar 5(a), (b), (c), (d), (e) dan (f). Jika dilihat dari pergerakan grafik, nilai akurasi skema klasifikasi pada Gambar (a), (b) dan (d) fluktuatif. Tidak ada pola naik atau turun, penambahan atau pengurangan jumlah dimensi PCA tidak berhubungan dengan nilai akurasi. Sedangkan pada Gambar (c) dan (d) nilai akurasi rata-rata cenderung bergerak naik seiring penambahan jumlah dimensi PCA. Sedangkan pada Gambar 5(f), yaitu metode TANH & Adam, nilai akurasi hanya

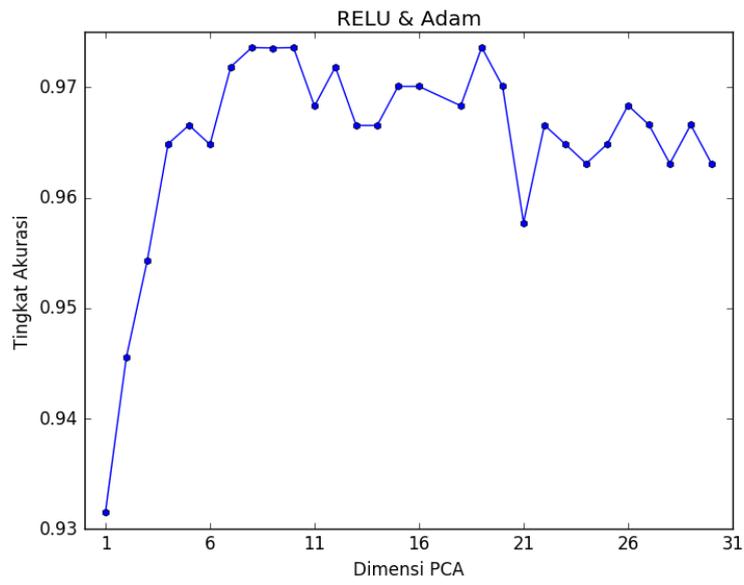
berkisar di rentang nilai 0.9. Namun jatuh pada saat penggunaan 23 dimensi PCA.



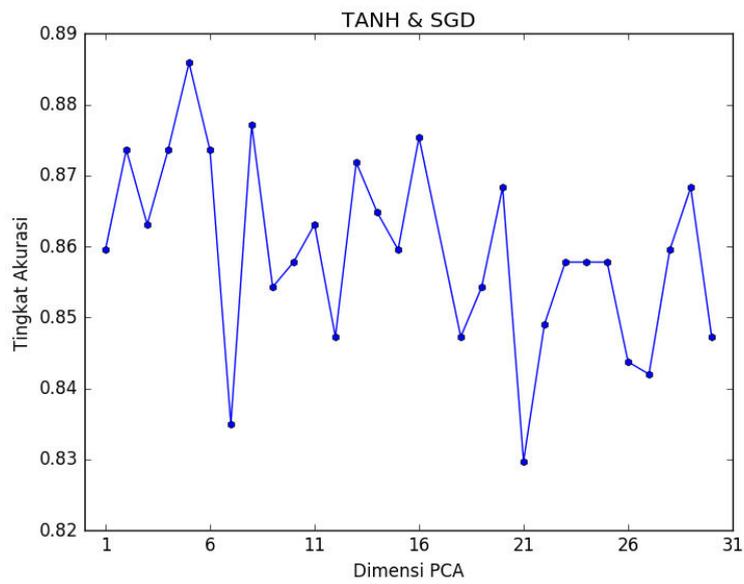
(a)



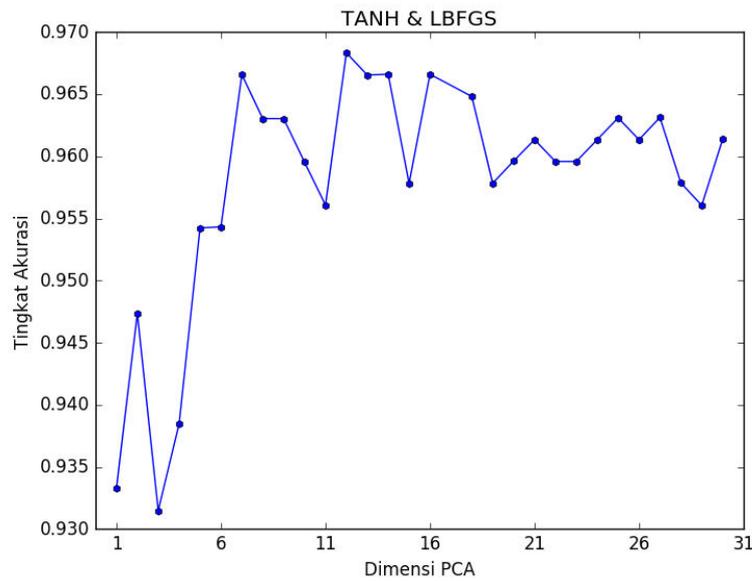
(b)



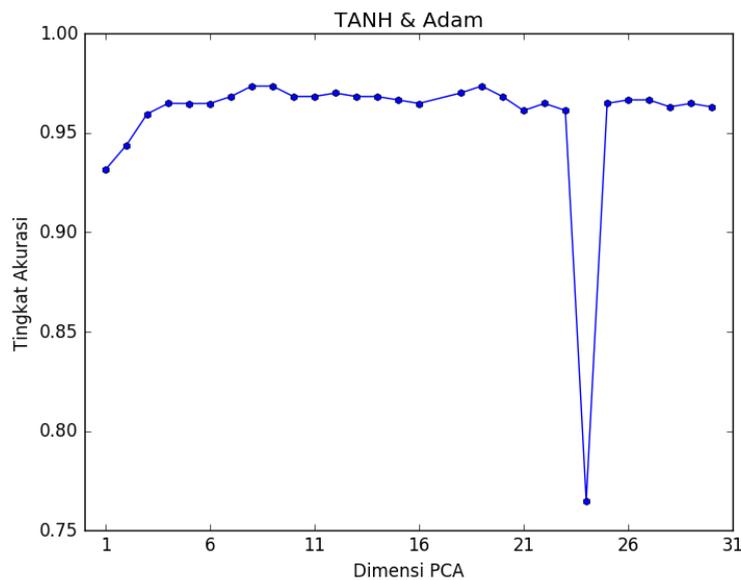
(c)



(d)



(e)



(f)

Gambar 5 Grafik Pengaruh PCA Terhadap Akurasi Metode (a)ReLU & SGD, (b)ReLU & L-BFGS, (c)TANH & SGD, (d)TANH & L-BFGS, (e) TANH & Adam

SIMPULAN

Penelitian ini bertujuan untuk mendiagnosa penyakit kanker payudara dengan diagnose berbantuan computer (*Computer Aided Diagnosis*). Metode yang digunakan adalah MLP sebagai klasifier dan PCA sebagai metode reduksi fitur. Kombinasi fungsi aktivasi, fungsi optimisasi dan jumlah dimensi PCA diteliti untuk menentukan kombinasi optimal dari ketiga

parameter tersebut. Adapun fungsi aktivasi yang diujikan adalah RELU dan TANH, sedangkan fungsi optimisasi yang diujikan adalah SGD, Adam dan L-BFGS. Untuk rentang jumlah dimensi PCA yang digunakan adalah 1-30. Hasil eksperimen menunjukkan bahwa kombinasi antara fungsi aktivasi RELU dengan metode optimisasi adam dan kombinasi antara fungsi aktivasi TANH dengan metode

optimisasi adam menghasilkan tingkat akurasi tertinggi, yaitu dengan tingkat akurasi 0.974.

Penelitian ke depannya dapat memanfaatkan data dari modalitas lainnya untuk melakukan CAD pada kanker payudara

REFERENSI

1. Negnevitsky M. Artificial intelligence: a guide to intelligent systems. 2nd ed. Harlow, England; New York: Addison-Wesley; 2005. 415 hlm.
2. Kshirsagar P, Rathod N. Artificial neural network. *Int J Comput Appl.* 2012;
3. Seidel P, Seidel A, Herbarth O. Multilayer perceptron tumour diagnosis based on chromatography analysis of urinary nucleosides. *Neural Netw.* 2007;20(5):646–651.
4. Ganpiseti R, Chandluri P, Lakshmi B, Swami PA. *WORLD JOURNAL OF PHARMACEUTICAL AND MEDICAL RESEARCH.*
5. Ferlay J, Héry C, Autier P, Sankaranarayanan R. Global burden of breast cancer. Dalam: *Breast cancer epidemiology.* Springer; 2010. hlm. 1–19.
6. Jeleń \Lukasz, Fevens T, Krzyżak A. Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies. *Int J Appl Math Comput Sci.* 2008;18(1):75–83.
7. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph.* 2007;31(4):198–211.
8. Lichman M. UCI Machine Learning Repository [Internet]. University of California, Irvine, School of Information and Computer Sciences; 2013.

Tersedia pada:
<http://archive.ics.uci.edu/ml>

9. Bro R, Smilde AK. Principal component analysis. *Anal Methods.* 2014;6(9):2812–2831.
10. Awada W, Khoshgoftaar TM, Dittman D, Wald R, Napolitano A. A review of the stability of feature selection techniques for bioinformatics data. Dalam: *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on.* IEEE; 2012. hlm. 356–363.
11. Janecek A, Gansterer W, Demel M, Ecker G. On the relationship between feature selection and classification accuracy. Dalam: *New Challenges for Feature Selection in Data Mining and Knowledge Discovery.* 2008. hlm. 90–105.