Analisis Perbandingan Klasifikasi *Microarray* menggunakan *Naïve Bayes* dan *Support Vector Machine* (SVM) untuk Deteksi Kanker dengan *Feature Extraction* PCA

Vina Mutiara Purnama[1], Widi Astuti[2], Adiwijaya[3]

[1,2,3] Universitas Telkom, Bandung
[1]vinamutiara@students.telkomuniversity.ac.id, [2]astutiwidi@telkomuniversity.ac.id, [3]adiwijaya@telkomuniversity.ac.id

Abstrak

Kanker merupakan salah satu penyebab kematian manusia terbanyak di dunia. Diperkirakan penderita kanker terus meningkat setiap tahunnya. Kanker yang dapat terdeteksi lebih dini memiliki probabilitas lebih tinggi untuk mendapatkan penanganan yang lebih cepat dan tepat. Salah satu caranya dengan menggunakan teknologi *Microarray*. Teknologi *Microarray* dapat menganalisis ribuan profil gene expression dalam waktu yang bersamaan. Dengan melakukan analisa terhadap data *Microarray* selanjutnya dapat diketahui apakah seseorang terkena kanker atau tidak. Namun, permasalahan dalam data *Microarray* adalah jumlah atribut yang jauh lebih banyak dibandingkan sampel sehingga perlu dilakukannya reduksi dimensi. Untuk mengatasi hal tersebut, penulis menggunakan salah satu teknik reduksi dimensi yaitu *Principal Component Analysis* (PCA) dan menggunakan 2 metode klasifikasi yaitu *Naïve Bayes* dan *Support Vector Machine* (SVM), yang selanjutnya akan dibandingkan dan dianalisa hasil performansi dari kedua metode tersebut untuk mencari mana yang lebih baik. Akurasi dari hasil penelitian ini menunjukkan 4 dari 5 data kanker mendapatkan akurasi sebesar 77-96% sedangkan 1 data lainnya yaitu data *breast cancer* mendapatkan akurasi terkecil yaitu 54.6%.

Kata kunci : Kanker, Microarray, Reduksi Dimensi, Principal Component Analysis (PCA), Naïve Bayes, Support Vector Machine (SVM).

Abstract

Cancer is one of the leading causes of human death in the world. It is estimated that cancer patients continue to increase every year. Cancer that can be detected early has a higher probability of getting a faster and more appropriate treatment. One way is by using Microarray technology. Microarray technology can analyze gene expression profiles at the same time. By analyzing the Microarray data, it can then be known whether a person has cancer or not. However, the problem with Microarray data is that there are far more attributes than samples, so dimension reduction is necessary. To overcome this, the authors use one of the dimensional reduction techniques, namely Principal Component Analysis (PCA) and 2 classification methods, namely Naïve Bayes and Support Vector Machine (SVM), which will then be compared and analyzed the performance results of the two methods to find out which one is the best. better. The accuracy of the results of this study shows that 4 out of 5 cancer data get an accuracy of 77-96% while 1 other data, namely breast cancer data, gets the smallest value, which is 54.6%.

Keywords: Cancer, Microarray, Dimension Reduction, Principal Component Analysis (PCA), Naïve Bayes, Support Vector Machine (SVM).

1. Pendahuluan

Kanker adalah salah satu penyakit paling mematikan di dunia. Menurut Organisasi Kesehatan Dunia (WHO), kanker menyebabkan 9,6 juta kematian pada tahun 2018 [1]. Diagnosis kanker merupakan suatu hal yang sulit untuk dilakukan. Kesulitan dalam diagnosis kanker adalah karena banyaknya informasi klinis yang terdapat dari seorang pasien dan adanya faktor subjektif dalam interpretasi data tersebut. Oleh karena itu, teknologi yang cepat dan akurat diperlukan untuk membantu pendeteksian penyakit kanker.

Terdapat teknologi terkenal yang dapat mendeteksi kanker yaitu DNA *Microarray*. DNA *Microarray* merupakan kumpulan dari ribuan DNA mikroskopis berupa fragmen DNA yang ditempatkan pada sebuah chip [28]. Teknologi DNA *microarray* digunakan untuk menentukan tingkat ekspresi ribuan gen yang dilakukan dalam satu percobaan, dan secara bersamaan menyatukan proses biologi yang sedang berlangsung [2]. Teknologi ini dapat menghasilkan prediksi dan klasifikasi gen untuk digolongkan ke dalam kanker atau bukan. Namun, permasalahan utama yang sering dijumpai ketika mengklasifikasikan data microarray adalah jumlah gen yang

sangat banyak (high dimensional), dibandingkan dengan jumlah sampel yang terbatas. Maka dari itu, diperlukan reduksi dimensi pada data microarray [3].

Terdapat dua jenis reduksi dimensi, yaitu *feature selection* dan *feature extraction*. Pada penelitian ini digunakan reduksi dimensi *feature extraction*. *Feature extraction* adalah salah satu cara untuk mereduksi dan mengkompres atribut yang bertujuan untuk menghasilkan satu set fitur yang memiliki dimensi lebih kecil dari dimensi data asli tetapi tetap mempertahankan karakteristiknya [4]. *Feature extraction* berpeluang lebih kecil untuk terkena *overfitting* dan melakukan akurasi yang baik untuk klasifikasi dibandingkan dengan metode *feature selection* [30]. Pada penelitian ini digunakan *feature extraction* PCA, *Principal Component Analysis* (PCA) atau analisis komponen utama adalah metode analisis eksplorasi multivariat yang berguna untuk memisahkan variasi sistematis dari adanya *noise*. Metode ini memungkinkan untuk dapat menentukan ruang dengan dimensi yang dikurangi tapi tetap mempertahankan informasi yang relevan dari data asli [6].

Penelitian ini akan menganalisis data DNA *microarray* untuk mengetahui metode klasifikasi mana yang dapat meningkatkan hasil akurasi dalam pengklasifikasian *gene expression* tersebut. Metode klasifikasi yang digunakan adalah *Naïve Bayes* dan *Support Vector Machine* (SVM). Metode klasifikasi yang dipakai didasarkan pada penelitian-penelitian sebelumnya yang menyebutkan bahwa metode-metode tersebut memiliki performansi yang baik dalam pengklasifikasian data DNA *microarray*. Data yang digunakan adalah data *microarray* yang didapatkan dari *Kent-Ridge Biomedical Dataset* yang terdiri dari data *breast cancer, colon tumor, lung cancer, ovarian cancer,* dan leukimia [5]. Pengujian performansi dilakukan dengan cara membandingkan hasil akurasi klasifikasi data antara penggunaan kedua metode klasifikasi tersebut, mana yang lebih baik jika digabungkan dengan *feature extraction Principal Component Analysis* (PCA). Penelitian ini juga mempertimbangkan hasil performansi dengan *precission, F1-score,* dan *recall.*

2. Studi Terkait

Terdapat beberapa penelitian sebelumnya yang menganalisis data DNA *microarray* dengan menggunakan metode klasifikasi *Naïve Bayes* dan *Support Vector Machine* (SVM). Haseeb Azzawi, et al. [10] menggunakan *Support Vector Machine* (SVM) dan *Relief-F* pada dataset *lung*. Hasil akurasi yang didapatkan adalah 97.33%. Adiwijaya, Untari N. Wisesty, E. Lisnawati, A. Aditsania dan Dana S. Kusumo [26] menganalisa penggunaan reduksi dimensi *Principal Component Analysis* (PCA) dengan metode klasifikasi *Support Vector Machine* (SVM) dan Levenberg Marquardt based Back Propagation (LMBP) untuk klasifikasi DNA microarray. Penelitian ini mendapatkan akurasi 86.67% untuk dataset Central Nervous System, 92.86% untuk dataset colon, 96.88% untuk dataset *lung* dan 100% untuk dataset leukimia, *ovarian*, dan prostat.

C. ArunKumar et al. [11] menganalisa penggunaan reduksi dimensi SVM-RFE dan FSS dengan metode klasifikasi *Support Vector Machine* (SVM) pada dataset breast dan leukimia. Hasil akurasi dengan menggunakan SVM-RFE adalah 83.51% untuk dataset *breast* dan 97.22% untuk dataset leukimia. Sedangkan hasil akurasi dengan menggunakan FSS adalah 65.98% untuk dataset *breast* dan 81.94% untuk dataset leukimia.

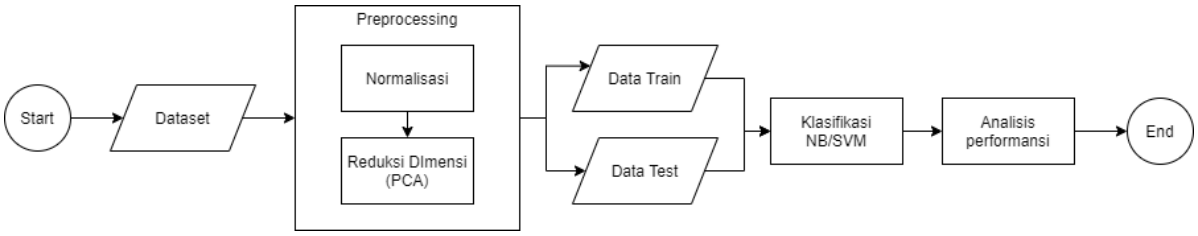
Prayoga, B [7] menganalisa penggunaan *Hybrid feature selection* yaitu IG-GA dengan metode klasifikasi *Naïve Bayes* pada dataset *colon, ovarian, breast, lung,* dan *prostate.* Hasil akurasi yang didapatkan adalah 91.8% untuk dataset *colon,* 58.94% untuk dataset *prostate*, 100% untuk dataset *lung,* 83.47% untuk dataset *breast,* dan 100% untuk dataset *ovarian*.

Terdapat penelitian juga menggunakan metode *feature selection* PCA untuk data DNA *microarray*. Rizky Pujianto, Adiwijaya dan Aniq A. Rahmawati pada penelitian [8] dengan judul "Analisis Ekstraksi Fitur Principle Component Analysis pada Klasifikasi Microarray Data Menggunakan Classification and Regression Trees" menganalisis penggunaan reduksi dimensi *Principal Component Analysis* (PCA) dengan metode klasifikasi *Classification and Regression Tree* (CART) untuk klasifikasi 8 DNA *microarray*. Dalam penelitian ini, hasil akurasi yang didapatkan adalah 93.9% untuk dataset *lung* dan 97% untuk dataset leukimia.

3. Sistem yang Dibangun

3.1 Skema Umum

Pada penelitian ini digunakan metode *Principal Component Analysis* (PCA) untuk proses reduksi dimensi. Sedangkan untuk metode klasifikasi digunakan *Naïve Bayes* dan *Support Vector Machines* (SVM) yang nantinya akan dibandingkan untuk mencari metode mana yang hasil performansinya lebih optimal.



Gambar 1 *Flowchart* Sistem

Untuk mengklasifikasikan kanker berdasarkan data microarray, skema umum dibentuk untuk mendukung proses klasifikasi. Pada *flowchart* diatas proses pertama yang dilakukan adalah persiapan dataset yang akan digunakan. Dataset yang dipakai untuk mengklasifikasikan microarray pada tugas akhir ini merupakan dataset kanker. Proses selanjutnya adalah tahap *preprocessing*. Dalam penelitian ini digunakan dua tahap *preprocessing* yaitu normalisasi dan *feature extraction*. Dataset yang digunakan untuk tugas akhir ini akan dinormalisasi menggunakan *min-max normalization*. Kemudian dilanjut dengan *feature extraction* menggunakan *Principal Component Analysis* (PCA) yang bertujuan untuk mengurangi kompleksitas data. Data yang telah melaluis proses *feature extraction* selanjutnya akan diklasifikasi menggunakan *Naïve Bayes* dan *Support Vector Machine* (SVM). Pengujian performansi dilakukan dengan cara membandingkan hasil akurasi antara penggunaan kedua metode klasifikasi tersebut untuk mencari metode klasifikasi mana yang lebih baik jika digabungkan dengan *feature extraction* PCA. Hasil dari *precision, F1-Score,* dan *recall* juga akan dianalisis untuk mendapatkan gambaran performansi secara keseluruhan.

3.2 Dataset

Pada penelitian ini penulis menggunakan data *microarray* yang berasal dari Kent-Ridge Biomedical Dataset [5]. Data microarray yang diambil terdiri dari data *breast cancer, colon tumor, lung cancer, ovarian cancer* dan leukimia.

Tabel 1 Spesifikasi Data *Microarray* [5]

Data	Jumlah Sampel	Jumlah Fitur	Jumlah Kelas
Breast Cancer	97	24482	2 (relapse, non-relapse)
Colon Tumor	62	2000	2 (negative, positive)
Lung Cancer	181	12533	2 (mesothelioma, ADCA)
Ovarian Cancer	252	15154	2 (normal, cancer)
Leukimia	72	7129	2 (AML,ALL)

3.3 *Pre-processing*

Proses *pre-processing* diperlukan karena terdapat beberapa permasalahan dari dataset yang digunakan yaitu atribut yang sangat banyak, dan nilai data yang harus dilakukan standarisasi. Masalah tersebut dapat menyebabkan hasil performansi data menjadi kurang optimal. Untuk mengatasi hal tersebut, data DNA *microarray* pada penelitian ini akan dinormalisasi dan direduksi dimensinya menggunakan *Principal Component Analysis* (PCA).

Normalisasi

Normalisasi digunakan untuk menghilangkan dan mengurangi redudansi data agar kinerja model klasifikasi menjadi lebih baik. Pada penelitian ini digunakan metode *Min-Max Normalization*. *Min-Max Normalization* dilakukan untuk mengubah skala range nilai kedalam range 0 sampai 1 [9]. Rumus untuk *Min-Max Normalization* terdapat pada persamaan (1).

$$X = \frac{Xi - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

Dimana X adalah nilai data setelah dilakukan normalisasi, Xi adalah nilai data ke-I, X_{min} adalah nilai minimum pada data, dan X_{max} adalah nilai maksimum pada data dalam atribut yang dinormalisasi. Data yang telah dinormalisasi ini tidak akan terlihat perbedaan nilai yang sangat jauh, karena nilai terbesarnya menjadi satu dan nilai terkecilnya nol.

Feature Extraction menggunakan PCA

Principal Component Analysis (PCA) adalah algoritma untuk mereduksi dimensi data yang dilakukan dengan cara menggabungkan atau memproyeksikan inti setiap fitur dengan membentuk subset fitur baru sehingga dimensi

fitur menjadi lebih kecil dan mengubah kumpulan dimensi yang saling berkolerasi menjadi dimensi yang tidak saling berkorelasi [29]. *Principal Component Analysis* (PCA) dapat menyederhanakan data dengan cara mentransformasi data menjadi suatu variabel set baru. *Principal Component Analysis* (PCA) juga dapat mendeteksi beberapa jenis kanker dan memilih fitur mana yang relevan [12]. Metode *feature extraxction* PCA bekerja mengurangi dimensi data tetapi tanpa menghilangkan informasi yang relevan dari data tersebut.

Hal yang dilakukan ketika menggunakan algoritma PCA ini adalah dengan mencari data $X^*_{i,j}$ berdimensi *m x n*, dimana *m* adalah banyaknya sample dan *n* adalah banyak atributnya. Dengan mengurangi semua nilai menggunakan teknik zero-mean, nilai $X_{i,j}$ pada matriks X, dengan nilai rata-rata nilai matriks $\bar{X}$ [13]. Teknik zero-mean adalah proses agar data tersebut berdistribusi normal standar. Ini dilakukan karena menurut teorema limit pusat, jika data yang diambil mendekati jumlah populasi maka datanya semakin mendekati distribusi normal [5]. Jadi hasil dari perhitungan dapat mewakili data sejumlah populasi.

$$X^*_{i,j} = X_{i,j} - \bar{X} \qquad (2)$$

Setelah itu adalah mencari nilai kovarian dari matriks $X_{i,j}$, C_x merupakan nilai dari kovarian yang dicari.

$$C_x = \frac{1}{m-1} \cdot X^{*T}_{i,j} \cdot X^*_{i,j} \qquad (3)$$

C_x adalah matriks kovariansi dari *j x j*, dan *m* adalah jumlah dari sample. Kemudian selanjutnya adalah mencari nilai eigen

$$|C_x - \lambda I| = 0 \text{ dan } (C_x - \lambda I) \cdot v = 0 \qquad (4)$$

dimana I adalah matriks identitas, $\boldsymbol{\lambda}$ adalah nilai eigen dan v adalah vektor eigen. Vektor eigen inilah yang menjadi komponen utama untuk menentukan variabel baru. Untuk menentukan jumlah variabel baru yang digunakan tergantung dari persentasi kontribusi komulatif variansi Vr,

$$V_r = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{D} \lambda_j} \cdot 100\% \qquad (5)$$

dimana *D* adalah jumlah atribut awal dan *r* adalah jumlah komponen yang dipilih [13].

3.4 Klasifikasi

Klasifikasi menggunakan *Naïve Bayes*

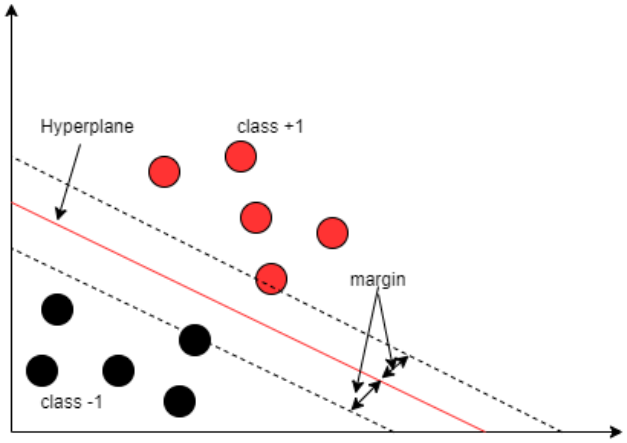
Klasifikasi *Naïve Bayes* adalah klasifikasi sederhana dimana setiap data memiliki sifat saling independent [16]. Metode *Naïve Bayes* telah terbukti dapat bekerja dengan baik untuk masalah klasifikasi [14]. Pada penelitian ini digunakan tipe *Gaussian Naïve Bayes*, karena tipe ini memiliki kinerja lebih baik untuk data yang bersifat kontinu [20]. Tahapan dari proses klasifikasi ini yaitu menghitung jumlah kelas, lalu mengetahui berapa banyak kelas, menghitung jumlah kasus dari masing-masing kelas, mengalikan semua variabel pada kelas dan membandingkan hasil diantara masing-masing kelas [27]. Berikut merupakan persamaan *Gaussian Naïve Bayes*:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}} \qquad (6)$$

Dimana x_i merupakan data pelatihan berisi atribut kontinu, μ_y merupakan rata-rata dari nilai dalam x terkait dengan kelas y, dan σ_y^2 merupakan standar deviasi dari nilai-nilai dalam x terkait dengan kelas y.

Klasifikasi menggunakan SVM

Support Vector Machine (SVM) merupakan salah satu metode *supervised learning*. SVM bertujuan untuk mencari lokasi *hyperplane* optimal yang digunakan sebagai pemisah kelas pada ruang input dengan cara memaksimalkan jarak antar kelas dengan mencari margin atau biasa disebut *support vector* [19]. Terdapat ilustrasi dalam pencarian *hyperplane* terbaik, sebagai berikut:



Gambar 2 Ilustrasi pencarian *hyperplane* terbaik

Pada gambar 2, menjelaskan mengenai ilustrasi pencarian *hyperplane* terbaik untuk memisahkan dua buah kelas. Terdapat dua buah kelas yaitu -1 dan +1 yang disebut dengan *pattern*. Sedangkan jarak terdekat antara *pattern* dengan *hyperplane* untuk tiap kelasnya disebut dengan *margin* dan diilustrasikan dengan garis putus-putus. *Hyperplane* terbaik dapat ditentukan dengan mencari *pattern* yang mempunyai jarak terdekat dengan hyperplane untuk kedua kelasnya. Jarak terdekat antara *hyperplane* dengan *pattern* itulah yang disebut sebagai support vector [25]. Pada penelitian ini digunakan SVM dengan linear kernel, karena linear kernel bekerja paling baik pada komponen yang kecil maupun besar dibandingkan dengan kernel lainnya [21]. Dengan parameter yang digunakan adalah C = 1 dan random state = 0. Berikut merupakan persamaan dari SVM:

$$\vec{w} \cdot \vec{x} + b = 0 \quad (8)$$

Sehingga,
Untuk kelas +1

$$\vec{w} \cdot \vec{x} + b \geq 1 \quad (9)$$

Untuk kelas -1

$$\vec{w} \cdot \vec{x} + b \geq -1 \quad (10)$$

Dimana $\vec{w}$ sebagai normal bidang, x adalah data input dan b adalah posisi bidang relatif. Agar dapat menemukan *hyperplane* yang optimal, diperlukan pencarian titik minimal, yaitu sebagai berikut.

$$min_w \frac{1}{2} (\left|\overrightarrow{|w|}\right|)^2 \quad (11)$$

$$y_i(\overrightarrow{x_i}, \vec{w} + b) - 1 \geq 0 \quad (12)$$

Dimana $\overrightarrow{x_i}$ merupakan data masukan ke-i untuk mencari nilai parameter w dan b, sedangkan nilai y_i merupakan data keluaran. Lalu fungsi keputusan dengan persamaan:

$$Linear : K(\overrightarrow{x_i}, \overrightarrow{x_j}) = \phi(\overrightarrow{x_i})\phi\overrightarrow{x_j}) \quad (13)$$

$$Non - Linear : K(\overrightarrow{x_i}, \overrightarrow{x_j}) = \exp(-\frac{|\overrightarrow{x_i}, \overrightarrow{x_j}|}{2\sigma^2})^2 \quad (14)$$

Dimana x_i dan x_j adalah vector dari ruang fitur dan σ adalah parameter bebas.

3.5 Analisis Performansi

Setelah semua proses dilakukan, proses terakhir adalah analisis performansi yang digunakan untuk menganalisis apakah sistem yang dibuat dan metode yang diusulkan sudah efisien. Digunakan perhitungan akurasi, yang mengacu pada seberapa sering pengklasifikasi memprediksi kelas yang benar [15]. Selain akurasi, penelitian juga mempertimbangkan hasil dari *precision, recall, dan f1-score*. *Precision* merupakan nilai kesesuaian antara informasi yang dibutuhkan dengan hasil keseluruhan data. *Recall* merupakan nilai keberhasilan sistem dalam menemukan kembali sebuah informasi. *F1-Score* merupakan perbandingan rata-rata antara *precision* dan *recall* yang dibobotkan. Perhitungan *akurasi, precision, recall, dan f1-score* dapat dilihat pada persamaan berikut.

$$Akurasi = \frac{jumlah\ prediksi\ benar}{jumlah\ data} \times 100\% \quad (15)$$

$$Precision = \frac{jumlah\ prediksi\ benar\ data\ kanker}{jumlah\ data\ yang\ diprediksi\ benar} \times 100\% \quad (16)$$

$$Recall = \frac{jumlah\ prediksi\ benar\ data\ kanker}{jumlah\ data\ yang\ memang\ benar} \times 100\% \quad (17)$$

$$F1 - score = \frac{2x(precisionxrecall)}{precision+recall} \times 100\% \quad (18)$$

4. Evaluasi Sistem

Dataset yang tertera pada tabel 1 akan melalui tahap feature extraction menggunakan PCA untuk mengurangi redudansinya. Setelah itu, dilakukan implementasi model klasifikasi menggunakan metode *Naïve Bayes* dan SVM. Perhitungan nilai akurasi, *precision, recall,* dan *f1-score* dilakukan dengan cara *k-fold cross-validation*. Agar hasil skor optimal, nilai K ditetapkan dengan K=5.

4.1 Perbandingan Performansi antara *Naïve Bayes* dan SVM

Performa PCA dan *Naïve Bayes*

Pada skenario ini, akan digunakan *feature extraction* PCA yang dipadukan dengan metode klasifikasi *Naïve Bayes.* Terdapat 6 skenario banyak n-komponen yang berbeda yang diujikan yang dipilih secara empiris, dimana banyak komponen yang diuji adalah 3, 5, 10, 20, 30, dan 40. Setelah itu akan dilakukan *K-Fold Cross Validation* dengan jumlah K yang akan diuji didasari oleh penelitian sebelumnya [7], yaitu K=5.

Tabel 2 Hasil Akurasi Uji Data *Microarray* Menggunakan PCA dan NB

Data	**Akurasi**					
	Banyak Komponen					
	3	**5**	**10**	**20**	**30**	**40**
Colon	64.5%	80.8%	80.8%	**81.2%**	79.4%	79.4%
Lung	89.7%	**90.8%**	90.2%	89.2%	85.3%	85.9%
Ovarian	75.6%	85.9%	92.1%	**94.5%**	92.9%	92.1%
Breast	**56.3%**	55.3%	54.3%	55.6%	53.3%	52.3%
Leukimia	73.8%	80.7%	**90.4%**	69.1%	66.2%	66.2%

Tabel 3 Hasil *Precision* Uji Data *Microarray* Menggunakan PCA dan NB

Data	***Precision***					
	Banyak Komponen					
	3	**5**	**10**	**20**	**30**	**40**
Colon	37.5%	**70.3%**	61%	61%	69.3%	69.3%
Lung	**89.6%**	**89.6%**	88.9%	87%	85.4%	86.2%
Ovarian	67.8%	72.9%	74.7%	69.2%	74.3%	**75.4%**
Breast	**40%**	**40%**	40%	10%	**40%**	20%
Leukimia	38%	42%	**56%**	42.9%	38.6%	38.6%

Tabel 4 Hasil *Recall* Uji Data *Microarray* Menggunakan PCA dan NB

Data	***Recall***					
	Banyak Komponen					
	3	**5**	**10**	**20**	**30**	**40**
Colon	37.3%	56%	58.7%	**67.5%**	45.3%	45.3%
Lung	92.3%	99.4%	99.4%	**100%**	**100%**	**100%**
Ovarian	66.4%	76.8%	**77.6%**	68%	76.2%	75.8%
Breast	**6.7%**	5.7%	6.3%	4%	2.7%	1%
Leukimia	38%	45.7%	47%	50%	**58.7%**	57.3%

Tabel 5 Hasil *F1-score* Uji Data *Microarray* Menggunakan PCA dan NB

Data	***F1-score***					
	Banyak Komponen					
	3	**5**	**10**	**20**	**30**	**40**
Colon	35.8%	58.7%	58.7%	**61.9%**	49%	49%
Lung	92.4%	**93%**	92.6%	90.7%	89.6%	90.9%
Ovarian	66.3%	74.1%	**75.7%**	68.5%	75%	75.2%
Breast	**11.1%**	9.7%	10.9%	10%	9.9%	1.9%
Leukimia	35.7%	42.6%	**50.9%**	45.6%	43.6%	43.6%

Hasil pengujian data DNA *Microarray* dengan menggunakan metode PCA dan *Naïve Bayes* menunjukan bahwa 3 dataset *cancer* yang terdiri dari *lung cancer, ovarian cancer,* dan leukimia menggunakan kurang dari 30 komponen mendapatkan akurasi lebih dari 90%. Akurasi tertinggi didapatkan dari dataset *ovarian cancer* sebesar 94.5% pada pemilihan 20 komponen. Akurasi terendah didapatkan dari dataset *breast cancer* sebesar 56.3% pada pemilihan 10 komponen. Pada dataset *colon tumor*, nilai akurasi terbesarnya adalah 81.2% yang diperoleh ketika menggunakan pemilihan 20 komponen. Selanjutnya pada dataset *lung cancer*, nilai akurasi terbesarnya adalah 90.8% yang diperoleh ketika menggunakan pemilihan 5 komponen. Dan pada dataset leukimia, nilai akurasi terbesarnya adalah 90.4% ketika menggunakan pemilihan 10 komponen. Hasil akurasi dari 5 dataset diatas mengalami penurunan ketika menggunakan pemilihan lebih dari 20 komponen.

Sebagai tambahan, hasil *precision* paling besar dimiliki oleh dataset *lung cancer* yaitu sebesar 89.6%. Hasil *recall* dan *F1-score* terbaik juga didapatkan oleh dataset *lung cancer* dengan tingkat *recall* mencapai 100% dan *F1-score* sebesar 93%. Analisis dari setiap data menunjukan bahwa nilai n-komponen yang lebih besar tidak selalu meningkatkan akurasi. Dataset *breast cancer* memiliki nilai *presicion, recall,* dan *F1-score* yang sangat rendah, ini karena dataset tersebut *noise* (data yang tidak informatif) digabungkan ke dalam data yang dihasilkan dari proses reduksi dimensi. Data *noise* memiliki nilai eigen (varians) yang relative kecil, sehingga dapat mengganggu data informatif lainnya dan mengurangi hasil klasifikasi.

Performa PCA dan *Support Vector Machine*

Pada skenario ini, akan digunakan *feature extraction* PCA yang dipadukan dengan metode klasifikasi *Support Vector Machine.* Terdapat 5 skenario banyak n-komponen yang berbeda yang diujikan yang dipilih secara empiris, dimana banyak komponen yang diuji adalah 3, 5, 10, 20, 30, dan 40. Setelah itu akan dilakukan *K-Fold Cross Validation* dengan jumlah K yang akan diuji didasari oleh penelitian sebelumnya[7], yaitu K=5.

Tabel 6 Hasil Akurasi Uji Data *Microarray* Menggunakan PCA dan SVM

Data	**Akurasi**					
	Banyak Komponen					
	3	**5**	**10**	**20**	**30**	**40**
Colon	59.7%	71%	76%	**77.7%**	**77.7%**	**77.7%**
Lung	89.7%	90.8%	**95.7%**	94%	90.2%	90.2%
Ovarian	74.9%	83.1%	90.6%	91.8%	**92.5%**	**92.5%**
Breast	44.6%	45.6%	47.7%	**48.8%**	45.7%	48.7%
Leukimia	59%	63%	75.3%	**86.1%**	81.9%	80.6%

Tabel 7 Hasil *Precision* Uji Data *Microarray* Menggunakan PCA dan SVM

Data	***Precision***					
	Banyak Komponen					
	3	**5**	**10**	**20**	**30**	**40**
Colon	30%	**75%**	70%	71%	71%	71%
Lung	89.6%	90%	**93.3%**	91.6%	87.8%	87.8%
Ovarian	67.4%	74.3%	74.7%	75.7%	**76.9%**	**76.9%**
Breast	**56.7%**	54.2%	50.6%	54.2%	54.2%	54.2%
Leukimia	13.3%	39.3%	**50%**	**50%**	43.9%	43%

Tabel 8 Hasil *Recall* Uji Data *Microarray* Menggunakan PCA dan SVM

Data	***Recall***					
	Banyak Komponen					
	3	**5**	**10**	**20**	**30**	**40**
Colon	15.3%	37.3%	37.3%	**41.3%**	**41.3%**	**41.3%**
Lung	98.3%	99.4%	**100%**	**100%**	**100%**	**100%**
Ovarian	63%	74.2%	78.4%	**78.8%**	**78.8%**	**78.8%**
Breast	18%	21.8%	**27.8%**	24%	22.3%	22.8%
Leukimia	18.3%	32.3%	41.7%	**47%**	**47%**	44.3%

Tabel 9 Hasil *F1-score* Uji Data *Microarray* Menggunakan PCA dan SVM

Data	***F1-score***					
	Banyak Komponen					
	3	**5**	**10**	**20**	**30**	**40**

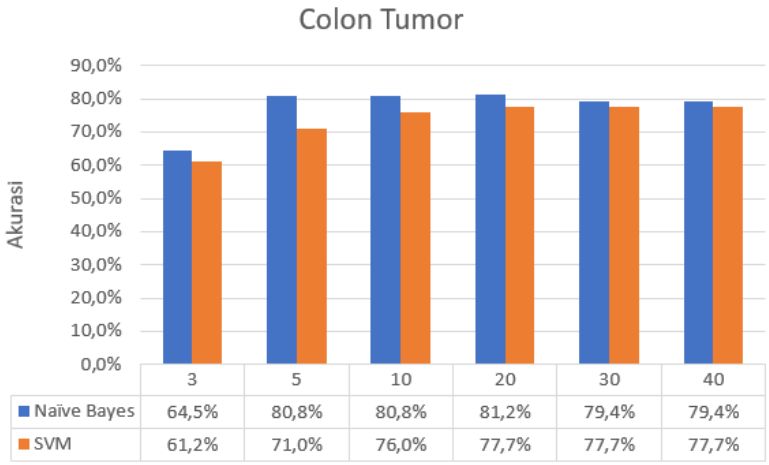
Colon	16.6%	47.4%	45.7%	**48.4%**	**48.4%**	**48.4%**
Lung	92.4%	93.1%	**96%**	94.9%	92.5%	92.5%
Ovarian	66.9%	73.9%	76.1%	**77%**	**77%**	**77%**
Breast	22.8%	29.2%	31.8%	**32.5%**	27.1%	29.3%
Leukimia	13.7%	25.8%	42.2%	**47.2%**	44.2%	42.6%

Hasil pengujian data DNA *Microarray* dengan menggunakan metode PCA dan *Support Vector Machine* menunjukan bahwa 2 dataset *cancer* yang terdiri dari *lung cancer* dan *ovarian cancer* menggunakan pemilihan 10 hingga 40 komponen mendapatkan akurasi lebih dari 90%. Akurasi tertinggi didapatkan dari dataset *lung cancer* sebesar 95.7% pada pemilihan 10 komponen. Akurasi terendah didapatkan dari dataset *breast cancer* sebesar 48.8% pada pemilihan 10 komponen. Pada dataset *colon tumor*, nilai akurasi terbesarnya adalah 77.7% yang diperoleh dengan stabi; ketika menggunakan pemilihan 20 hingga 40 komponen. Selanjutnya pada dataset *ovarian cancer*, nilai akurasi terbesarnya adalah 92.5% yang diperoleh ketika menggunakan pemilihan 30 dan 40 komponen. Dan pada dataset leukimia, nilai akurasi terbesarnya adalah 86.1% ketika menggunakan pemilihan 20 komponen.

Sebagai tambahan, hasil *precision* paling besar dimiliki oleh dataset *lung cancer* yaitu sebesar 93.3%. Hasil *recall* dan *F1-score* terbaik juga didapatkan oleh dataset *lung cancer* dengan tingkat *recall* mencapai 100% dan *F1-score* sebesar 96%. Analisis dari setiap data menunjukan bahwa nilai n-komponen yang lebih besar tidak selalu meningkatkan akurasi. Dataset *breast cancer* memiliki nilai *presicion*, *recall*, dan *F1-score* yang sangat rendah, ini karena dataset tersebut *noise* (data yang tidak informatif).

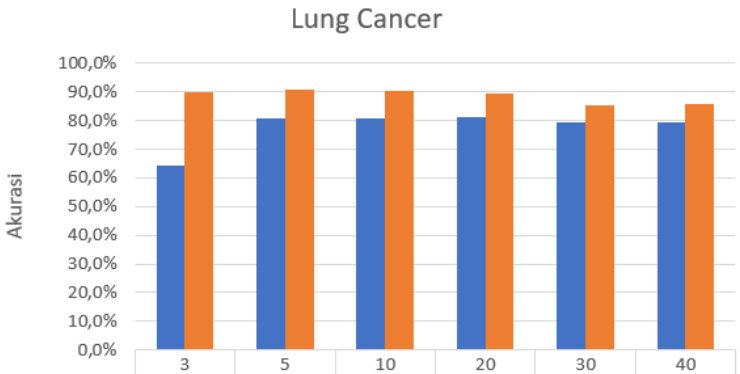
4.2 Pengaruh Metode *Naïve Bayes* dan SVM terhadap akurasi

Dilakukan pengujian lima dataset kanker yang diuji menggunakan parameter yang telah ditentukan sebelumnya. Nilai parameter yang diujikan yang dipilih secara empiris. Hasil penelitian dapat dilihat pada gambar berikut.



Gambar 3 Perbandingan Akurasi NB dan SVM menggunakan PCA pada dataset *Colon Tumor*

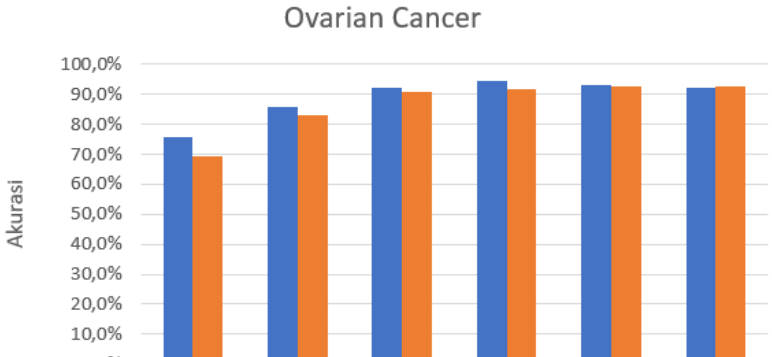
Pada gambar diatas, dapat terlihat bahwa hasil pengujian pada lima data kanker dengan penggunaan metode *Naïve Bayes* dan SVM dapat meningkatkan akurasi. Data yang kinerja klasifikasinya paling baik adalah dataset *Lung Cancer* dan dataset *Ovarian Cancer*. Hasil akurasi pada semua data mengalami kenaikan dan juga penurunan yang berbeda. Berdasarkan gambar 3, akurasi data *Colon Tumor* mengalami kenaikan pada penggunaan 3 hingga 20 komponen, lalu dengan menggunakan metode *naïve bayes* mengalami penurunan pada penggunaan 30 hingga 40 komponen sedangkan dengan menggunakan metode SVM mengalami akurasi yang stabil pada penggunaan 20 hingga 40 komponen. Dengan menggunakan metode klasifikasi *naïve bayes* akurasi paling tinggi didapatkan ketika menggunakan 20 komponen yaitu sebesar 81.2%, sedangkan dengan metode klasifikasi SVM akurasi paling tinggi didapatkan ketika menggunakan 20 hingga 40 komponen yaitu sebesar 77.7%.



	3	5	10	20	30	40
Naïve Bayes	64,5%	80,8%	80,8%	81,2%	79,4%	79,4%
SVM	89,7%	90,8%	90,2%	89,2%	85,3%	85,9%

Gambar 4 Perbandingan Akurasi NB dan SVM menggunakan PCA pada dataset *Lung Cancer*

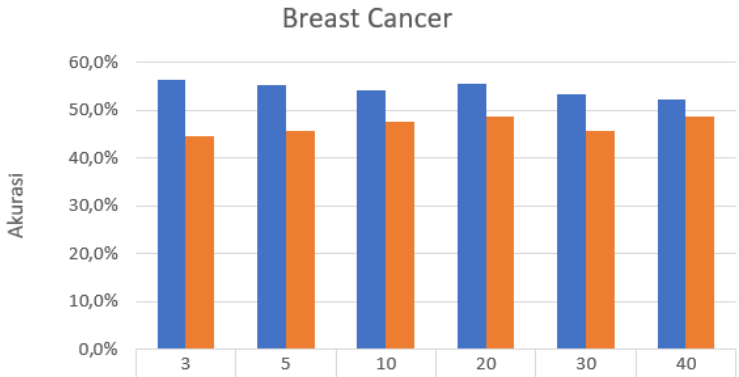
Berdasarkan gambar 4, dengan menggunakan metode klasifikasi *naïve bayes* akurasi data *Lung Cancer* mengalami kenaikan pada penggunaan 3 hingga 5 komponen, lalu stabil hingga 10 komponen dan mengalami penurunan pada penggunaan 20 komponen lalu kembali stabil pada penggunaan 30 hingga 40 komponen. Akurasi tertinggi pada penggunaan metode ini sebesar 81.2% ketika menggunakan 20 komponen. Dengan menggunakan metode klasifikasi SVM akurasi data *Lung Cancer* mengalami kenaikan pada penggunaan 3 hingga 5 komponen, lalu mengalami penurunan pada penggunaan 10 hingga 30 komponen dan kembali naik pada penggunaan 40 komponen. Akurasi tertinggi pada penggunaan metode ini sebesar 90.8% pada penggunaan 5 komponen.



	3	5	10	20	30	40
Naïve Bayes	75,6%	85,9%	92,1%	94,5%	92,9%	92,1%
SVM	69,3%	83,1%	90,6%	91,8%	92,5%	92,5%

Gambar 5 Perbandingan Akurasi NB dan SVM menggunakan PCA pada dataset *Ovarian Cancer*

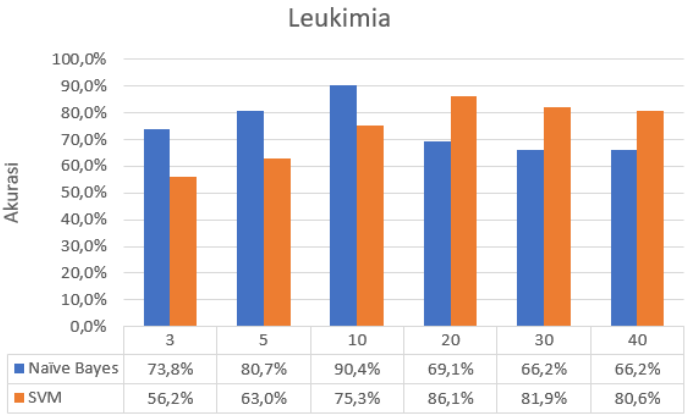
Berdasarkan gambar 5, dengan menggunakan metode klasifikasi *naïve bayes* akurasi data *Ovarian Cancer* mengalami kenaikan pada penggunaan 3 hingga 20 komponen lalu mengalami penurunan pada penggunaan 30 hingga 40 komponen. Akurasi tertinggi pada penggunaan metode ini sebesar 94.5% ketika menggunakan pemilihan 20 komponen. Dengan menggunakan metode klasifikasi SVM akurasi data *Ovarian Cancer* juga mengalami kenaikan pada penggunaan 3 hingga 20 komponen, lalu stabil pada penggunaan 30 hingga 40 komponen. Akurasi tertinggi pada penggunaan metode ini sebesar 92.5% pada penggunaan 30 hingga 40 komponen.



	3	5	10	20	30	40
Naïve Bayes	56,3%	55,3%	54,3%	55,6%	53,3%	52,3%
SVM	44,50%	45,60%	47,70%	48,80%	45,70%	48,70%

Gambar 6 Perbandingan Akurasi NB dan SVM menggunakan PCA pada dataset *Breast Cancer*

Berdasarkan gambar 6, dengan menggunakan metode klasifikasi *naïve bayes* akurasi data *Breast Cancer* mengalami penurunan pada penggunaan 3 hingga 10 komponen lalu mengalami penaikan pada penggunaan 20 komponen dan kembali mengalami penurunan pada penggunaan 30 hingga 40 komponen. Dengan menggunakan klasifikasi SVM akurasi data *Breast Cancer* mengalami kenaikan pada penggunaan 3 hingga 20 komponen, lalu mengalami penurunan pada penggunaan 30 komponen dan kembali naik pada penggunaan 40 komponen. Akurasi pada data *Breast Cancer* dengan menggunakan kedua metode tersebut terbilang rendah, yaitu *naïve bayes* sebesar 56.3% dan SVM sebesar 48.8%.



	3	5	10	20	30	40
Naïve Bayes	73,8%	80,7%	90,4%	69,1%	66,2%	66,2%
SVM	56,2%	63,0%	75,3%	86,1%	81,9%	80,6%

Gambar 7 Perbandingan Akurasi NB dan SVM menggunakan PCA pada dataset Leukimia

Berdasarkan gambar 7, dengan menggunakan metode klasifikasi *naïve bayes* akurasi data Leukimia mengalami kenaikan pada penggunaan 3 hingga 10 komponen lalu mengalami penurunan pada penggunaan 20 komponen. Akurasi stabil pada penggunaan 30 hingga 40 komponen. Akurasi tertinggi pada pengunaan metode ini sebesar 90.4% ketika menggunakan pemilihan 10 komponen. Dengan menggunakan metode klasifikasi SVM akurasi data Leukimia mengalami kenaikan pada penggunaan 3 hingga 20 komponen, lalu mengalami penurunan pada penggunaan 20 hingga 40 komponen. Akurasi tertinggi pada penggunaan metode ini sebesar 86.1% pada penggunaan 20 komponen.

Dari hasil yang didapatkan, akurasi mengalami kenaikan dan penurunan yang berbeda pada setiap metode klasifikasi yang digunakan. Dikarenakan tidak semua dataset cocok dengan klasifikasi yang digunakan. Jika grafik yang tertera pada gambar 3 hingga gambar 7 diperhatikan secara seksama, hasil klasifikasi pada dataset *colon tumor, ovarian cancer, breast cancer* dan leukimia lebih baik jika menggunakan metode klasifikasi *Naïve Bayes*. Sedangkan pada dataset *lung cancer* hasil klasifikasinya lebih baik jika menggunakan metode klasifikasi SVM. Klasifikasi *Naïve Bayes* dan SVM memiliki cara kerja yang berbeda dan keduanya sensitif terhadap optimasi parameter (pemilihan parameter yang berbeda dapat mengubah outputnya secara signifikan). Oleh karena itu, parameter yang digunakan dalam penelitian ini menunjukan bahwa kinerja klasifikasi *Naïve Bayes* memiliki hasil yang lebih baik dibandingkan dengan SVM.

Lebih jauh dari itu, pemilihan komponen juga merupakan salah satu parameter penting dalam penelitian ini, sehingga ukuran komponen yang telah ditentukan berpengaruh terhadap kinerja klasifikasi. Dari hasil pengujian yang telah dilakukan menunjukan bahwa semakin besar nilai komponen yang digunakan belum tentu menghasilkan akurasi yang lebih tinggi. Ketika akurasi mengalami kenaikan, maka jumlah komponen yang digunakan adalah komponen yang optimal.

Perlu diperhatikan juga bahwa, kecocokan antara metode *feature extraction* dengan metode klasifikasinya berpengaruh terhadap besarnya akurasi. Seperti pada dataset breast cancer, hasil pengujian yang didapatkan dengan menggunakan metode klasifikasi *Naïve Bayes* dan SVM terbilang sangat rendah yaitu 48.8% untuk *Naïve bayes* dan 56.3% untuk SVM yang berarti kedua metode tersebut belum mampu untuk menangani dataset ini, dikarenakan dataset ini adalah data yang memiliki kompleksitas data yang paling tinggi dibandingkan data lainnya. Dataset *breast cancer* berbasis microarray memiliki sample yang sangat sedikit dengan jumlah fitur yang jauh lebih besar dibandingkan dengan dataset lainnya [22], dikarenakan untuk melakukan pengujiannya memerlukan waktu yang lama dan harga yang mahal [23]. Ukuran sampel juga berkurang karena faktor klinis yang hilang dalam dataset microarray [24]. Oleh karena itu, metode yang diusul belum mampu bekerja dengan baik untuk metode klasifikasi data *breast cancer*.

5. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, metode *feature extraction* PCA yang digabungkan dengan metode klasifikasi *Naïve Bayes* dan SVM mampu mengklasifikasikan data DNA microarray untuk deteksi kanker. Metode klasifikasi yang paling baik bergantung pada dataset yang digunakan. Hasil dari penelitian menunjukan empat dari lima dataset yang digunakan yaitu *colon tumor, lung cancer, ovarian cancer, breast cancer* dan leukimia mendapatkan hasil akurasi terbaik dari model *Naïve Bayes*. Sehingga dapat disimpulkan bahwa *Naïve* Bayes digabungkan dengan PCA lebih baik dibandingkan dengan SVM yang digabungkan dengan PCA. Tingkat akurasinya sebesar 77-96%, hanya saja pada data *breast cancer* mendapatkan akurasi yang masih rendah yaitu 54.6%. Metode ini baik digunakan untuk data yang memiliki akurasi lebih dari 90% seperti data *lung cancer* (akurasi 90.2%), *ovarian cancer* (akurasi 94.5%), dan leukimia (akurasi 90.4%). Tetapi metode klasifikasi ini belum mampu untuk menangani data *breast cancer* karena mendapatkan akurasi hanya 54.6% maka diperlukan penelitian lebih lanjut. Saran untuk penelitian selanjutnya adalah mengubah nilai-nilai parameter yang sudah ditentukan, serta membandingkan metode klasifikasi lainnya sebagai bentuk pengujian dan perbandingan.

Daftar Pustaka

[1] World Health Organization. Cancer Key Facts. Geneva: World Health Organization; 2018. Available from: https://www.who.int/en/newsroom/fact-sheets/detail/cancer.

[2] Tan Ching Siang, Ting Wai Soon, Shahreen Kasim, Mohd Saberi Mohamad, Chan Weng Howe, Safaai Deris, Zalmiyah Zakaria, Zuraini Ali Shah, and Zuwairie Ibrahim, "A review of cancer classification software for gene expression data," *International Journal of Bio-Science and Bio-Technology*, 7(4):89–108, 2015.

[3] H. Aydadenta and Adiwijaya, "A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest," *Journal of Information Processing Systems*, vol. 14, no. 5, pp. 1167 – 1175, 2018.

[4] Zhao, Huimin, et al. Fault diagnosis method based on principal component analysis and broad learning system. IEEE Access, 2019, 7: 99263-99272.

[5] Elvira biomedical Dataset Repository, "Kent Ridge Biomedical Data Set Repository," 2005. [Online]. Available: https://leo.ugr.es/elvira/DBCRepository//. [Accessed: 02-Feb-2021].

[6] GELADI, Paul; LINDERHOLM, Johan. Principal Component Analysis☆. 2020.

[7] Peryoga, Bintang, Adiwijaya Adiwijaya, and Widi Astuti, "Deteksi Kanker Berdasarkan Data Microarray Menggunakan Metode Naïve Bayes dan Hybrid Feature Selection," JURNAL MEDIA INFORMATIKA BUDIDARMA, 2020, 4.3: 486-494.

[8] Pujianto Rizky; Adiwijaya; Rohmawati, Aniq Atiqi, "Analisis Ekstraksi Fitur Principle Component Analysis pada Klasifikasi Microarray Data Menggunakan Classification And Regression Trees," *eProceedings of Engineering*, 2019, 6.1.

[9] M. Babu and K. Sarkar, "A comparative study of gene selection methods for cancer classification using microarray data," 2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2016.

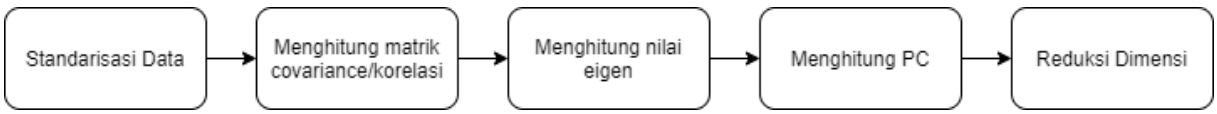
[10] C. ArunKumar, M. P. Sooraj, and S. Ramakrishnan, "A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Microarray Datasets," Procedia Computer Science, vol. 115, pp. 209 – 217, 2017.

[11] Tan Ching Siang, Ting Wai Soon, Shahreen Kasim, Mohd Saberi Mohamad, Chan Weng Howe, Safaai Deris, Zalmiyah Zakaria, Zuraini Ali Shah, and Zuwairie Ibrahim. A review of cancer classification software for gene expression data. International Journal of Bio-Science and Bio-Technology, 7(4):89–108, 2015.

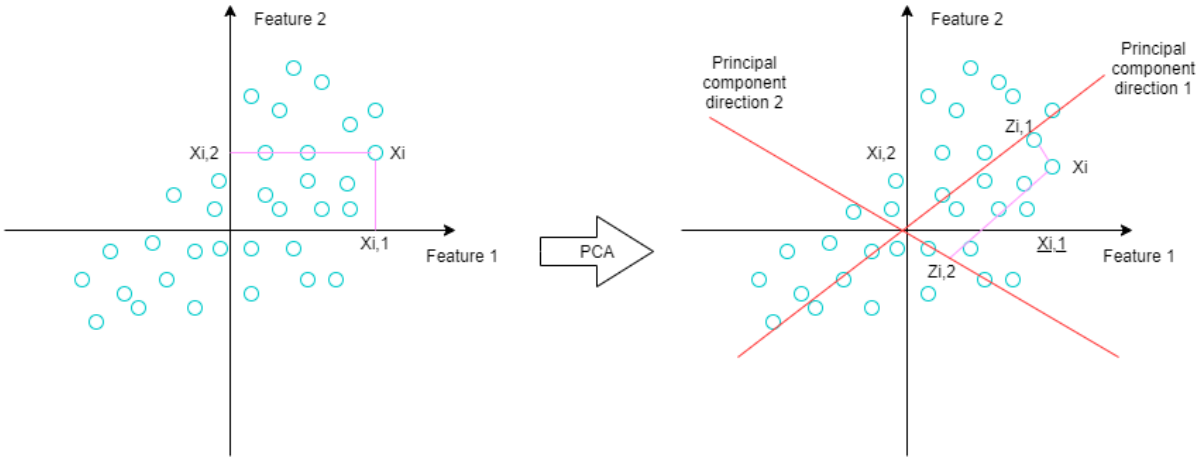
[12] Clayman, Carly L.; Srinivasan, Satish M.; Sangwan, Raghvinder S, "K-means Clustering and Principal Components Analysis of Microarray Data of L1000 Landmark Genes," *Procedia Computer Science*, 2020, 168: 97-104.

[13] X. Jin, A. Xu, R. Bie1, and P. Guo, "Machine Learning Techniques and ChiSquare Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles," *Data Mining for Biomedical Applications*, pp. 106 – 115, 2006.
[14] M. S. Mubarok, A. Adiwijaya, and M. D. Aldhi, "Aspect-based sentiment analysis to review products using Naïve Bayes," in AIP Conference Proceedings, 2017.
[15] M. Nuruddin Qaisar Bhuiyan, M. Shamsujjoha, S. H. Ripon, F. H. Proma, and F. Khan, "Transfer learning and supervised classifier based prediction model for breast cancer," in *Big Data Analytics for Intelligent Healthcare Management, Elsevier*, 2019, pp. 59–86
[16] R. Aziz, C. K. Verma, and N. Srivastava, "A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data," *Genomics Data*, 2016.
[17] Putri, Laila, Mubarok Mohamad, Adiwijaya, "Klasifikasi Sentimen Pada Ulasan Buku Berbahasa Inggris Menggunakan Information Gain Dan Naïve Bayes," *eProceedings of Engineering*, 2017, 4.3.
[18] ALPAYDIN, Ethem. Introduction to machine learning. MIT press, 2020.
[19] C. Campbell, Support Vector Machine and Kernel Methods. 2005.
[20] B. M. and C. P., "An automated technique using Gaussian naïve Bayes classifier to classify breast cancer," *Int. J. Comput. Appl.*, vol. 148, no. 6, pp. 16–21, 2016.
[21] F. A. Ma'Ruf, Adiwijaya, and U. N. Wisesty, "Analysis of the influence of Minimum Redundancy Maximum Relevance as dimensionality reduction method on cancer classification based on microarray data using Support Vector Machine classifier," in *Journal of Physics: Conference Series*, May 2019, vol. 1192, no. 1
[22] A. Saini, J. Hou, and W. Zhou, "Breast cancer prognosis risk estimation using integrated gene expression and clinical data," *Biomed Res. Int.*, vol. 2014, p. 459203, 2014.
[23] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," Proc. Natl. Acad. Sci. U. S. A., vol. 103, no. 15, pp. 5923–5928, 2006.
[24] M. Shi and B. Zhang, "Semi-supervised learning improves gene expression-based prediction of cancer recurrence," Bioinformatics, vol. 27, no. 21, pp. 3017–3023, 2011.
[25] K. Sembiring, "Ide Dasar Support Vector Machine," no. September, pp. 1–28, 2007.
[26] Adiwijaya, Wisesty Untari, et al, "Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification," *Journal of Computer Science*, 2018, 14.11: 1521-1530.
[27] N. Russelia Wassi and M. Dwifebri Purbolaksono, "Classification of Personality based on Beauty Product Reviews Using the TF-IDF and Naïve Bayes (Case Study : Female Daily)," *Journal of Data Science and Its Applications*, vol. 1, no. 1, pp. 57–066, 2020, doi: 10.34818/JDSA.2020.3.61.
[28] T. Kayla Amory and W. Astuti, "Comparative Analysis of Support Vector Machine-Recursive Feature Elimination and Chi-Square on Microarray Classification for Cancer Detection with Naïve Bayes," *Journal of Data Science and Its Applications*, vol. 3, no. 2, pp. 48–056, 2020, doi: 10.34818/JDSA.2020.3.62.
[29] I. Priyono and A. Aditsania, "Cancer Detection based on Microarray Data Classification Using Principal Component Analysis and Functional Link Neural Network," *Journal of Data Science and Its Applications*, vol. 3, no. 2, pp. 85–096, 2020, doi: 10.34818/JDSA.2020.3.52.
[30] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 56–70, May 2020, doi: 10.38094/jastt1224.

Lampiran



Tahapan cara kerja *feature extraction* PCA



Ilustrasi *Principal Component Analysis*

datacolon

	0	1	2	3	4	5	6	7	8	9	
0	0.514957	0.465829	0.408125	0.286098	0.120009	0.413873	0.124713	0.357208	0.496659	0.320553	0.23
1	0.559306	0.608527	0.492652	0.251636	0.122510	0.442248	0.313901	0.361291	0.457976	0.337722	0.09
2	0.147436	0.637132	0.558977	0.349771	0.000000	0.047797	0.029595	0.091152	0.175006	0.159865	0.20
3	0.334197	0.734429	0.638846	0.277217	0.120690	0.102891	0.052773	0.140711	0.218571	0.354236	0.10
4	0.101503	0.263544	0.290522	0.226338	0.146502	0.181060	0.113394	0.393023	0.175494	0.188351	0.24
...	...	...	...	...	...	...	...	...	...	...	...
57	0.235886	0.318223	0.327029	0.037915	0.058303	0.234739	0.117591	0.104136	0.067825	0.224544	0.06
58	0.555304	0.620496	0.642528	0.194880	0.531339	1.000000	0.593787	0.669255	0.277484	0.433574	0.78
59	0.371551	0.238190	0.175834	0.142970	0.062025	0.324130	0.308209	0.000000	0.191386	0.200153	0.00
60	0.333285	0.298994	0.248661	0.198462	0.090765	0.281511	0.107829	0.077736	0.146931	0.106106	0.14
61	0.428749	0.258924	0.198840	0.229408	0.178729	0.464123	0.204328	0.229929	0.210754	0.150928	0.32

62 rows × 2001 columns

Contoh data sebelum menggunakan PCA

	PC 1	PC 2	PC 3	PC 4
0	-2.694492	-1.487057	1.328105	0.369621
1	1.537714	-3.836633	0.034512	-1.214968
2	-6.371380	-2.565739	0.873038	-0.622788
3	-3.902638	-4.353898	-0.113721	1.082370
4	-5.233409	1.457575	0.669368	-0.335411
..	...	...	...	...
57	-7.233186	-0.863575	0.189055	0.593880
58	0.203284	0.802870	-1.442806	2.877332
59	0.435363	-2.682416	-1.591947	-3.331308
60	-3.465826	0.884694	-0.307788	-0.094667
61	-0.898816	0.509663	-2.817999	-1.779404

[62 rows x 5 columns]

Contoh data setelah menggunakan PCA