Analisis Perbandingan Reduksi Dimensi Principal Component Analysis (PCA) dan Partial Least Square (PLS) untuk Deteksi Kanker menggunakan Data Microarray

Daniel Tanta Christopher Sirait[1], Adiwijaya[2], Widi Astuti[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]danielchristopher@students.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id,
[3]widiwdu@telkomuniversity.ac.id

Abstrak

Menurut data WHO (World Health Organization) pada tahun 2015, 8.8 juta kematian diakibatkan oleh kanker dimana angka kematian tersebut meningkat dan berkakibat fatal setiap tahunnya bila diagnosa tidak dilakukan lebih dini. Oleh karena itu , tidak heran penelitian dalam bidang kanker menjadi topik utama dalam penelitian di bidang medis dan bioinformatika dan terus berkembang hingga saat ini, termasuk teknologi DNA microarray. Banyak cara untuk mendeteksi kanker, salah satunya adalah teknik microarray. Microarray adalah teknologi yang mampu menyimpan ribuan ekspresi gen yang diambil dari beberapa jaringan manusia sekaligus. Dikarenakan oleh record data microarray yang banyak, komputasi yang dibutuhkan cukup berat. Untuk mengatasi masalah tersebut, dibutuhkan reduksi dimensi. Pada penelitian ini, sistem menggunakan dua fitur ekstrasi: Principal Component Analysis (PCA) dan Partial Least Square (PLS) dengan Support Vector Machine (SVM) sebagai classifier. Hal ini berguna untuk mengurangi attribute yang terlalu banyak. Sistem yang dibangun mampu mengklasifikasi kanker dan memperoleh nilai rata-rata 82% dengan PCA-SVM dan 55.17% untuk PLS-SVM.

Kata kunci : kanker, microarray, principal component analysis, partial least square, support vector machine.

Abstract

According to WHO data in 2015 (World Health Organization), 8.8 million deaths were caused by cancer where the mortality rate increased and was fatal every year if the diagnosis was not made earlier. Therefore, it is not surprising that research in the field of cancer has become a major topic in research in the medical and bioinformatics fields and continues to grow to date, including DNA microarray technology. There are many ways to detect cancer, one of which is the microarray technique. Microarray is a technology that can store thousands of gene expressions taken from several human tissues at once. Due to a large number of microarray data records, the computing required is quite heavy. To overcome this problem, dimension reduction is needed. In this study, the system uses two extraction features: Principal Component Analysis (PCA) and Partial Least Square (PLS) with Support Vector Machine (SVM) as a classifier. This is useful to reduce the large amount of attributes. The accuracy generated from this system averaged 82% with PCA-SVM and 55.17% for PLS-SVM.

Keywords: cancer, microarray, principal component analysis, partial least square, support vector machine.

1. Pendahuluan

Latar Belakang

Kanker merupakan salah satu penyakit yang paling berbahaya di seluruh dunia. Menurut data WHO pada tahun 2015, 8.8 juta kematian diakibatkan oleh kanker dimana angka kematian tersebut meningkat dan berkakibat fatal setiap tahunnya bila diagnosa tidak dilakukan lebih dini [26]. Kanker adalah istilah yang digunakan untuk penyakit di mana sel-sel abnormal membelah tanpa kontrol dan mampu menyerang sel jaringan lain. Sel-sel kanker dapat menyebar ke bagian lain dari tubuh melalui darah.

Oleh karena itu , tidak heran penelitian dalam bidang kanker menjadi topik utama dalam penelitian di bidang medis dan bioinformatika dan terus berkembang hingga saat ini, termasuk teknologi DNA *microarray*. Banyak cara untuk mendeteksi kanker, salah satunya adalah teknik *microarray*. *Microarray* adalah teknologi yang mampu menyimpan ribuan ekspresi gen yang diambil dari beberapa jaringan manusia sekaligus. Analisis *microarray* memiliki peran penting dalam melakukan diagnosa penyakit dikarenakan oleh kemampuannya untuk melihat tingkat

gen pada sampel sel dan memeriksa ribuan gen secara bersamaan agar jaringan pada tubuh manusia terkena kanker atau tidak [7]. Karena itu , sangat diperlukan suatu teknologi yang mampu mendeteksi penyakit kanker lebih awal dengan analisis yang akurat, sehingga penyakit kanker dapat ditangani sejak dini.

Teknik *microarray* mampu menghasilkan data yang dibutuhkan untuk proses prediksi dan klasifikasi gen yang diambil dari beberapa jaringan tertentu pada manusia untuk digolongkan ke dalam kanker atau bukan. Namun yang menjadi kendala utama dalam suatu data *microarray* adalah besarnya dimensi. Besarnya dimensi inilah yang menjadi masalah dalam suatu data *microarray* karena informasi penting dan berguna dihalangi dari data tersebut, yang mengakibatkan beban komputasi yang tidak stabil dan tingkat performansi yang rendah. Untuk mengatasi masalah tersebut, *Principal Component Analysis* (PCA) dan *Partial Least Square* (PLS) digunakan untuk mereduksi dimensinya. Kemudian, proses klasifikasi bertujuan untuk mengklasifikasikan data kanker atau bukan kanker, dengan menggunakan klasifikasi *Support Vector Machine* (SVM). Hal ini dibuktikan pada buku yang ditulis oleh Henry Horng, Shing Lu, Bernhard Scholkopf dan Hongyu Zhao [19], mengatakan bahwa "*Support Vector Machine* dapat mengatasi data yang memiliki dimensi besar dan berhasil diterapkan pada penelitian *microarray*".

Topik dan Batasannya

Proses reduksi dimensi terdiri dari dua yaitu ekstrasi fitur dan seleksi fitur. Seleksi fitur dapat digunakan untuk mengurangi biaya pengumpulan data, reduksi waktu komputasi, melakukan visualisasi data maupun menggali informasi tentang penyebab masalah [15]. Ekstraksi fitur umumnya melakukan pengurangan dimensi melalui data yang tidak relevan agar data-data tersebut tidak memperburuk tingkat akurasi ketika dilakukannya pengklasifikasian. Pada penelitian ini, penulis menggunakan dua ekstrasi fitur yaitu: *Principal Component Analysis* (PCA) dan *Partial Least Square* (PLS).

Satu set ekspresi gen data *microarray* dapat direpresentasikan dalam bentuk tabel, dimana setiap baris merupakan satu gen tertentu, setiap kolom memiliki sampel atau titik waktu, dan untuk setiap entri dari matriks adalah tingkat ekspresi yang diukur dari gen tertentu dalam sampel atau titik waktu masing – masing. *Data set microarray* yang diperoleh dengan melakukan percobaan pada beberapa sampel biasanya dirancang sebagai matriks dua dimensi dengan *n* sebagai baris dan *m* sebagai kolom. Tujuan dari klasifikasi pada data *microarray* adalah untuk memisahkan atau membedakan satu jenis sampel kanker atau bukan kanker untuk mendapatkan hasil prediksinya.

Tujuan

Dengan menggunakan beberapa parameter pada reduksi dimensi dan klasifikasi, akan dilakukan analisis terhadap performansi klasifikasinya menggunakan SVM yang telah direduksi terlebih dahulu menggunakan algoritma PCA dan PLS.

2. Studi Terkait

Kanker

Kanker merupakan penyakit akibat pertumbuhan tidak normal dari sel-sel jaringan tubuh yang berubah menjadi sel kanker dimana sel-sel ini berkembang dan menyebar ke bagian tubuh lainnya sehingga dapat menyebabkan kematian. Penyakit mematikan ini telah menelan banyak korban . Angka kematian yang diakibatkan kanker selalu bertambah. Laporan terbaru yang dirilis oleh *International Agency for Research on Cancer*, Organisasi Kesehatan Dunia (WHO) mengestimasi terdapat 18,1 juta kasus kanker baru dan 9,6 juta kematian yang terjadi pada tahun 2018 [9]. Untuk mengurangi angka kematian tersebut, diperlukan solusi yang tepat. Berdasarkan identifikasi yang telah dilakukan dalam jangka waktu lama, kanker terjadi karena gangguan gen. Penanganan yang diberikan saat ini berupa pencegahan, deteksi sejak dini, diagnosis, terapi dan rehabilitasi [12]. Dalam masa perawatan yang dilakukan, banyak pasien yang tidak tertolong karena terlambatnya deteksi kanker untuk menangani penyakit mereka. Oleh karena itu, deteksi kanker sejak dini sangat krusial untuk mengurangi jumlah kematian para pasien.

Microarray Data

Sebagian jumlah besar data yang berguna untuk memecahkan banyak masalah di bidang biologi dapat dilakukan dengan teknik yang disebut *microarray*. *Microarray* adalah teknologi yang mampu menyimpan ribuan ekspresi gen yang diambil dari beberapa jaringan manusia sekaligus. Dengan menganalisis data *microarray*, dapat diketahui apakah jaringan tersebut terkena kanker atau tidak. Studi ini memberikan performa kinerja yang cepat dan akurat untuk mendeteksi kanker berdasarkan klasifikasi data *microarray* [7].

Data yang digunakan adalah data *breast cancer*, *ovarian cancer*, *colon* dan *lung cancer* yang didapatkan dari *Kent Ridge Biomedical Data Repository* [18]. Spesifikasi dari data tersebut bisa dilihat pada tabel 1. Kolom jumlah kelas, merupakan kolom yang menginformasikan banyaknya kelas yang akan di klasifikasikan. Kolom *sample*, merupakan kolom yang menginformasikan jumlah data sampel yang tersedia disetiap data kanker. Sedangkan kolom *feature* (fitur), merupakan kolom yang menginformasikan jumlah atribut (penilaiaan) pada setiap *sample*.

Tabel 1. Detail Data Kanker dari Kent Ridge Biomedical Data Repository

Data	Jumlah Kelas	Sample	Feature
Lung Cancer	2	181 (31 Mesothelioma, 150 ADCA)	12533
Breast Cancer	2	78 (34 Relapse, 44 Non-Relapse)	24481
Ovarian Cancer	2	253 (91 Normal, 162 Cancer)	15154
Colon	2	62 (22 Positive, 40 Negative)	2000

Reduksi Dimensi

Data set DNA *microarray* memiliki data yang sangat terbatas dan untuk menambah penyediaan data diperlukan biaya yang tinggi. Masalah utama yang dijumpai dalam *data set* DNA *microarray* adalah *Curse of Dimensionality*. Masalah ini mempersulit proses pencarian informasi atau bahkan menghalangi informasi penting yang sangat dibutuhkan. Sehingga proses pengolahan data menjadi kurang efektif dan efisien yang akan menyebabkan beban komputasi menjadi tidak stabil [16]. Masalah dimensi ini mempengaruhi akurasi dalam melakukan klasifikasi. Pada dasarnya, dengan adanya penambahan atribut pada dataset maka akan mempermudah pengklasifikasian, namun jika atribut terus bertambah sedangkan jumlah sampel tidak bertambah, maka inilah yang menjadi masalah dan hal tersebut mengakibatkan akurasi pengklasifikasian menurun daripada meningkat. Oleh karena itu, diperlukan sebuah metode untuk menangani masalah dimensi yang tinggi ini. Diperlukan beberapa perhitungan dalam matriks dan matematika pada perhitungan reduksi dimensi [3] [2].

Berbagai macam metode digunakan untuk mereduksi dimensi data *microarray*. Pada tahun 2018, Husna dan Adiwijaya [6] meneliti reduksi dimensi dalam data *microarray*. Penelitian ini menggunakan Random Forest sebagai metode *feature selection* dan membandingkan hasil antara Random Forest dengan dan tanpa reduksi dimensi. Hasil hanya menggunakan Random Forest menghasilkan akurasi rata-rata 72,63%, sedangkan Random Forest dengan reduksi dimensi menghasilkan akurasi 91,24%. Adiwijaya, Untari N. Wisesty, E. Lisnawati, A. Aditsania dan Dana S. Kusumo [5] menggunakan *Principal Component Analysis* (PCA) sebagai metode *feature extraction* dan membandingkan hasil antara *Support Vector Machine* (SVM) dan *Levenberg-Marquardt Backpropagation* (LM-BP) sebagai *classifier*. Berdasarkan hasil yang diperoleh, kedua metode klasifikasi sangat tinggi. Metode LMBP menghasilkan akurasi 96.07%, sedangkan SVM menghasilkan akurasi 94.98%. Adiwijaya [4] menerapkan BPNN dan PCA pada beberapa jenis data kanker, didapatkan bahwa BPNN dan PCA mendapatkan hasil akurasi lebih dari 80% dengan waktu *training time* 0-4 detik. Deegala [13] melakukan reduksi fitur yang mana ada beberapa metode reduksi fitur yang digunakan, yaitu PCA, *Random Projection* (RP), *Partial Least Square* (PLS) dan *Information Gain* (IG). Hasil penelitian tersebut menunjukan bahwa PCA dan PLS mencapai akurasi terbaik dengan komponen yang lebih sedikit daripada dua metode lainnya. Berdasarkan beberapa penelitian sebelumnya [20] [8] [14], dapat disimpulkan bahwa reduksi dimensi dapat diterapkan dalam klasifikasi data *microarray* untuk meningkatkan kinerja sistem.

Principal Component Analysis (PCA)

Dimensi dan kompleksitas data *microarray* sangat besar. Oleh karena itu, sangat diperlukan suatu proses yang dapat mengurangi kompleksitas tersebut. Pengurangan pada kompleksitas data *microarray* bertujuan untuk meminimalkan kesalahan dalam klasifikasi proses. *Principal Component Analysis* (PCA) digunakan untuk mereduksi dimensi tanpa mengurangi karakteristik data secara signifikan. Metode ini mengubah sebagian besar variabel asli yang berkorelasi menjadi satu himpunan variabel baru yang lebih kecil dan saling bebas. [24]

Proses reduksi dimensi dilakukan untuk mengurangi dimensi *data set* pada data *microarray* dikarenakan banyaknya dimensi serta kompleksitas yang sangat besar. Diekspetasikan, dengan berkurangnya dimensi *data set* pada *microarray* dapat mengurangi kompleksitas dalam proses klasifikasi. Adapun tahapan untuk melakukan reduksi dimensi dengan menggunakan *Principal Component Analysis* (PCA) [22], diantaranya:

1. Data diatur sebagai satu set n data vektor $X_1 \ldots X_n$ dimana matriks input untuk PCA adalah $X_(i, j)$ untuk *training*, dengan i merupakan baris dan j merupakan kolom.
2. Menghitung nilai mean dari data menggunakan rumus (1):

$$\bar{X} = \sum_{i=1}^{n} \frac{1}{X_i} \tag{1}$$

dimana n adalah jumlah sampel atau jumlah data observasi dan X_i adalah data observasi.

3. Mencari matriks kovarian dari data menggunakan rumus (2):

$$C_X = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \tag{2}$$

dimana n adalah jumlah sampel atau jumlah data observasi, X_i adalah data observasi dan $\bar{X}$ adalah nilai rata-rata data.

4. Menghitung nilai *eigen* (v_m) dan vektor *eigen* (λ_m) menggunakan rumus (3):

$$C_X v_m = \lambda_m v_m \tag{3}$$

5. Sortir nilai eigen dalam urutan menurun.

6. *Principal Component* (PC) adalah kumpulan vektor eigen sesuai dengan nilai eigen yang telah diurutkan dalam langkah 5.

7. Dimensi *Principal Component* akan dikurangi berdasarkan pada nilai eigen.

8. Setelah itu pilih beberapa vektor eigen dengan nilai eigen tertinggi.

9. Selanjutnya, lakukan transformasi data menggunakan vektor eigen yang telah dipilih.

Partial Least Square (PLS)

PLS adalah kelas teknik untuk memodelkan hubungan antara blok fitur yang diamati dengan menggunakan fitur laten. Asumsi yang mendasari PLS adalah bahwa data yang diamati dihasilkan oleh sistem atau proses yang didorong oleh sejumlah kecil fitur laten (tidak diamati atau diukur secara langsung). Oleh karena itu, PLS bertujuan untuk menemukan transformasi linear tidak berkorelasi (komponen laten) dari fitur prediktor asli yang memiliki kovarians tinggi dengan fitur respons. Berdasarkan komponen laten ini, PLS memprediksi fitur respons y, tugas regresi, dan merekonstruksi matriks X asli, tugas pemodelan data, pada saat yang sama [17]. Untuk memudahkan, *Partial Least Square* mencari komponen X yang relevan terhadap Y [1]. *Partial Least Square* memodelkan suatu data dengan persamaan (4) dan (5) [21]:

$$X = TP^T \tag{4}$$

$$Y = UQ^T \tag{5}$$

dengan X adalah variabel prediktor, Y adalah variabel dependen, T dan U matriks skor atau komponen laten, P dan Q merupakan matriks *loading* [23].*Partial Least Square* mempunyai keuntungan dapat mereduksi kompleksitas *microarray* dengan membuat dimensi data *microarray* menjadi lebih kecil. Walaupun metode ini seperti *Principal Component Analysis* yang dalam mereduksi dimensinya diperoleh dari memaksimalkan kovarian tetapi metode ini merupakan metode *supervised* tidak seperti *Principal Component Analysis* yang merupakan metode *unsupervised* [11].

Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah klasifikasi linear yang berfungsi untuk mencari *hyperplane* terbaik yang memisahkan antar kelasnya [10]. Pada masalah non-linear, *Support Vector Machine* menggunakan trik kernel dalam data *training* sehingga dimensi itu menjadi tersebar luas. Setelah dimensi diubah, *Support Vector Machine* akan mencari *hyperplane* optimal yang dapat memisahkan sebuah kelas dari kelas yang lain. Proses untuk menemukan *hyperplane* menggunakan *Support Vector Machine* menurut Campbell dirinci di bawah ini [10]:

1. Terdapat data $\overrightarrow{X_1} \in (\overrightarrow{X_1}, \overrightarrow{X_2}, ...\overrightarrow{X_n})$ dengan X_i merupakan data yang terdiri dari M atribut dan kelas target $y_i \in$ +1,-1.

2. Diasumsikan bahwa kelas +1 dan -1 dapat terpisahkan secara sempurna oleh *hyperplane*, didefinisikan sebagai persamaan (6) berikut ini:

$$\overrightarrow{w}.\overrightarrow{x} + b = 0 \tag{6}$$

maka dari persamaan (6) diperoleh:

$$\overrightarrow{w}.\overrightarrow{x} + b \geq 1, untuk\, kelas + 1 \tag{7}$$

$$\vec{w}.\vec{x} + b \geq -1, untuk\, kelas - 1 \tag{8}$$

dengan $\vec{w}$ sebagai normal bidang, $\vec{x}$ adalah data input dan b posisi bidang relatif terhadap pusat koordinat.

3. *Support Vector Machine* bertujuan untuk menemukan *hyperplane* pemisah yang memaksimalkan jarak antara dua kelas. Memaksimalkan jarak adalah masalah *Quadratic Programming* (QP), yaitu dengan mencari titik minimal persamaan:

$$min_w \frac{1}{2}(\| \vec{w} \|)^2 \tag{9}$$

dengan kendala pada persamaan (10):

$$y_i(\vec{x}_i\vec{w} + b) - 1 \geq 0 \tag{10}$$

dengan $\vec{x}_i$ sebagai data input ke-i, y_i adalah target ke-i, $\vec{w}$ adalah normal bidang dan b adalah posisi bidang relatif terhadap pusat koordinat.

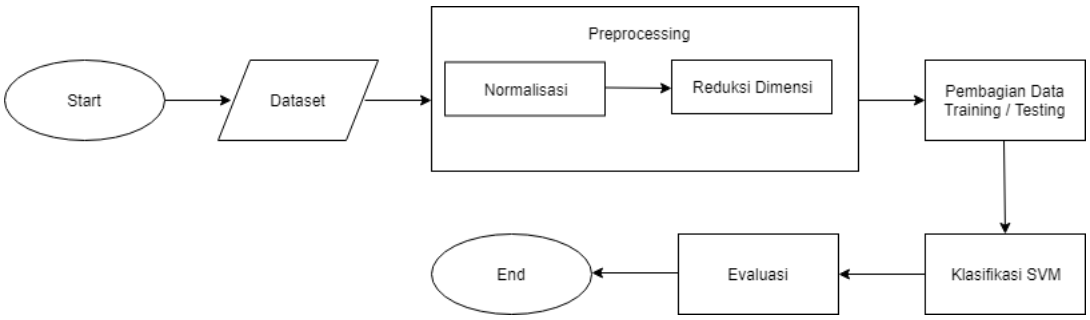
4. Membuat fungsi keputusan untuk persamaan (11) dan (12):

$$Linear : f(\vec{x}_d) = sign(\sum_{i=1}^{n} \alpha_i y_i(\vec{x}_i\vec{x}_d) + b) \tag{11}$$

$$Nonlinear : f(\vec{x}_d) = sign(\sum_{i=1}^{n} \alpha_i y_i K(\vec{x}_i\vec{x}_d) + b) \tag{12}$$

dengan n yaitu jumlah *support* vektor, $\vec{x}_i$ adalah *support* vektor ke-i, $\vec{x}_d$ adalah data uji, α merupakan bobot, y_i sebagai target ke-i dan b yaitu posisi bidang relatif terhadap pusat koordinat.

3. Sistem yang Dibangun



Gambar 1. Skema umum proses klasifikasi ekspresi gen.

Untuk mendeteksi kanker berdasarkan data *microarray* maka dibentuklah sebuah skema umum, dimana proses-prosesnya terdiri dari proses *preprocessing*, reduksi dimensi, dan proses klasifikasi. Proses pada skema umum dapat dilihat pada gambar 1. Pada penelitian ini metode yang digunakan adalah *Support Vector Machine* sebagai *classifier* untuk menentukan label kelas data *microarray*. Reduksi dimensi sangat diperlukan karena dimensi data *microarray* sangat besar. Reduksi dimensi yang digunakan pada penelitian ini adalah *Principal Component Analysis* (PCA) dan *Partial Least Square* (PLS), dimana peneliti akan membandingkan dan melakukan analisis terhadap kedua teknik reduksi dimensi tersebut.

Data *microarray* merupakan data yang memiliki dimensi sangat besar. Selain itu, salah satu masalah yang terdapat pada data *microarray* yang akan digunakan yaitu skala (*range*) pada setiap fiturnya memiliki perbedaan nilai yang cukup signifikan. Oleh karena itu, proses preprocessing normalisasi data yang membuat range nilai pada setiap atribut (*feature*) data *microarray* seragam atau berada pada range 0 sampai 1 sangat dibutuhkan. Setelah itu dilakukan proses reduksi dimensi yang bertujuan untuk menemukan gen ataupun fitur yang informatif dan mengurangi kompleksitas data yang yang akan digunakan sebagai masukan (*input*). Pada penelitian ini, penulis menggunakan metode *k-fold cross validation* (CV) yang berbeda tersedia untuk pemilihan sampel sebagai set data *training*. Metode CV digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dipisahkan menjadi dua subset yaitu data proses *training* dan data *testing*. Model atau algoritma dilatih oleh subset *training* dan *testing* oleh subset validasi. Selanjutnya pemilihan jenis CV dapat didasarkan pada ukuran dataset. Biasanya CV K-fold digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi. Selanjutnya

melakukan proses klasifikasi data *microarray* untuk menentukan kelas kanker atau tidak kanker pada setiap *sample*. Setelah diklasifikasikan, peneliti akan membandingkan kedua performansi antara PCA dan PLS untuk menentukan manakah yang memiliki hasil lebih baik.

Preprocessing

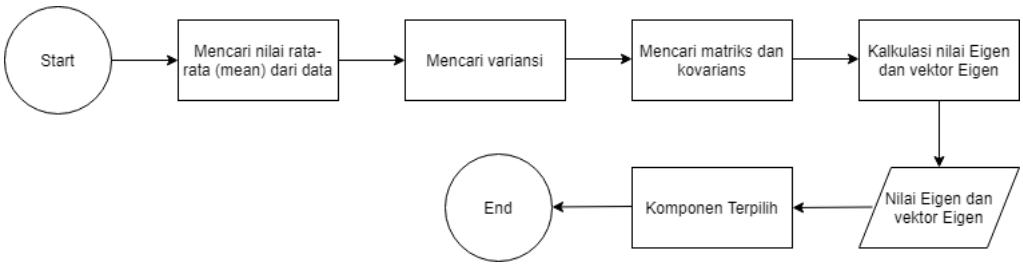
Seperti pada gambar 1, dapat dilihat bahwa sistem dimulai dengan memasukkan data *microarray* yang di dalamnya merupakan informasi dari sel kanker antara lain yaitu *breast cancer*, *ovarian cancer*, *colon* dan *lung cancer*. Kemudian data akan diolah dengan teknik *preprocessing* yang diaplikasikan untuk mengurangi kompleksitas data masukan (*input*) sehingga data menjadi lebih mudah digunakan. Banyak metode *preprocessing* yang dapat diimplementasikan pada data, yaitu normalisasi data dimana nilai pada data diubah menjadi interval 0-1 agar kompleksitas data pada saat data digunakan sebagai masukan (*input*) akan berkurang. Rumus umum untuk normalisasi data terdapat pada persamaan (13) berikut ini [25]:

$$Normalisasi = \frac{data - min(data)}{max(data) - min(data)} \tag{13}$$

Reduksi Dimensi

Untuk mengurangi dimensi besar pada data *microarray* diperlukan sebuah teknik reduksi dimensi agar kesalahan dapat diminimumkan dan *running time* ketika proses klasifikasi menjadi lebih baik. Pada tabel 1 terlihat bahwa jumlah fitur yang akan digunakan sebagai masukan (*input*) sangatlah besar, maka dari itu perlu dilakukan reduksi dimensi yang bertujuan untuk mengurangi jumlah fitur tanpa mengurangi karakteristik data secara signifikan. Pada penelitian ini, peneliti menggunakan dua teknik reduksi dimensi yaitu *Principal Component Analysis* (PCA) dan *Partial Least Square* (PLS).

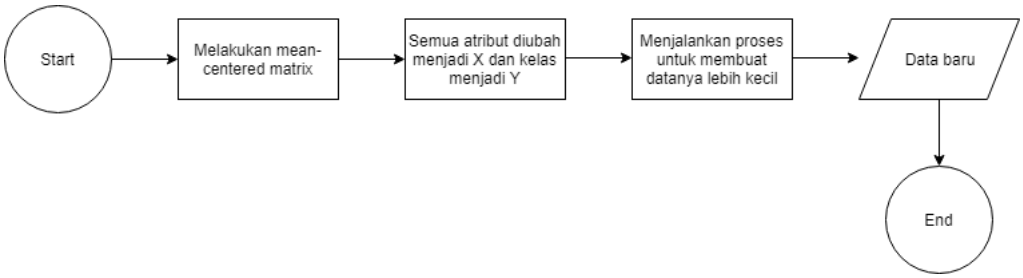
Principal Component Analysis (PCA)



Gambar 2. Proses reduksi dimensi PCA.

Pada PCA, proses reduksi dimensi dilakukan berdasarkan dengan nilai eigen dan vektor eigen. Nilai eigen dan vektor eigen didapatkan dari perhitungan matriks kovarians. Setelah nilai eigen dan vektor eigen didapatkan, maka kemudian akan diurutkan berdasarkan nilai eigen dari yang terbesar hingga yang terkecil dan vektor eigen akan dipilih berdasarkan nilai eigen untuk mengurangi fitur yang terdapat pada data *microarray*. Skema umum proses *Principal Component Analysis* (PCA) dapat dilihat pada gambar 2.

Partial Least Square (PLS)



Gambar 3. Proses reduksi dimensi PLS.

Partial Least Square (PLS) merupakan ekstraksi fitur, sehingga PLS akan membentuk data baru hasil dari olahan dari data asli. Tahapan awal pada PLS adalah melakukan *mean-centered matrix*. *Mean-centered matrix* adalah mencari rata-rata setiap atribut kemudian data dikurangi oleh rata-rata tersebut. Kemudian menjadikan semua atribut data train menjadi X dan label kelas menjadi Y. PLS akan memproses X dan Y sehingga menghasilkan data

baru yang jumlah attribut nya lebih kecil dibandingkan dengan data asli. Selanjutnya, data baru tersebut kemudian yang akan dijadikan data train. Skema umum proses *Partial Least Square* (PLS) dapat dilihat pada gambar 3.

Klasifikasi

Tahap terakhir dari penelitian ini adalah mengklasifikasi data kanker yang telah diolah menggunakan preprocessing dan reduksi dimensi, dimana akan diketahui data *microarray* termasuk ke dalam kelas klasifikasi kanker atau tidaknya terhadap seseorang berdasarkan *sample* yang diperoleh. Metode yang digunakan sebagai *classifier* adalah *Support Vector Machine* (SVM). *Support Vector Machine* (SVM) adalah salah satu *supervised learning* yang bertujuan untuk mencari hyperplane paling optimal dalam memisahkan kelas-kelas di ruang input.

Perhitungan Performansi (Akurasi)

Performansi sistem merupakan ukuran ketepatan nilai suatu sistem dapat mengenali dan melakukan klasifikasi dengan benar sehingga menghasilkan keluaran yang benar. Pada sistem ini perhitungan performansi dilakukan berdasarkan persamaan (14):

$$Performansi\,Akurasi = \frac{Jumlah\,Data\,Benar}{Jumlah\,Data\,Keseluruhan} \times 100\% \quad (14)$$

4. Evaluasi

4.1 Hasil Pengujian

Implementasi dilakukan pada 4 jenis data kanker dengan menerapkan metode *Principal Component Analysis* (PCA) dan *Partial Least Square* (PLS). Pada proses learning model menggunakan SVM, data pada *input* space ditransformasi dengan menggunakan kernel trick. Kernel yang digunakan yaitu linear, RBF dan polynomial dengan parameter (*d*) *degree* bernilai 3 dan *C* bernilai 1000 [5] dimana data di *split* atau dibagi menjadi 10 dan dilakukan sebanyak 10 kali iterasi. Ketiga kernel tersebut memegang peranan penting dalam proses pengklasifikasian keempat *dataset* kanker. Dari ketiga kernel, akan ada satu kernel terbaik yang memisahkan kedua buah kelas pada masing - masing *dataset*.

Skenario pengujian sistem pada PCA-SVM dan PLS-SVM menggunakan *10-Fold cross validation*, dimana data latih dan data uji dibagi menjadi 10 bagian, dan masing-masing fold akan dijadikan data uji pada setiap iterasi yang berbeda sebanyak 10 kali. Data latih dan data uji yang digunakan untuk evaluasi sistem akan tetap sama di setiap pengujian parameter yang berbeda. Parameter yang akan diuji antara lain jumlah komponen yang diperoleh setelah direduksi pada setiap *data set*. Hasil pengujian dengan 10-Fold validation ini ditunjukkan pada Tabel 2, tabel 3, tabel 4, tabel 5, tabel 6, dan tabel 7. Hasil rata-rata dari semua klasifikasi data kanker dapat dilihat pada gambar 4 dan gambar 5.

4.2 Analisis Hasil Pengujian PCA - SVM

Tabel 2. Skenario Pengujian PCA - SVM kernel Linear

Data	Komponen	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Rata-rata
Breast	4	66.66	55.56	55.56	55.56	42.86	85.71	85.71	71.43	85.71	42.86	64.76
	9	33.33	88.89	44.44	77.78	71.43	71.43	71.42	85.71	28.57	57.14	63.01
	19	55.56	66.67	55.56	77.78	71.43	57.14	100.00	85.71	14.29	71.43	65.56
	39	11.11	55.56	55.56	55.56	42.86	57.14	100.00	42.86	42.86	42.86	50.63
Colon	3	71.43	57.14	66.67	66.67	50.00	83.33	50.00	66.67	33.33	83.33	62.86
	5	57.14	85.71	83.33	100.00	100.00	100.00	83.33	66.67	66.67	66.67	80.95
	10	85.71	71.43	83.33	100.00	100.00	100.00	66.67	66.67	50.00	100.00	82.38
	20	85.71	85.71	66.67	100.00	100.00	83.33	66.67	50.00	50.00	66.67	75.48
Lung	30	100.00	100.00	94.44	94.44	88.89	100.00	88.89	100.00	88.89	94.44	95.00
	47	94.74	94.44	88.89	94.44	88.89	100.00	83.33	100.00	88.89	88.89	92.25
	72	94.74	100.00	94.44	88.89	88.89	94.44	88.89	94.44	88.89	83.33	91.70
	111	94.74	94.44	94.44	83.33	88.89	88.89	83.33	88.89	88.89	83.33	88.92
Ovarian	3	62.96	80.77	52.00	92.00	84.00	72.00	80.00	72.00	92.00	92.00	77.97
	4	77.78	84.62	80.00	92.00	92.00	92.00	92.00	84.00	96.00	88.00	87.84
	6	88.89	96.15	88.00	96.00	96.00	100.00	100.00	92.00	96.00	100.00	95.30
	13	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

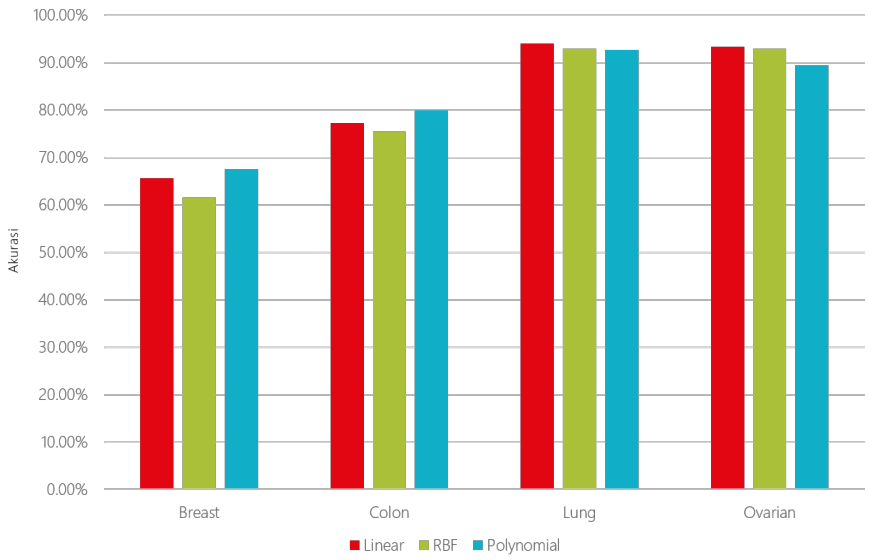
Dalam penelitian ini, pengujian dilakukan pada 4 dataset kanker menggunakan *Principal Component Analysis* (PCA) sebagai metode *feature extraction* dan SVM sebagai *classifier*. Setelah proses *feature extraction*, *Principle*

Tabel 3. Skenario Pengujian PCA - SVM kernel RBF

Data	Komponen	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Rata-rata
Breast	4	55.56	66.67	55.56	44.44	42.86	85.71	85.71	28.57	85.71	28.57	57.94
	9	33.33	77.78	44.44	66.67	71.43	85.71	100.00	71.43	28.57	85.71	66.51
	19	55.56	66.67	55.56	77.78	71.43	57.14	100.00	85.71	14.29	71.43	65.56
	39	11.11	66.67	44.44	55.56	42.86	57.14	100.00	57.14	57.14	42.86	53.49
Colon	3	57.14	57.14	66.67	66.67	50.00	66.67	83.33	66.67	33.33	66.67	61.43
	5	57.14	100.00	83.33	83.33	83.33	100.00	83.33	66.67	50.00	83.33	79.05
	10	85.71	85.71	100.00	100.00	100.00	100.00	83.33	66.67	50.00	66.67	83.81
	20	71.43	85.71	83.33	100.00	100.00	100.00	66.67	66.67	50.00	66.67	79.05
Lung	30	100.00	100.00	94.44	100.00	88.89	100.00	88.89	100.00	88.89	94.44	95.56
	47	94.74	94.44	88.89	94.44	88.89	94.44	83.33	100.00	88.89	88.89	91.70
	72	94.74	100.00	94.44	88.89	88.89	94.44	88.89	94.44	88.89	83.33	91.70
	111	94.74	94.44	94.44	83.33	88.89	88.89	83.33	88.89	88.89	88.89	89.47
Ovarian	3	70.37	61.54	64.00	88.00	72.00	72.00	76.00	72.00	96.00	92.00	76.39
	4	77.78	76.92	76.00	92.00	88.00	92.00	92.00	80.00	96.00	88.00	85.87
	6	88.89	96.15	92.00	100.00	96.00	100.00	100.00	96.00	100.00	100.00	96.90
	13	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Tabel 4. Skenario Pengujian PCA - SVM kernel Polynomial

Data	Komponen	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Rata-rata
Breast	4	55.55	55.55	66.66	55.55	57.14	57.14	57.14	57.14	57.14	57.14	57.615
	9	22.22	88.88	33.33	77.77	57.14	85.71	85.71	28.57	71.4	100	65.073
	19	55.55	66.66	44.44	88.88	71.42	57.14	85.71	100	28.57	85.71	68.408
	39	22.22	66.66	44.44	66.66	71.4	85.7	100	57.14	57.14	57.14	62.85
Colon	3	57.14	57.14	66.66	66.66	66.66	83.33	83.33	66.66	33.33	66.66	64.757
	5	71.4	100	66.66	83.33	100	100	100	66.66	66.66	83.33	83.804
	10	85.71	100	83.33	100	100	100	100	66.66	50	100	88.57
	20	85.71	85.71	83.33	83.33	83.33	100	100	66.66	50	100	83.807
Lung	30	94.73	100	88.88	88.88	88.88	100	88.88	100	88.88	88.88	92.801
	47	94.73	100	88.88	83.33	88.88	94.44	88.88	100	88.88	88.88	91.69
	72	94.73	94.44	94.44	83.33	88.88	94.44	94.44	94.44	88.88	88.88	91.69
	111	84.21	88.88	83.33	83.33	88.88	88.88	83.33	88.88	88.88	83.33	86.193
Ovarian	3	70.37	73.03	64	88	80	76	80	72	92	92	78.74
	4	81.48	80.76	72	92	84	92	100	84	96	92	87.424
	6	92.5	96.15	80	96	96	100	100	96	100	96	95.265
	13	100	100	100	100	100	100	100	96	100	100	99.6



Gambar 4. Grafik rata-rata klasifikasi PCA-SVM

Component (PC) diperoleh. Jumlah PC tergantung pada *Proportion of Variance* (PPV). Setelah dilakukan klasifikasi, hasil rata-rata yang diperoleh menggunakan kernel linear ialah 82.595%, kernel RBF sebesar 80.754% dan

kernel polynomial sebesar 82.4%. Dapat dilihat pada tabel 2 dan tabel 3 bahwa untuk akurasi terbaik yaitu 100% pada *ovarian cancer* dengan kernel Linear, RBF dan juga 99.6% untuk polynomial.

Hasil pengujian terhadap nilai PPV pada setiap data testing kanker terhadap akurasi sistem menggunakan metode SVM. Hasil tes menunjukkan bahwa ada perubahan dalam akurasi untuk setiap nilai PPV (60-90 %) yang digunakan untuk setiap data kanker. Oleh karena itu, untuk mengukur kinerja metode klasifikasi di setiap dataset, penulis juga mengukur akurasi rata-rata yang dihasilkan untuk setiap PPV yang ditentukan. Analisis setiap data berdasarkan PPV, menurut Tabel 3, menunjukkan bahwa nilai PPV yang lebih besar tidak selalu meningkatkan hasil akurasi, mis. dalam data *colon* dan *breast*. Pemilihan PPV tergantung pada karakteristik data yang digunakan. Ini karena data *noise* (data yang tidak informatif) digabungkan ke dalam data yang dihasilkan dari proses reduksi dimensi. Data *noise* memiliki nilai eigen (varians) yang relatif kecil, sehingga dapat mengganggu data informatif lainnya dan mengurangi hasil klasifikasi.

4.3 Analisis Hasil Pengujian PLS - SVM

Dalam penelitian ini, pengujian dilakukan pada 4 dataset kanker menggunakan *Partial Least Square* (PLS) sebagai metode *feature extraction* dan SVM sebagai *classifier*. Setelah dilakukan klasifikasi, hasil rata-rata yang diperoleh menggunakan kernel linear ialah 48.81%, kernel RBF sebesar 58.52% dan kernel polynomial sebesar 56.67%. Jumlah *component* disesuaikan dengan PC yang diperoleh pada percobaan PCA - SVM sebelumnya untuk memberikan perbandingan yang lebih sesuai. Dapat dilihat pada tabel 6 dan tabel 7 bahwa untuk akurasi terbesar diperoleh oleh *ovarian* dengan kernel RBF sebesar 81.57%, dan kernel polynomial sebesar 81.57% sedangkan *lung cancer* mendapatkan akurasi terbesar dengan kernel linear yaitu 78.45%.

Tabel 5. Skenario Pengujian PLS - SVM kernel Linear

Data	Komponen	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Rata-rata
Breast	4	55.56	55.56	55.56	55.56	57.14	57.14	57.14	57.14	57.14	57.14	56.51
	9	55.56	55.56	55.56	55.56	57.14	42.86	57.14	57.14	57.14	57.14	55.08
	19	55.56	22.22	22.22	33.33	42.86	42.86	14.29	28.57	14.29	57.14	33.33
	39	55.56	11.11	22.22	33.33	42.86	42.86	28.57	14.29	14.29	57.14	32.22
Colon	3	57.14	57.14	66.67	66.67	66.67	66.67	50.00	50.00	66.67	66.67	61.43
	5	57.14	57.14	66.67	66.67	66.67	66.67	50.00	66.67	66.67	66.67	63.10
	10	57.14	57.14	33.33	66.67	33.33	50.00	50.00	66.67	66.67	66.67	54.76
	20	42.86	57.14	16.67	83.33	50.00	50.00	33.33	50.00	66.67	66.67	51.67
Lung	30	78.95	83.33	83.33	83.33	61.11	61.11	83.33	83.33	83.33	83.33	78.45
	47	73.68	72.22	83.33	77.78	50.00	27.78	83.33	83.33	77.78	83.33	71.26
	72	78.95	72.22	83.33	77.78	27.78	22.22	61.11	55.56	77.78	77.78	63.45
	111	63.16	61.11	55.56	66.67	22.22	50.00	83.33	88.89	88.89	88.89	66.87
Ovarian	3	62.96	65.38	64.00	64.00	64.00	64.00	64.00	64.00	64.00	64.00	64.03
	4	62.96	65.38	64.00	64.00	64.00	64.00	64.00	64.00	64.00	64.00	64.03
	6	62.96	65.38	64.00	64.00	64.00	64.00	64.00	64.00	64.00	64.00	64.03
	13	62.96	62.96	65.38	64.00	64.00	64.00	64.00	64.00	64.00	64.00	64.00

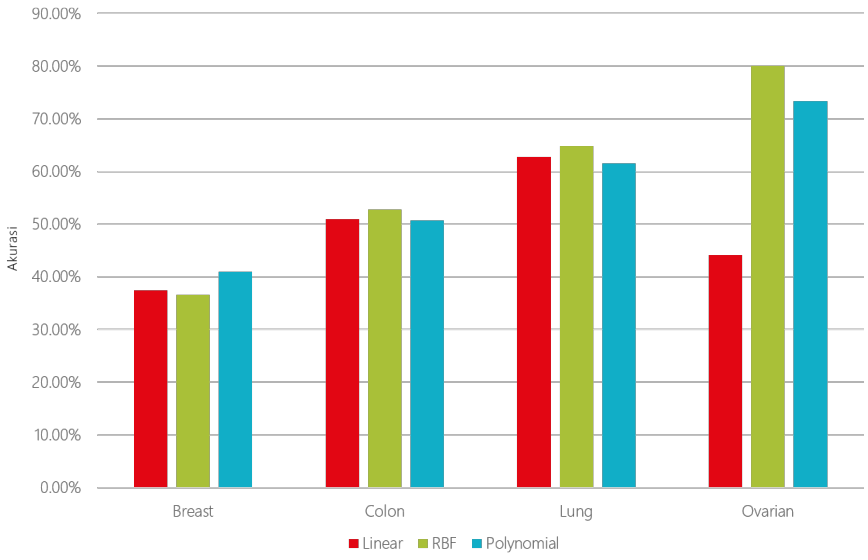
Tabel 6. Skenario Pengujian PLS - SVM kernel RBF

Data	Komponen	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Rata-rata
Breast	4	77.78	33.33	66.67	66.67	42.86	57.14	28.57	42.86	85.71	57.14	55.87
	9	55.56	44.44	33.33	55.56	42.86	42.86	42.86	42.86	42.86	57.14	46.03
	19	55.56	11.11	22.22	44.44	42.86	42.86	14.29	14.29	14.29	42.86	30.48
	39	55.56	11.11	22.22	33.33	42.86	42.86	28.57	14.29	14.29	57.14	32.22
Colon	3	42.86	57.14	33.33	66.67	66.67	66.67	50.00	50.00	66.67	66.67	56.67
	5	57.14	57.14	33.33	66.67	50.00	50.00	50.00	50.00	66.67	66.67	54.76
	10	57.14	57.14	0.00	50.00	16.67	33.33	50.00	66.67	66.67	50.00	44.76
	20	57.14	71.43	16.67	83.33	50.00	50.00	50.00	50.00	50.00	66.67	54.52
Lung	30	78.95	83.33	83.33	83.33	50.00	55.56	83.33	83.33	83.33	83.33	76.78
	47	78.95	83.33	83.33	77.78	44.44	16.67	83.33	83.33	77.78	83.33	71.23
	72	78.95	83.33	83.33	77.78	22.22	16.67	77.78	72.22	83.33	83.33	67.89
	111	68.42	55.56	55.56	72.22	33.33	44.44	83.33	83.33	83.33	88.89	66.84
Ovarian	3	37.04	34.62	84.00	100.00	96.00	92.00	100.00	92.00	88.00	92.00	81.57
	4	40.74	30.77	72.00	92.00	80.00	96.00	80.00	88.00	88.00	92.00	75.95
	6	37.04	34.62	92.00	88.00	92.00	96.00	88.00	92.00	92.00	100.00	78.37
	13	44.44	34.62	88.00	88.00	88.00	80.00	92.00	96.00	80.00	96.00	78.71

Tabel 7. Skenario Pengujian PLS - SVM kernel Polynomial

Data	Komponen	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Rata-rata
Breast	4	77.78	44.44	66.67	66.67	42.86	57.14	28.57	42.86	85.71	57.14	56.98
	9	55.56	33.33	44.44	44.44	42.86	42.86	42.86	28.57	28.57	57.14	42.06
	19	55.56	11.11	22.22	22.22	42.86	28.57	28.57	14.29	14.29	57.14	29.68
	39	55.56	11.11	22.22	33.33	42.86	42.86	28.57	14.29	14.29	57.14	32.22
Colon	3	57.14	57.14	66.67	66.67	66.67	66.67	50.00	50.00	66.67	66.67	61.43
	5	57.14	57.14	50.00	66.67	50.00	66.67	50.00	66.67	66.67	66.67	59.76
	10	57.14	42.86	0.00	66.67	33.33	33.33	33.33	66.67	66.67	66.67	46.67
	20	57.14	71.43	33.33	83.33	66.67	50.00	50.00	50.00	50.00	66.67	57.86
Lung	30	78.95	83.33	83.33	83.33	61.11	66.67	83.33	83.33	83.33	83.33	79.01
	47	78.95	83.33	83.33	77.78	61.11	38.89	83.33	83.33	83.33	83.33	75.67
	72	78.95	83.33	83.33	77.78	38.89	22.22	83.33	83.33	83.33	83.33	71.78
	111	78.95	83.33	77.78	77.78	11.11	27.78	83.33	83.33	83.33	88.89	69.56
Ovarian	3	37.04	34.62	84.00	88.00	84.00	80.00	88.00	84.00	88.00	92.00	81.57
	4	40.74	30.77	80.00	84.00	80.00	96.00	80.00	80.00	80.00	84.00	73.55
	6	37.04	34.62	88.00	88.00	88.00	92.00	92.00	92.00	80.00	92.00	81.17
	13	48.15	34.62	80.00	84.00	88.00	76.00	92.00	80.00	72.00	92.00	74.68

Salah satu kesulitan menggunakan PLS adalah pemilihan komponen. Tabel 5 menunjukkan bahwa tidak semua bagian atas komponen PLS yang diperoleh berguna untuk klasifikasi sehingga nilai performansi tidak optimal. Dalam penelitian selanjutnya, metode ini dapat dikembangkan bila metode reduksi dimensi disertai dengan parameter yang lebih sesuai untuk meningkatkan nilai akurasi. Penelitian selanjutnya dapat melakukan komparasi performansi persamaan PLS antar jurnal ataupun komparasi PLS dengan *classifier* lain.



Gambar 5. Grafik rata-rata klasifikasi PLS-SVM

5. Kesimpulan

Data *microarray* memiliki dimensi yang sangat tinggi dan mempunyai banyak noise sehingga reduksi dimensi diperlukan sebelum melakukan proses klasifikasi. Pada penelitian ini reduksi dimensi yang digunakan adalah *Principal Component Analysis* (PCA) dan *Partial Least Square* (PLS) yang merupakan ekstraksi fitur dan klasifikasi yang digunakan adalah *Support Vector Machine* (SVM). Dalam menguji metode klasifikasi SVM, parameter yang paling berpengaruh adalah jenis kernel yang digunakan. Sistem yang menggunakan PCA-SVM menghasilkan akurasi yang cenderung bagus untuk klasifikasi data *microrarray* dengan rata-rata sebesar 82% dengan nilai akurasi rata-rata tertinggi sebesar 100% yang diperoleh pada data *ovarian cancer* dengan menggunakan kernel linear dan RBF, sedangkan PLS-SVM mendapatkan nilai akurasi rata-rata yang lebih kecil yaitu sebesar 54.67% dengan nilai akurasi rata-rata tertinggi sebesar 81.57% yang diperoleh pada data *ovarian cancer* menggunakan kernel RBF dan Polynomial. Dalam penelitian ini, berbagai parameter klasifikasi dan metode reduksi dimensi

digunakan untuk mengetahui kinerja sistem terbaik. PLS-SVM memiliki nilai akurasi yang sangat rendah dibandingkan dengan PCA-SVM dikarenakan oleh kurang optimalnya penghapusan noise dan menyimpan informasi yang penting dalam dataset. Kinerja performansi PCA-SVM juga dapat ditingkatkan dengan pemilihan arsitektur lebih terstruktur. Jadi pada penelitian selanjutnya ada baiknya untuk mencoba menggunakan algoritma lain supaya mendapatkan performansi dengan hasil optimal.

Daftar Pustaka

[1] H. Abdi. *Partial Least Square Regression*. Encyclopedia of Measurement and Statistic, 2007.

[2] Adiwijaya. *Aplikasi Matriks dan Ruang Vektor*. Graha Ilmu, 2014.

[3] Adiwijaya. *Matematika Diskrit Dan Aplikasinya*. alfabeta, 2016.

[4] Adiwijaya. Deteksi kanker berdasarkan klasifikasi microarray data. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 2(4):181–186, 2018.

[5] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, and D. S. Kusumo. Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification. *Journal of Computer Science*, 14(11):1521–1530, Nov. 2018.

[6] H. A. Adiwijaya. A clustering approach for feature selection in microarray data classification using random forest. *Journal of Information Processing System*, pages 1167–1175, 2018.

[7] A. A. Adiyasa, N; Adiwijaya; Suryani. Cancer Detection Based On Microarray Data Classification Using PCA and Modified Backpropagation. *Far East Journal of Electronics and Communications*, pages 269–281, 2015.

[8] W. Astuti and Adiwijaya. Principal component analysis sebagai ekstraksi fitur data microarray untuk deteksi kanker berbasis linear discriminant analysis. *Media Informatika Budidarma*, 2(4):181–186, 2018.

[9] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, Sept. 2018.

[10] C. Campbell. *Support Vector Machine and Kernel Methods*. 2005.

[11] L. R. D. Dai, J. J.; Lieu. Dimension reduction for classification with gene expression microarray data. *Statistical Application in Genetics Molecular Biology*, 2006.

[12] P. D. dan Informasi. *Stop Kanker*. Kementrian Kesehatan RI, 2015.

[13] S. Deegalla and H. Boström. Classification of microarrays with kNN: Comparison of dimensionality reduction methods. In *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, pages 800–809. Springer Berlin Heidelberg, 2007.

[14] A. F. B. Firmansyah and S. Pramana. Ensemble based gustafson kessel fuzzy clustering.

[15] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[16] S. R. S. K. Kumar, Mukesh; Singh. Classification of Microarray data using Functional Link Neural Network. *Procedia Computer Science*, pages 727–737, 2015.

[17] G.-Z. Li, X.-Q. Zeng, J. Y. Yang, and M. Q. Yang. Partial least squares based dimension reduction with gene selection for tumor classification. In *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*. IEEE, Oct. 2007.

[18] J. Li. *Kent-ridge bio-medical data set repository, School of Computer Engineering Nanyang Technological University, Singapore*, 2013.

[19] H. S. B. Lu, Henry Horng-Shing; Zhao. *Handbook of Statistical Bioinformatics*. Springer, 2011.

[20] A. Manik, D. Q. Utama, et al. Classification of electrocardiogram signals using principal component analysis and levenberg marquardt backpropagation for detection ventricular tachyarrhythmia. *Journal of Data Science and Its Applications*, 2(1):78–87, 2019.

[21] K. S. Ng. A simple explanation of partial least squares. 2013.

[22] F. Nurfalah, A.; Umbara. Colorectal cancer classification using pca and fisherface feature extraction data from pathology microscopic image. *ISICO*, 2013.

[23] R. Nurhasanah; Subianto, M; Fitriani. Perbandingan Metode Partial Least Square (pls) dengan Regresi Komponen Utama untuk mengatasi Multikolinearitas. *STATISTIKA: Forum Teori dan Aplikasi Statistika*, pages 33–42, 2012.

[24] R. Pujianto, A. Adiwijaya, and A. A. Rohmawati. Analisis ekstraksi fitur principle component analysis pada klasifikasi microarray data menggunakan classification and regression trees. *eProceedings of Engineering*, 6(1), 2019.

[25] P. Stafford. *Methods in microarray normalization*. CRC Press, 2008.

[26] WHO. Cancer Facts and Tableures. *American Cancer Society*, 2015.