Implementasi Partial Least Square dan K-Nearest Neighbor - Support Vector Machines Untuk Klasifikasi Data Microarray

A Rakha Ahmad Taufiq[1], **Adiwijaya**[2], **Annisa Aditsania**[3]

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

Abstrak

Kanker menjadi salah satu penyebab kematian paling banyak di dunia. Diperkirakan setiap tahun jumlahnya akan terus bertambah. Salah satu pendeteksiannya adalah menggunakan ekspresi gen. *Microarray* dapat mengoleksi kumpulan besar ekspresi gen dalam satu waktu, sehingga DNA *microarray* mempunyai karakteristik data tersendiri, yaitu mempunyai dimensi data yang sangat besar dibanding dengan jumlah datanya. Oleh karena itu, dibutuhkannya sistem untuk menyelesaikan masalah tersebut. Pada penelitian ini, dibangun sistem yang mengimplementasikan ekstraksi fitur *Partial Least Square* (PLS) dan metode klasifikasi *K-Nearest Neighbor - Support Vector Machines* (KNN-SVM). Ekstraksi fitur berguna untuk mengurangi dimensi *microarray* yang sangat besar dengan membentuk data baru yang merupakan representasi data asli. Performansi sistem diukur menggunakan akurasi. PLS berhasil menaikkan akurasi dari *classifier* KNN-SVM. Nilai akurasi tertinggi yang didapatkan oleh PLS KNN-SVM adalah sebesar 96.17%

Kata kunci: k-nearest neighbor, support vector machines, partial least square, microarray.

Abstract

Cancer is one of the most common causes of death in the world. Estimated every year the number will continue to grow. One of the detection is using gene expression. Microarray can collect a large number of gene expression at the same time, DNA Microarray have their own data characteristic, which have a very large data dimension compared with the amount of data. Therefore, a system needed to solve the problem. In this research, we built a system that implements Partial Least Square (PLS) feature extraction and K-Nearest Neighbor - Support Vector Machines (KNN-SVM) for the classification. Feature extraction is useful for reducing very large dimension of microarray by forming new data. System performance is measured using accuracy. PLS managed to increase the accuracy of the KNN-SVM classifier. The highest accuracy obtained by PLS KNN-SVM is 96.17%.

Keywords: k-nearest neighbor, support vector machines, partial least square, microarray.

1. Pendahuluan

Latar Belakang

Menurut World Health Organization (WHO), kanker menjadi salah satu penyebab kematian paling banyak di dunia. Pada tahun 2015, 8.8 juta orang meninggal karena kanker [22], jumlah ini diperkirakan akan terus meningkat tiap tahunnya. Kematian akibat penyakit kanker dapat ditanggulangi dengan adanya pendeteksian dini, Terdapat beberapa cara untuk mendeteksi kanker, seperti rontgen, pemeriksaan fisik maupun pendeteksian melalui ekspresi gen.

Setiap orang mempunyai ribuan gen yang memiliki keunikannya masing-masing. Oleh karena itu, kondisi setiap orang dapat dideteksi melalui gen yang dipunyai. DNA *microarray* dikembangkan untuk dapat mengoleksi kumpulan besar ekspresi gen dalam satu waktu [26]. Hal tersebut telah menjadi alasan beberapa penelitian dalam melakukan deteksi kondisi menusia termasuk deteksi kanker melalui ekspresi gen.

Gen yang telah terkumpul pada *microarray* kemudian diklasifikasikan berdasarkan kelas masing-masing untuk menentukan bagaimana kondisi orang tersebut. Karakteristik dari *microarray* adalah jumlah dimensinya yang sangat besar dibandingkan dengan jumlah datanya. Hal ini menjadi tantangan tersendiri bagi peneliti, yaitu bagaimana membangun sistem klasifikasi yang mempunyai performansi tinggi baik dalam segi akurasi maupun waktu [20]. Salah satu solusinya adalah melakukan reduksi dimensi sebelum melakukan klasifikasi terhadap *Microarray Data*.

Pada tugas akhir ini metode yang digunakan untuk melakukan reduksi dimensi data microarray adalah dengan menggunakan *Partial Least Squares* (PLS). PLS mempunyai keuntungan dapat mereduksi kompleksitas *microarray* dengan membuat dimensi data *microarray* menjadi lebih kecil [10] dan PLS merupakan *supervised learning* sehingga komponen PLS mencari kovarian maksimal antara variabel prediktor dan variabel respon [19]. *Classifier* yang digunakan adalah *K-Nearest Neighbor–Support Vector Machines* (KNN-SVM). Metode ini merupakan gabungan antara KNN dan SVM dan bertujuan untuk meningkatkan performansi dari SVM serta KNN-SVM mempunyai *misclassification rate* yang lebih kecil dibanding KNN [28].

Perumusan dan Batasan Masalah

Perumusan masalah pada penelitian ini yaitu bagaimana mengimplementasikan reduksi dimensi dan klasifikasi terhadap data *Microarray*, bagaimana menganalisis performansi dari sistem yang kami buat, dan bagaimana pengaruh reduksi dimensi terhadap klasifikasi data *Microarray*.

Data yang digunakan berasal dari Kent-Ridge Repository. Kami menggunakan data yang berasal dari Kent-Ridge Repository karena *repository* tersebut telah lazim digunakan untuk penelitian data *Microarray*. Kami hanya menggunakan lima jenis data, yaitu Breast Cancer, Colon Tumor, Leukemia, Lung Cancer, dan Ovarian.

Tujuan

Pada penelitian ini kami membangun sistem yang terdiri dari reduksi dimensi dan klasifikasi. Proses reduksi dimensi menggunakan metode *Partial Least Square* dan proses klasifikasi menggunakan metode *K-Nearest Neighbor–Support Vector Machines* (KNN-SVM). Hasil dari sistem yang telah dibangun akan dihitung performansinya menggunakan Akurasi, kemudian kami membandingkan performansi PLS KNN-SVM dengan KNN-SVM tanpa reduksi dimensi.

Organisasi Tulisan

Jurnal ini disusun sebagai berikut. Penjelasan terkait studi terkait pada bab kedua. Bab tiga berisi perancangan sitem yang dibangun termasuk penggabungan KNN dan SVM. Hasil serta analisis hasil pengujian terdapat pada bab keempat. Bab lima berisi kesimpulan dan saran.

2. Studi Terkait

Tahun 1999, Brown [9] melakukan penelitian klasifikasi *microarray* menggunakan *Support Vector Machines* dan menghasilkan akurasi rata-rata sebesar 99,4 %. Satu tahun setelah itu, Li [15] melakukan penelitian menggunakan Algoritma Genetika dan *K-Nearest Neighbor* untuk mengklasifikasikan data *Microarray* keuntungan dari metode yang dipakai adalah metode tersebut dapat memilih gen prediktif diantara banyak data *noise*. Nguyen [19] dalam penelitiannya membandingkan reduksi dimensi *Partial Least Square* (PLS) dan *Principal Components Analysis* (PCA) untuk mereduksi data *microarray* hasil yang diperoleh menyatakan bahwa PLS terbukti lebih baik dari PCA. Tahun 2018, Munzir [17] melakukan penelitian klasifikasi pada *Microarray data* menggunakan *MBP Powell Beale* dengan hasil akurasi tertinggi sebesar 100%. Pada tahun yang sama, Adiwijaya [6] melakukan penelitian dengan menggunakan *Principal Component Analysis* sebagai reduksi dimensi pada *microarray data*.

Microarray

Microarray dikembangkan agar dapat mengumpulkan gen dengan jumlah yang sangat banyak secara bersamaan [26]. Sample yang berisi DNA atau RNA ditaruh didalam *chip* gen [24]. *Microarray* banyak digunakan dalam dunia kesehatan untuk mendeteksi penyakit terutama kanker, dimana terjadi abnormalitas pada sel yang terkena. Keuntungan *microarray* memungkinkan peneliti untuk dapat menganalisa gen dalam jumlah banyak sekaligus [24]. Hal lain yang menjadikan *microarray* banyak dijadikan sebagai topik penelitian karena *microarray data* merupakan data pasien yang disimpan dalam bentuk media elektronik dan pesatnya perkembangan internet saat ini memudahkan penyebaran media elektornik [5].

Reduksi Dimensi

Dataset yang berdimensi besar mempunyai banyak tantangan matematis, salah satunya adalah tidak semua variabel dalam data merupakan variabel penting. Reduksi dimensi banyak diminati karena dapat membuat model data prediksi dengan akurasi yang tinggi dari dataset asli yang berdimensi besar. Reduksi dimensi dapat dibagi menjadi *feature selection* dan *feature extraction*.

Feature Selection (FS) adalah teknik reduksi dimensi dengan memilih subset-subset penting dari data asli [27]. Metode FS dintaranya adalah *Ant Colony Optimization* dan *Algoritma Genetika*. *Feature Extraction* (FE) adalah teknik reduksi dimensi dengan membentuk model data baru dari tur yang relevan [13]. Metode FE antara lain adalah *Principal Component Analysis* dan *Partial Least Square*. FS dan FE mempunyai perbedaan dalam cara

mereduksi dimensi, tetapi FS dan FE mempunyai tujuan yang sama, yaitu menghilangkan *noise* dari data untuk meningkatkan nilai akurasi [8].

Partial Least Squares

Partial Least Square (PLS) adalah salah satu teknik untuk mereduksi dimensi data *microarray*. Tujuan dari PLS adalah memprediksi atau menganalisa satu set variabel dependen dari satu set variabel independen atau prediktor [3].

Secara teknis tujuan dari PLS adalah untuk memprediksi Y dari X dan mendeskripsi struktur umum mereka, dengan Y adalah variabel dependen dan X adalah variabel prediktor. Secara lebih mudah, PLS mencari komponen X yang relevan terhadap Y [3]. PLS memodelkan suatu data dengan persamaan [18]:

$$X = TP^T \quad (1)$$

$$Y = UQ^T \quad (2)$$

dengan X adalah variabel prediktor, Y adalah variabel dependen, T dan U matriks skor atau komponen laten, P dan Q merupakan matriks loading [21].

Gagasan utama PLS adalah mendekomposisikan X dan Y dengan mengambil informasi satu sama lain. Salah satu caranya adalah dengan menukar t dan u untuk memperbarui nilai dari w dan q kemudian penukaran nilai dan permbaruan nilai tersebut dilakukan secara perulangan [18]. w adalah vektor bobot dari X, t merupakan komponen kolom dari matriks T dan u merupakan kolom dari matriks U, sesuai dengan prosedur [14]:

$u = y_j$ for some j

Loop

$$w = X^T u / ||X^T u|| \quad (3)$$

$$t = Xw \quad (4)$$

$$q = Y^T t / ||Y^T t|| \quad (5)$$

$$u = Yq \quad (6)$$

until t stop changing

jika Y hanya terdiri dari satu dimensi, maka persamaan 5 dan 6 dapat diabaikan. kemudian hitung matriks loading untuk X :

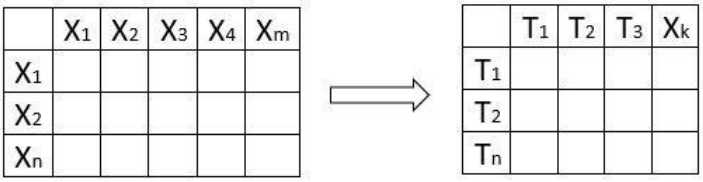
$$p = X^t t / ||t^t t|| \quad (7)$$

Prosedur diatas memberikan nilai pertama komponen dan matriks loading PLS. Untuk mencari nilai komponen selanjutnya, X dan Y diatur sebagai berikut [18]:

$$X = X - tp^T \quad (8)$$

$$Y = Y - uq^T \quad (9)$$

kemudian ulangi langkah yang sama. Setelah itu, simpan p, t, q, dan u pada matriks yang sesuai. Pada akhirnya kita akan mendapatkan matriks nilai P, T , Q, dan U . T merupakan matriks data hasil dari reduksi dimensi PLS.



Gambar 1. Ilustrasi reduksi dimensi oleh PLS. PLS mereduksi dimensi data asli X dengan membentuk data baru T yang merupakan representasi data asli dengan jumlah $k < m$.

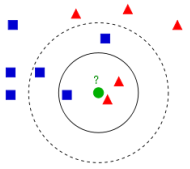
PLS mempunyai keuntungan dapat mereduksi kompleksitas *microarray* dengan membuat dimensi data *microarray* menjadi lebih kecil. Walaupun metode ini seperti PCA yang dalam mereduksi dimensinya diperoleh dari memaksimalkan kovarian tetapi metode ini merupakan metode *supervised* tidak seperti PCA yang merupakan metode *unsupervised* [10].

K-Nearest Neighbor

Metode K-Nearest Neighbor (KNN) sudah ada sejak sekitar tahun 1950 [26]. Metode ini merupakan algoritma klasifikasi dengan cara mengumpulkan infromasi dari tetangga. KNN mengklasifikan data dengan mencari kemiripan / jarak terdekat data tes terhadap data latih biasanya menggunakan *Euclidean distance* 10:

$$d = \sqrt{\sum |(x-y)|^2} \tag{10}$$

euclidean distance adalah teknik paling simpel dan populer tetapi memiliki kekurangan ruang yang akan masuk kedalam *k* tetangga terdekat berbentuk bulat [12]. Label kelas terbanyak dari tetangga sejumlah *K* terdekat yang kemudian dipilih menjadi lebel kelas data tes tersebut [4].



Gambar 2. Ilustrasi KNN [25]

Contoh, pada Gambar 2 obyek bewarna hijau merupakan obyek yang akan diklasifikasikan kedalam sebuah kelas. Dapat diihat jika K=3 maka obyek lingkaran hijau masuk ke dalam kelas segitiga merah, tetapi jika K=5 maka obyek lingkaran hijau masuk ke dalam kelas persegi biru.

KNN merupakan salah satu algoritma yang paling umum digunakan karena merupakan algoritma yang mudah digunakan serta dapat beradaptasi dengan mudah terhadap informasi baru. Namun, metode ini ketika menghadapi permasalah komplek menjadi tidak bagus akurasi prediksinya [26].

Support Vector Machines

Support Vector Machines (SVM) merupakan salah satu teknik pembelajaran mesin yang paling menjanjikan [7]. Metode ini diperkenalkan sekitar tahun 1970 oleh Vladimir Vapnik dan Alxei Chervonenkis [26]. Metode ini mempunyai akurasi yang tinggi dan dapat mengatasi data yang sangat kompleks tetapi membutuhkan komputasi yang intensif [26].

Ide dasar dari SVM adalah mencari garis pemisah terbaik antar kelas, garis tersebut dinamakan *hyperplane*. Letak pemisah diperoleh dari margin yang merupakan jarak terdekat antara vektor dengan *hyperplane*. Margin antara *hyperplane* dengan vektor disuatu kelas dan margin *hyperplane* dengan vektor di kelas lain adalah sama. Label dari masing masing kelas dapat dinotasikan dengan $y_i \in \{+1, -1\}$ untuk $i = 1, 2, 3, \ldots, l$ dengan *l* adalah jumlah data [23]. *Hyperplane* dapat dinotasikan dengan persamaan [23]:

$$\vec{w} . \vec{x} + b = 0 \tag{11}$$

Hyperplane adalah pemisah antar kelas. Oleh karena itu, dalam menentukan kelas $\vec{x}$ menggunakan pertidaksamaan [11]:

$$\vec{w} . \vec{x_i} + b \leq -1 \tag{12}$$

$$\vec{w} . \vec{x_i} + b \geq +1 \tag{13}$$

Hyperplane terbaik diperoleh dari margin terbesar yang memisahkan dua kelas [16]. Margin terbesar diperoleh dari memaksimalkan jarak *hyperplane* dan vektor dengan persamaan $\frac{1}{\|\vec{w}\|}$. Kondisi memaksimalkan margin ini ekivalen dengan meminimalkan fungsi obyektif [23]:

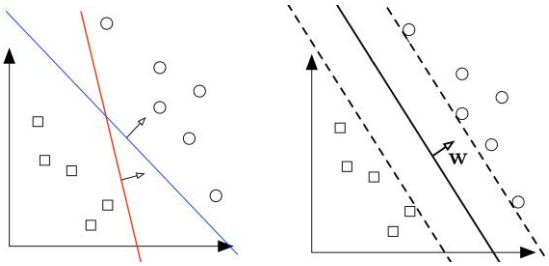
$$\Phi(w) = \frac{1}{2} \|w\|^2 \tag{14}$$

dengan *constraint* yang harus dipenuhi sesuai dengan persamaan berikut:

$$y_i (\vec{x_i} . \vec{w} + b) - 1 \geq 0, \forall i \tag{15}$$

Lagrange Multiplier adalah salah satu cara untuk mengatasi permasalahan optimasi ini. *Hyperplane* dapat dilihat pada Gambar 3

Pada Gambar 3 (kiri) terdapat beberapa *hyperplane* yang diperoleh untuk memisahkan kelas. Tetapi *hyperplane* terbaik pada Gambar 3 (kanan) diperoleh dari menghitung margin antara *hyperplane* dengan kelas.

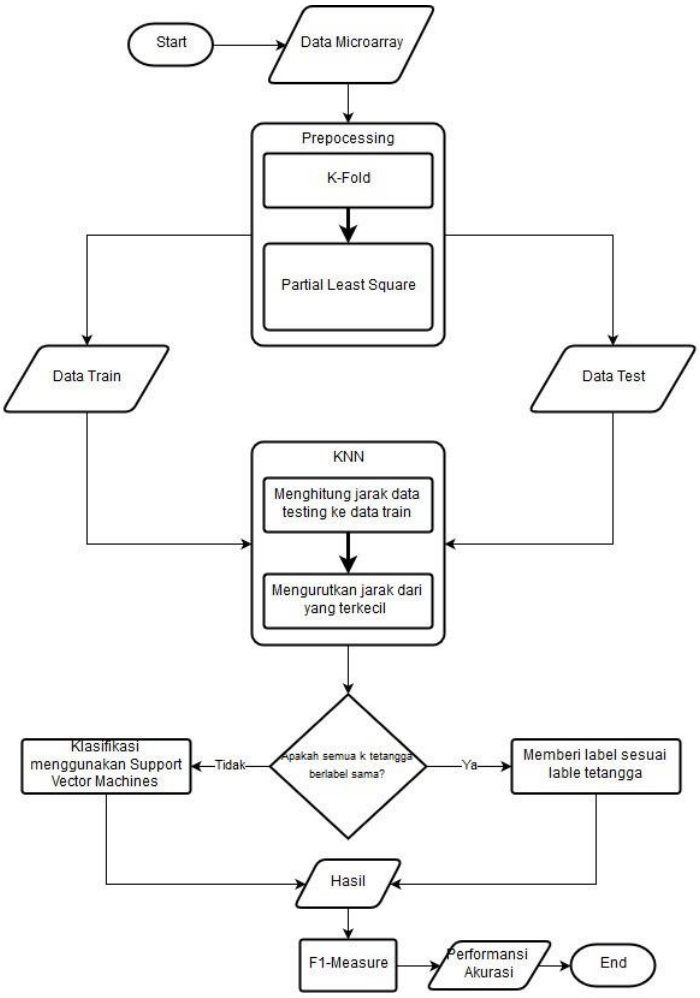


Gambar 3. Hyperplane [1]

Tidak semua data dapat dipisahkan secara linier. Sehingga pada SVM terdapat fungsi kernel. Fungsi ini juga memungkinkan SVM untuk menemukan *hyperplane* pada data berdimensi yang lebih besar.

3. Sistem yang Dibangun

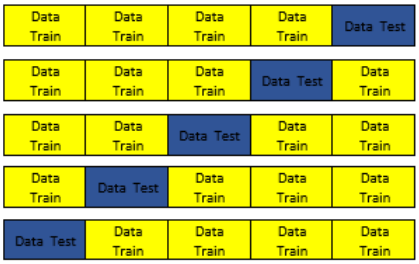
Pada penelitian ini metode yang digunakan adalah *K-Nearest-Neighbot – Support Vector Machines* (KNN-SVM) sebagai *classifier* untuk menentukan label kelas data *microarray* yang belum terdefinisi kelasnya. Reduksi dimensi sangat diperlukan karena dimensi data *microarray* sangat besar. Reduksi dimensi yang digunakan pada tugas akhir ini adalah *Partial Least Square* (PLS). Kemudian data dibagi menjadi dua bagian yatu data *train* dan data *test*. Data *test* digunakan sebagai estimasi performansi sistem.



Gambar 4. Alur skema PLS dan KNN-SVM sebagai sistem klasifikasi *microarray data*

Preprocessing

Seperti pada alur klasifikasi pada Gambar 4, dapat dilihat bahwa sistem dimulai dengan memasukkan data *microarray*. Kemudian data akan diproses dengan K-Fold. K-Fold ini kemudian yang akan membagi data menjadi data *train* dan data *test*. Pada penelitian ini nilai *K* pada K-Fold diatur 5. K-Fold akan membagi data sebanyak *K* dan membagi data manjadi data *train* dan data *test* dengan perbandingan sebesar 4:1. Data akan diiterasi sebanyak *K* dengan posisi data *test* akan berbeda disetiap iterasinya.



Gambar 5. Ilustrasi K-Fold dengan *K* = 5

Data-data tersebut kemudian dilakukan *preprocessing* dengan mereduksi dimensi. Pada penelitian ini, metode reduksi dimensi yang digunakan adalah adalah *Partial Least Square* (PLS). 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*, sedangkan untuk data *test* nya sendiri merupakan pemrosesan antara data *test* asli dengan matriks *loading* atau *P* hasil dari pemrosesan PLS data *train*.

Klasifikasi

Setelah melalui proses *preprocessing* sistem akan melakukan klasifikasi pada data. Pada penelitian ini karena menggunakan metode KNN-SVM. Oleh karena itu, terdapat beberapa data yang melewati metode KNN dan metode SVM seperti yang dapat dilihat pada gambar diatas. Terdapat beberapa tahapan penting dari proses ini, yaitu klasifikasi KNN, pelatihan SVM, klasifikasi menggunakan metode SVM. Berikut tahapan pada proses klasifikasi:

1. Data yang belum terdefinisi kelasnya (data *test*) akan diberi label kelas menggunakan metode KNN. Metode ini menghitung jarak antara data dengan semua data train menggunakan *euclidean distance*.
2. Kemudian jarak hasil proses sebelumnya akan diurutkan dari yang terkecil ke yang terbesar.
3. Jika sejumlah k tetangga terdekat mempunyai label kelas yang sama maka data tersebut diberi label sesuai label k tetangga terdekat.
4. Jika tidak maka data train sejumlah k tetangga terdekat akan diproses menggunakan metode SVM dengan *euclidean distance* sebagai kernel.
5. Metode SVM kemudian akan menentukan data tersebut masuk kedalam kelas mana.
6. Kemudian data tersebut diberi label sesuai dengan kelas hasil dari klasifikasi metode SVM.

Perhitungan Performansi

Performansi sistem merupakan ukuran ketepan nilai suatu sistem dapat mengenali dan melakukan klasifikasi dengan benar sehingga menghasilkan keluaran yang benar. Pada sistem ini perhitungan performansi menggunakan metode *Score* dan menggunakan *True Positif*, *True Negative* hasil dari *confusion matrix*.

Accuracy merupakan nilai ketepatan sistem dalam melakukan klasifikasi. *True Positive* (TP) adalah nilai keberhasilan sistem jika sistem memberi label benar dan label asli adalah benar. *True Negative* (TN) adalah nilai keberhasilan sistem jika sistem memberi label tidak dan label asli adalah tidak. Contoh, sistem mendapat nilai TP jika sistem berhasil menebak data kenker sebagai kanker dan diberi nilai TN jika sistem berhasil menebak data bukan kanker sebagai bukan kanker. Untuk mendapatkan *Accuracy* dapat dilihat pada persamaan 16.

Tabel 1. Confusion Matrix

Actual / Classified	*Classified Positive*	*Classified Negative*
Actual Positive	*True Positive (TP)*	*False Negative (FN)*
Actual Negative	*False Positive (FP)*	*True Negative (TN)*

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

4. Hasil dan Analisis

Penelitian ini menggunakan data yang berasal dari Kent-Ridge Repository [2]. Berikut data yang digunakan Tabel 2:

Tabel 2. Data yang digunakan

Nama Data	Jumlah Data	Dimensi Data	Kelas *Positive*	Kelas *Negative*
Leukemia	72	7129	AML	ALL
Breast Cancer	97	24481	relapse	non-relapse
Colon Tumor	62	2000	positive	negative
Lung Cancer	181	12533	Mesothelioma	ADCA
Ovarian Cancer	253	15154	Cancer	Normal

Penelitian ini menggunakan K-Fold dengan k=5, sehingga performansi sistem dihitung dari rata-rata setiap iterasi. Penelitian ini menggunakan K-Neighbor KNN-SVM 2,3,4,5,6,7,8,9, dan 10 untuk semua dataset dan Kernel yang dipakai adalah euclidean distance. Pada penelitian ini pembagian hasil berdasarkan dataset kanker. Pada penelitian ini kami membandingkan sistem yang kami bangun (PLS KNN-SVM) dengan KNN-SVM tanpa reduksi dimensi untuk mengetahui bagaimana pengaruh reduksi dimensi terhadap hasil klasifikasi.

Hasil Sistem pada Data Microarray Leukemia

Berikut hasil pengujian terhadap sistem yang diaplikasikan pada data Leukemia Tabel 3:

Tabel 3. Hasil Pengujian pada Data Leukemia

PLS KNN-SVM		KNN-SVM	
Neighbor KNN	*Accuracy*	*Neighbor* KNN	*Accuracy*
2	90.10%	2	64.00%
3	87.52%	3	69.71%
4	86.19%	4	66.86%
5	86.19%	5	69.71%
6	86.29%	6	69.71%
7	83.43%	7	69.71%
8	83.43%	8	69.71%
9	86.29%	9	69.71%
10	86.19%	10	65.43%

Dari Tabel 3 dapat dilihat bahwa untuk akurasi terbesar diperoleh oleh PLS KNN-SVM dimensi pada k = 2 dengan nilai sebesar 90.10% . KNN-SVM tanpa reduksi dimensi memperoleh nilai akurasi terbesar 71.05% pada k = 3, k = 5, k = 6, k = 7, k = 8, k = 9. Pada data Leukemia, PLS KNN-SVM mendapatkan nilai akurasi lebih tinggi dibandingkan dengan KNN-SVM tanpa reduksi dimensi. Dimensi data baru hasil reduksi dimensi PLS untuk iterasi ke-1, ke-2, ke-3, ke-4 dan ke-5 berturut adalah berjumlah 57, 57, 58, 58, dan 58 dimensi.

Hasil Sistem pada Data Microarray Colon Tumor

Berikut hasil pengujian terhadap sistem yang diaplikasikan pada data Microarray Colon Tumor Tabel 4:

Tabel 4. Hasil Pengujian pada Data Colon Tumor

PLS KNN-SVM		KNN-SVM	
Neighbor KNN	*Accuracy*	*Neighbor* KNN	*Accuracy*
2	79.36%	2	51.79%
3	66.41%	3	54.87%
4	74.49%	4	54.87%
5	74.49%	5	54.87%
6	68.21%	6	56.41%
7	82.44%	7	56.41%
8	75.77%	8	56.41%
9	83.59%	9	56.41%
10	80.38%	10	56.41%

Dari Tabel 4 dapat dilihat bahwa untuk akurasi terbesar diperoleh oleh PLS KNN-SVM dimensi pada k = 9 dengan nilai sebesar 83.59%. KNN-SVM tanpa reduksi dimensi memperoleh nilai akurasi terbesar 56.41% pada k = 6 hingga k = 10. Pada data Colon Tumor, PLS KNN-SVM mendapatkan nilai akurasi lebih tinggi dibandingkan dengan KNN-SVM tanpa reduksi dimensi. Dimensi data baru hasil reduksi dimensi PLS untuk iterasi ke-1, ke-2, ke-3, ke-4 dan ke-5 berturut adalah berjumlah 48, 48, 49, 49 dan 49 dimensi.

Hasil Sistem pada Data Microarray Lung Cancer

Berikut hasil pengujian terhadap sistem yang diaplikasikan pada data Microarray Lung Cancer Tabel 5:

Tabel 5. Hasil Pengujian pada Data Lung Cancer

PLS KNN-SVM		KNN-SVM	
Neighbor KNN	*Accuracy*	*Neighbor* KNN	*Accuracy*
2	95.06%	2	83.39%
3	95.62%	3	83.39%
4	95.06%	4	83.39%
5	95.06%	5	83.39%
6	94.50%	6	83.39%
7	96.17%	7	83.39%
8	93.96%	8	83.39%
9	96.73%	9	83.39%
10	93.41%	10	83.39%

Dari Tabel 5 dapat dilihat bahwa untuk akurasi terbesar diperoleh oleh PLS KNN-SVM dimensi pada k = 9 dengan nilai sebesar 96.73% . KNN-SVM tanpa reduksi dimensi memperoleh nilai akurasi yang stagnan dengan nilai 83.39% sejak k = 1 hingga k = 10. Pada data Lung Cancer, PLS KNN-SVM mendapatkan nilai akurasi lebih tinggi dibandingkan dengan KNN-SVM tanpa reduksi dimensi. Dimensi data baru hasil reduksi dimensi PLS untuk iterasi ke-1, ke-2, ke-3, ke-4 dan ke-5 berturut adalah berjumlah 144, 145, 145, 145 dan 145 dimensi.

Hasil Sistem pada Data Microarray Ovarian Cancer

Berikut hasil pengujian terhadap sistem yang diaplikasikan pada data Microarray Ovarian Cancer Tabel 6:

Tabel 6. Hasil Pengujian pada Data Ovarian Cancer

PLS KNN-SVM		KNN-SVM	
Neighbor KNN	*Accuracy*	*Neighbor* KNN	*Accuracy*
2	85.38%	2	47.38%

3	91.72%	3	58.45%
4	88.55%	4	51.36%
5	90.52%	5	64.00%
6	86.96%	6	59.23%
7	90.13%	7	65.62%
8	90.13%	8	62.82%
9	89.73%	9	65.62%
10	88.16%	10	61.64%

Dari Tabel 6 dapat dilihat bahwa untuk akurasi terbesar diperoleh oleh PLS KNN-SVM dimensi pada k = 2 dengan nilai sebesar 91.72%. KNN-SVM tanpa reduksi dimensi memperoleh nilai akurasi terbesar 65.62% pada k = 7 dan k = 9. Pada data Ovarian Cancer, PLS KNN-SVM mendapatkan nilai akurasi lebih tinggi dibandingkan dengan KNN-SVM tanpa reduksi dimensi. Dimensi data baru hasil reduksi dimensi PLS untuk iterasi ke-1, ke-2, ke-3, ke-4 dan ke-5 berturut adalah berjumlah 204, 204, 204, 205 dan 205 dimensi.

Hasil Sistem pada Data Microarray Breast Cancer

Berikut hasil pengujian terhadap sistem yang diaplikasikan pada data Microarray Breast Cancer Tabel 7:

Tabel 7. Hasil Pengujian pada Data Breast Cancer

PLS KNN-SVM		KNN-SVM	
Neighbor KNN	*Accuracy*	*Neighbor* KNN	*Accuracy*
2	46.58%	2	47.53%
3	57.84%	3	52.68%
4	52.68%	4	52.68%
5	60.79%	5	51.58%
6	62.95%	6	52.58%
7	55.58%	7	52.58%
8	56.74%	8	49.53%
9	58.68%	9	52.58%
10	53.42%	10	50.63%

Dari Dari Tabel 7 dapat dilihat bahwa untuk akurasi terbesar diperoleh oleh PLS KNN-SVM dimensi pada k = 6 dengan nilai sebesar 62.98% . KNN-SVM tanpa reduksi dimensi memperoleh nilai akurasi terbesar 52.68% pada k = 3 dan k = 4. Pada data Breast Cancer, PLS KNN-SVM mendapatkan nilai akurasi lebih tinggi dibandingkan dengan KNN-SVM tanpa reduksi dimensi. Dimensi data baru hasil reduksi dimensi PLS untuk iterasi ke-1, ke-2, ke-3, ke-4 dan ke-5 berturut adalah berjumlah 77, 77, 78, 78 dan 78 dimensi.

Analisis Umum

Hasil observasi menunjukkan bahwa PLS KNN-SVM mempunyai akurasi keseluruhan yang lebih bagus di-banding dengan KNN-SVM tanpa reduksi dimensi dengan rata-rata hasil akurasi dapat dilihat pada tabel. PLS KNN-SVM cenderung bagus pada data Microarray Lung Cancer dan pada data tersebut PLS KNN-SVM menca-tatkan nilai akurasi tertinggi.

Sistem PLS KNN-SVM yang diaplikasikan pada data Microarray Breast Cancer cenderung kecil dengan nilai rata-rata 56.16% , walaupun dengan reduksi dimensi PLS nilai akurasi lebih bagus dibandingkan dengan tanpa reduksi dimensi, hal ini dapat terjadi karena data Microarray Breast Cancer cenderung banyak *noise* sehingga menghasilkan nilai akurasi yang kecil pula.

PLS KNN-SVM mempunyai rata-rata 80.51% dan KNN-SVM tanpa reduksi dimensi memperoleh rata-rata sebesar 63.60% untuk semua data. Nilai rata-rata keseluruhan yang diperoleh PLS KNN-SVM lebih besar sekitar 17% dibandingkan dengan KNN-SVM tanpa reduksi dimensi. KNN-SVM tanpa reduksi dimensi mendapatkan nilai rata-rata keseluruhan yang kecil dapat terjadi karena KNN-SVM membatasi jumlah data yang diproses oleh SVM, hal ini sangat memungkinkan terjadinya kesalahan klasifikasi mengingat dimensi data yang sangat besar

Tabel 8. Rata-rata Hasil dari Penelitian

Data	PLS KNN-SVM	KNN-SVM
Leukemia	86.18%	68.29%
Colon	76.13%	55.38%
Lung	95.06%	83.39%
Ovarian	89.03%	59.57%
Breast	56.14%	51.37%

tetapi dengan jumlah data yang diproses cenderung sedikit. PLS membuktikan bahwa reduksi dimensi dapat menaikkan akurasi sistem termasuk pada masing masing data PLS mendapatkan nilai akurasi tertinggi dan hal ini menunjukkan bahwa data yang dipakai pada penelitian mempunyai *noise* yang kemudian perlu adanya *preprocessing* seperti reduksi dimensi untuk meingkatkan akurasi.

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 *Partial Least Squar*e (PLS) yang merupakan ekstraksi fitur dan klasifikasi yang digunakan adalah *K-Nearest Neighbor Support Vector Machines* (KNN-SVM). Sistem yang menggunakan PLS KNN-SVM menghasilkan akurasi yang cenderung bagus untuk klasifikasi data Micorarray dengan rata-rata sebesar 80.51% dengan nilai akurasi tertinggi sebesar 96.73% yang diperoleh pada data Lung Cancer, sedangkan KNN-SVM tanpa reduksi dimensi mendapatkan nilai akurasi rata-rata yang lebih kecil yaitu sebesar 63.60% dengan nilai akurasi tertinggi sebesar 83.59% yang diperoleh pada data Lung Cancer. Penelitian selanjutnya dapat melakukan komparasi performansi persamaan PLS antar jurnal ataupun komparasi PLS dengan *classifier* lain.

Daftar Pustaka

[1] Lecture 9: Svm. `http://www.cs.cornell.edu/courses/cs4780/2017sp/lectures/lecturenote09.html`. [Accessed 7 November 2017].

[2] Elvira biomedical data set repository. `http://leo.ugr.es/elvira/DBCRepository/`, February 2005. [Accessed 18 October 2017].

[3] H. Abdi. Partial least square regression (plsr). *Encyclopedia of Measurement and Statistic*, 2007.

[4] Adiwijaya, M. N. Aulia, M. S. Mubarok, W. U. Novia, and F. Nhita. A comparative study of mfcc-knn and lpc-knn for hijaiyyah letters pronounciation classification system. In *2017 5th International Conference on Information and Communication Technology (ICoIC7)*, pages 1–5, May 2017.

[5] Adiwijaya, T. A. B. Wirayuda, S. D. Winanjuar, and U. Muslimah. The multiple watermarking on digital medical image for mobility and authenticity. pages 457–462. Springer International Publishing, 2014.

[6] 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, 2018.

[7] S. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Network*, pages 783–789, 1999.

[8] H. Aydadenta and Adiwijaya. A clustering approach for feature selection in microarray data classification using random forest. *Journal of Information Processing System*, 14(5), 2018.

[9] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furrey, M. A. Jr, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS*, 97:262–267, 2000.

[10] J. J. Dai, L. Lieu, and D. Rocke. Dimension reduction for classification with gene expression microarray data. *Statistical Application in Genetics Molecular Biology*, 5, 2006.

[11] R. Diani, U. N. Wisesty, and A. Aditsania. Analisis pengaruh kernel support vector machine (svm) pada klasifikasi data microarray untuk deteksi kanker. *Indonesian Journal on Computing*, pages 109–118, 2017.

[12] A. Fauzi Bagus Firmansyah and S. Pramana. Ensemble based gustafson kessel fuzzy clustering. *Journal of Data Science and Its Applications*, 1:1–9, 07 2018.

[13] I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh, and (Eds.). *Feature Extraction*. Springer Science & Business Media, 2006.

[14] S. D. Jong. Simpls, an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251–163, 1993.

[15] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen. Gene selection for sample classification based on gene ekspression data: Study of sensivity to choice of parameters of the ga/knn method. *BIOINFORMATCICS*, 17:1131–1142, 2001.

[16] C.-F. Lin and S.-D. Wang. Fuzzy support vector machines. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 13:464–470, 2002.

[17] A. F. H. MUNZIR, A. , and A. ADITSANIA. Analisis reduksi dimensi pada klasifikasi microarray menggunakan mbp powell beale. *E-Jurnal Matematika*, pages 17–24, 2018.

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

[19] D. V. Nguyen and D. M. Rocke. Tumor classification by partial least squares using microarray gene expression data. *BIOINFORMATCICS*, 18:39–50, 2002.

[20] A. Nurfalah, Adiwijaya, and A. Ardiyanti. Cancer detection based on microarray data classification using pca and modified back propagation. *Far East Journal of Electronics and Communications*, 16:269–281, 05 2016.

[21] Nurhasanah, M. Subianto, and R. Fitriani. Perbandingan metode partial least square (pls) dengan regresi komponen utama untuk mengatasi multikolinearitas. *STATISTIKA: Forum Teori dan Aplikasi Statistika*, 12:33–42, 2012.

[22] W. H. Organization. Cancer fact sheet. http://www.who.int/, February 2017. [Accessed 14 October 2017].

[23] H. G. S., Adiwijaya, and K. Maulana. Analisis performansi klasifikasi email menggunakan support vector machines. *JURNAL PENELITIAN DAN PENGEMBANGAN TELEKOMUNIKASI*, 12:50–55, 2007.

[24] Firmansyah, A.F.B. and Pramana, S., 2018. Ensemble Based Gustafson Kessel Fuzzy Clustering. *Journal of Data Science and Its Applications*, *1*(1), pp.1-9.

[25] R. K. Singh and D. M. Sivabalakrishnan. Feature selection of gene expression data for cancer classification: A review. *Procedia Computer Science*, pages 52–57, 2015.

[26] Manik, A., Adiwijaya, A. and Utama, D.Q., 2019. 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), pp.29-37.

[27] Hadiana A., 2018. Designing Interface of Mobile Parental Information System based on Users' Perception Using Kansei Engingeering, Journal of Data Science and Its Applications (JDSA), 1(1), pp.10-19.

[28] W.-K. Yip, S. B. Amin, and C. Li. A survey of classification techniques for microarray data analysis. *Handbook of Statistical Bioinformatics*, pages 193–223, 2011.

[29] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. *International Conference on Machine Learning*, pages 1–8, 2003.

[30] Manuel B., Tricahyono D., 2018. Classifying Electronic Word of Mouth and Competitive Position in Online Game Industry, Journal of Data Science and Its Applications (JDSA), 1(1), pp.20-27.

[31] H. Zheng, A. Berg, M. Maire, and J. Malik. Svm-knn : Discriminative nearest neighbor classification for visual category. IEEE Computer Vision and Pattern Recognition (CVPR), 2:2126–2136, 2006.

[32] Alamsyah, A. and Syawiluna, M., 2018. Mapping Organization Knowledge Network and Social Media Based Reputation Management. Journal of Data Science and Its Applications, 1(1), pp.39-48.