OPEN ACCESS
Indo-JC
Indonesian Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

IND. JOURNAL ON COMPUTING
VOL. 2, ISSUE. 1, MARET 2017. PP. 109-118
DOI:10.21108/INDOJC.2017.21.169

Analisis Pengaruh Kernel Support Vector Machine (SVM) pada Klasifikasi Data *Microarray* untuk Deteksi Kanker

Rima Diani[#1], Untari Novia Wisesty[#2], Annisa Aditsania[#3]

 # School of Computing, Telkom University
Jl. Telekomunikasi No. 1, Terusan Buah Batu Bandung 40257 Indonesia

[1] rianidima@telkomuniversity.ac.id
[2]untarinw@telkomuniversity.ac.id
[2]aaditsania@telkomuniversity.ac.id

Abstract

Based on data from the Center for Data and Information Ministry of Health, in 2012 about 8.2 million cases of death caused by cancer. Recent developments show that DNA microarray technology is able to handle cancer detection problems early on, but the main disadvantage of microarray is the problem of curse of dimensionality. Analysis of Variance (ANOVA) is one of the feature selection methods that can overcome the weakness of microarray. ANOVA can find an informative gene pair that can assist in the classification process performed by the Support Vector Machine (SVM). In SVM, the kernel trick when learning model is helpful in overcoming the feature space problem. The selection of the kernel affects the resulting accuracy. Through a series of processes such as correlation calculations, feature selection and classification using SVM, accuracy is obtained from the four datasets used. For leukemia and ovarian cancer datasets, the greatest accuracy is generated by the polynomial kernel at 100% and 97.54% with the parameter values of $C = 1.5$ $d = 1$ and $C = 1.5$ $d = 2$. As for the largest lung cancer accuracy dataset obtained from linear kernel that is equal to 100% with the parameter value $C = 1.0$ and for the dataset colon tumor the greatest accuracy is obtained from the RBF kernel of 85.15% with the parameter value $C = 1.5$ $\sigma = 0.5$. The kernel difference that produces the highest accuracy on each dataset is highly dependent on the characteristics of the cancer dataset itself.

Keywords: *cancer detection, DNA microarray, dimension reduction, correlation, analysis of variance, support vector machine, kernel trick*

Abstrak

Berdasarkan data dari Pusat Data dan Informasi Kementrian Kesehatan RI, di tahun 2012 sekitar 8,2 juta kasus kematian disebabkan oleh kanker. Perkembangan terakhir menunjukan bahwa teknologi DNA *microarray* mampu menangani masalah deteksi kanker sejak dini, namun kelemahan utama dari *microarray* adalah masalah *curse of dimensionality.* Analysis of Variance (ANOVA) merupakan salah satu metode seleksi fitur yang dapat mengatasi kelemahan *microarray*. ANOVA dapat menemukan pasangan gen informatif yang dapat membantu dalam proses pengklasifikasian yang dilakukan oleh Support Vector Machine (SVM). Dalam SVM, kernel *trick* saat *learning* model sangat membantu dalam mengatasi masalah *feature space*. Pemilihan kernel berpengaruh terhadap akurasi yang dihasilkan. Melalui serangkaian proses seperti perhitungan korelasi, seleksi fitur dan pengklasifikasian menggunakan SVM, didapatkan akurasi dari empat *dataset* yang digunakan. Untuk *dataset* leukimia dan *ovarian cancer*, akurasi terbesar dihasilkan oleh kernel polynomial yaitu sebesar 100% dan 97,54% dengan nilai parameter $C = 1.5$ $d = 1$ dan $C = 1.5$ $d = 2$. Sedangkan untuk *dataset lung cancer* akurasi terbesar diperoleh dari kernel linear yaitu sebesar 100% dengan nilai parameter $C = 1.0$ dan untuk *dataset colon tumor* akurasi terbesar diperoleh dari kernel RBF sebesar 85,15% dengan nilai parameter $C = 1.5$ $\sigma = 0.5$. Perbedaan

RECEIVED ON FEB 2017. ACCEPTED ON MAR 2017

kernel yang menghasilkan akurasi tertinggi pada setiap *dataset* sangat bergantung kepada karakteristik *dataset* kanker itu sendiri.
Kata Kunci: deteksi kanker, DNA *microarray*, reduksi dimensi, korelasi, analysis of variance, support vector machine, *kernel trick*

I. Pendahuluan

Setiap tahun angka kematian yang diakibatkan kanker terus mengalami kenaikan. Dalam banyak kasus, sel kanker baru ditemukan pada pertumbuhan tumor. Sel kanker tersebut dapat menyusup ke jaringan sekitar tumor sehingga tumor tersebut sudah diklasifikasikan kedalam kanker [5]. Dibandingkan dengan mendeteksi kanker secara tradisional yaitu berdasarkan analisis kemunculan tumor, deteksi kanker melalui ekspresi gen jauh lebih terpercaya dan akurat [3]. Mendeteksi kanker melalui ekspresi gen akan sangat membantu para ahli medis dalam penanganan bagi pasien yang menderita kanker sehingga dapat menekan angka kematian yang kian bertambah.

DNA mengandung sifat dan informasi suatu makhluk hidup, sehingga suatu penyakit dapat diprediksi dari ekspresi DNA [1]. Perkembangan terakhir dalam diagnosis menunjukan DNA *microarray* dapat menggolongkan kanker pada tingkat gen [1]. DNA *microarray* memiliki kemampuan memantau ribuan ekspresi gen secara bersamaan dalam satu kali percobaan. Teknologi ini membantu para peneliti dalam mempelajari berbagai penyakit, terutama kanker.

Beberapa tahun terakhir, DNA *microarray* telah menunjukan pengaruh besar dalam menentukan gen yang menjadi penyebab kanker. Kekurangan utama dalam DNA *microarray* yaitu masalah dimensi (*curse of dimensionality*) [2]. Data DNA *microarray* ini mengandung jumlah gen yang melebihi jumlah sampel, sehingga diperlukan metode seleksi fitur untuk menentukan gen informatif [4]. Gen informatif yang dipilih, akan digunakan untuk melatih *classifier*. Kemudian *classifier* ini akan menggolongkan sampel data *microarray* kedalam kelasnya masing-masing berdasarkan model klasifikasi yang telah dibuat.

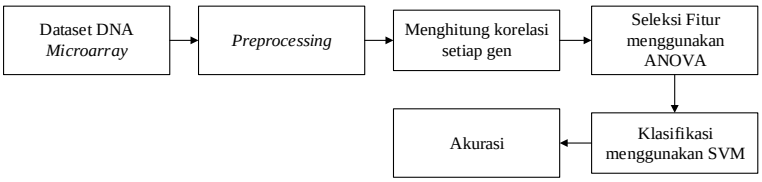
Dalam penelitian ini, Support Vector Machine (SVM) bersama tiga kernel dasar yaitu Linear, Polynomial dan Radial Basis Function (RBF) akan diterapkan untuk membagi *dataset microarray* menggunakan Analysis of Variance (ANOVA) sebagai metode seleksi fitur. ANOVA dipiilh sebagai metode seleksi fitur karena menjadi pasangan terbaik bersama *classifier* SVM dibandingkan dengan pasangan lainnya [1]. Juga dengan memainkan kernel pada SVM, dapat diketahui kernel mana yang menghasilkan akurasi paling besar untuk klasifikasi. Berdasarkan kajian yang telah dilakukan oleh Bharathi dan Natarajam, akurasi terbaik yang dihasilkan SVM dan ANOVA yaitu sebesar 97,91%, lebih besar bila dibandingkan dengan T-Test dan BPN [1], dan berdasarkan kajian yang dilakukan oleh Devi Arockia Vanitha dan tim, akurasi terbesar dihasilkan oleh SVM dengan kernel linear yaitu sebesar 97,7% [3]

II. Studi Literatur

Berbagai penelitian telah dilakukan oleh para ahli untuk menangani masalah dimensi tinggi yang dimiliki oleh data *microarray* juga metode yang tepat untuk pengklasifikasian data tersebut. Berikut beberapa rujukan yang melakukan pengujian dengan menggunakan data *microarray*. Mukesh Kumar et. All menggunakan metode seleksi fitur T-test dan Functional Link Neural Network (FLNN) sebagai *classifier* dengan tingkat akurasi yang dihasilkan sebesar 97,22% [1]. Bharathi and Natarajan menggunakan seleksi fitur Analysis of Variance (ANOVA) dan Support Vector Machine (SVM) sebagai *classifier* dengan tingkat akurasi 97,91% [3]. Selanjutnya Diaz et. All menggunakan metode Random Forest dan menghasilkan akurasi sebesar 95% [6], serta Devi Arockia et. All menggunakan Mutual Information (MI) sebagai seleksi fitur dan SVM sebagai *classifier* dengan tingkat akurasi yang dihasilkan yaitu sebesar 97,77% [2]. Berdasarkan rujukan tersebut, dapat terlihat bahwa ANOVA dan SVM merupakan metode dengan tingkat akurasi tertinggi, sehingga diharapkan akurasi yang dihasilkan pada pengujian ini akan lebih baik dengan memainkan beberapa kernel dan nilai parameter pada SVM.

III. Metodologi Penelitian

Gambaran umum sistem yang akan dibuat dalam penelitian ini dapat dilihat pada diagram blok di bawah ini.



Gambar. 1. Diagram Blok Sistem

Berdasarkan Gambar. 1, sistem yang akan dibuat dalam penelitian yaitu keempat *dataset* yang dimiliki masuk kedalam tahap *preprocessing* terlebih dahulu. *Preprocessing* ini akan membantu dalam penanganan *missing value* pada *dataset*. Kemudian dilakukan perhitungan korelasi antar gen agar masukan pada ANOVA tidak terlalu banyak. Setelah itu melakukan seleksi fitur menggunakan ANOVA. Pasangan gen yang dihasilkan oleh ANOVA kemudian diklasifikasikan oleh SVM kedalam kelas masing-masing berdasarkan model yang telah dihasilkan oleh *classifier* tersebut.

A. *Dataset* DNA *Microarray*

Dataset yang digunakan dalam penelitian ini terdiri dari empat *dataset* DNA *microarray* yaitu leukimia, *colon tumor*, *lung cancer* dan *ovarian cancer* yang nantinya akan menjadi masukan. Dalam satu kali proses pengerjaan sampai pada tahap akhir seperti pada Gambar. 1, *dataset* yang digunakan hanya satu *dataset*. Sehingga diperlukan empat kali proses pengulangan untuk mendapatkan hasil akurasi dari empat *dataset* tersebut.

Tabel. 1. *Dataset Microarray*

Dataset	**Jumlah Gen**	**Jumlah Class**	**Jumlah Sample**		**Ukuran Data**
Leukimia	7129	2	47 ALL	25 AML	1.8 MB
Colon tumor	2000	2	40 Negatif	22 Positif	0.9 MB
Lung cancer	12533	2	31 Mesothelioma	150 ADCA	9 MB
Ovarian cancer	15154	2	91 Negatif	162 Positif	24.7 MB

B. *Preprocessing*

Data DNA *microarray* akan dinormalisasi. *Missing value* dari sebuah fitur *dataset* DNA *microarray* terhubung menggunakan nilai rata – rata dari masing – masing fitur [1]. Dengan melakukan normalisasi, hasil yang didapatkan akan lebih baik. Rentang nilai atribut yang dianjurkan untuk digunakan yaitu [0,1] dengan menggunakan teknik normalisasi Min – Max [7].

$$Normalized(x) = \frac{x - \min(X_i)}{\max(X_i) - \min(X_i)} \quad (1)$$

Dimana $\min(X_i)$ dan $\max(X_i)$ mewakili nilai minimum dan maksimum untuk *dataset* X_i masing – masing. Jika dalam situasi $\max(X_i)$ sama dengan $\min(X_i)$, maka $Normalized(x)$ akan diubahh menjadi 0,5.

C. Perhitungan Korelasi Setiap Gen

Menghitung korelasi pada setiap gen dapat membantu melihat seberapa besar hubungan antara dua variabel tersebut. Hal ini bermanfaat untuk proses seleksi fitur selanjutnya, sebab tanpa melakukan perhitungan korelasi pada setiap gen akan membuat *output* ANOVA menjadi jutaan pasangan gen yang berdampak pada efisiensi waktu perhitungan proses seleksi fitur. Berikut merupakan persamaan untuk mencari korelasi pearson.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} \quad (2)$$

Besarnya nilai korelasi yang dihasilkan berkisar antara nilai $-1 \le \mu \le 1$. Apabila nilai korelasi yang dihasilkan semakin mendekati -1 atau 1, berarti hubungan kedua variabel sangat kuat. Namun apabila nilai korelasi mendekati 0, maka hubungan kedua korelasi tersebut sangat lemah.

D. Seleksi Fitur menggunakan Analysis of Variance (ANOVA)

Seleksi fitur dilakukan untuk mereduksi dimensi yang ada pada *dataset*, selain itu untuk menemukan gen informatif dalam *dataset* dan mengetahui interaksi antar gen serta pengaruh terhadap suatu perilaku. Jenis ANOVA yang digunakan yaitu *two way* ANOVA.

Langkah pertama yang dilakukan dalam ANOVA yaitu penentuan hipotesis nol (H_0). Kemudian data harus memenuhi empat asumsi seperti pengecekan normalitas, homogenitas, pengecekan kelompok yang independen dan data yang digunakan harus bersifat aditif. Data yang sudah memenuhi asumsi akan melakukan proses perhitungan ANOVA.

Tabel. 2. Tabel Perhitungan ANOVA

Sumber Keragaman (SK)	Jumlah Kuadrat (JK)	Derajat Kebebasan (db)	Kuadrat Tengah (KT)	F Hitung
Kolom (K)	$SSA = \left(\sum_{k=1}^{K} \frac{T_k^{\,2}}{n_k} \right) - \frac{T^2}{N}$	$DFA = k - 1$	$MSA = \frac{SSA}{DFA}$	$Fhitung(kolom) = MSA / MSAxB$
Baris (B)	$SSB = \left(\sum_{b=1}^{B} \frac{T_b^{\,2}}{n_b} \right) - \frac{T^2}{N}$	$DFB = b - 1$	$MSB = \frac{SSB}{DFB}$	$Fhitung(baris) = MSB / MSAxB$
Galat (G)	$SSAxB = SST - (SSA + SSB)$	$DFAxB = (k-1)(b-1)$	$MSAxB = \frac{SSAxB}{DFAxB}$	
Total (T)	$SST = \left(\sum_{b=1}^{B} \sum_{k=1}^{K} X_{bk}^{\,2} \right) - \frac{T^2}{N}$	$DFT = N - 1$		

Keterangan:

SST	: jumlah kuadrat keseluruhan	n_b	: jumlah data dalam masing – masing baris
B	: baris	$T_b^{\,2}$	: kuadrat jumlah masing – masing baris
K	: kolom	DFA	: derajat bebas kolom
X_{bk}	: data dalam baris-b dan kolom-k	DFB	: derajat bebas baris
N	: jumlah data keseluruhan	$DFAxB$	: derajat bebas galat
T^2	: kuadrat jumlah keseluruhan	DFT	: derajat bebas keseluruhan
SSA	: jumlah kuadrat antar kolom	MSA	: Kuadrat rata – rata kolom
n_k	: jumlah data dalam masing – masing kolom	MSB	: Kuadrat rata – rata baris
$T_k^{\,2}$	: kuadrat jumlah masing – masing kolom	$MSAxB$	: Kuadrat rata – rata galat
SSB	: jumlah kuadrat antar baris		

Kesimpulan yang dapat diambil setelah mengetahui tabel ANOVA yaitu penerimaan atau penolakan hipotesis nol (H_0). Apabila hipotesis nol (H_0) ditolak, maka langkah selanjutnya yaitu melakukan uji lanjut pada data tersebut untuk menemukan pasangan gen informatif yang akan menjadi masukan pada proses klasifikasi. Uji Tukey's HSD merupakan uji lanjut yang akan dilakukan dalam penelitian ini.

Setelah mengetahui selisih rata – rata antar gen, maka dilakukan perbandingan $|\mu_i - \mu_j| > HSD_{(\alpha)}$. Dimana nilai $\mu_i - \mu_j$ merupakan selisih antar gen dan nilai $HSD_{(\alpha)}$ didapat dari persamaan:

$$HSD_{(\alpha)} = q_{\alpha(p,v)} \sqrt{\frac{MSAxB}{n}} \qquad (3)$$

Dapat disimpulkan apabila pasangan gen tersebut merupakan gen informatif, sedangkan apabila maka pasangan gen tersebut bukan pasangan gen informatif dan tidak menjadi masukan untuk proses klasifikasi.

E. Klasifikasi menggunakan Support Vector Machine (SVM)

Setelah mendapatkan pasangan gen informatif, pada tahap ini akan dibuat model *hyperplane* terbaik untuk memisahkan kedua kelas berdasarkan pasangan gen tersebut. Sebelumnya, data yang digunakan untuk membuat model adalah data *training* yang dibagi menggunakan metode cross validation. Jenis SVM yang digunakan yaitu *binary class*, karena kelas pada data hanya bernilai 1 atau -1 (kanker atau *non* kanker). Kemudian model yang dihasilkan akan diuji dengan menggunakan data *testing*.

1. *Learning* Model dengan Support Vector Machine (SVM)

Pada proses *learning* model inilah akan dicari *hyperplane* terbaik yang akan memisahkan data kedalam dua buah kelas yang berbeda. *Hyperplane* terbaik diperoleh dengan memaksimalkan *margin* pada *support vector*. Untuk mendapatkan nilai optimal maka dapat dihitung dengan meminimumkan L terhadap $\vec{w}$ dan b dan memaksimumkan L terhadap α_i.

Karena pada proses klasifikasi ini menggunakan *kernel trick*, sehingga perhitungan *dot product* $\vec{x}$ pada setiap persamaan akan berubah dengan menambahkan fungsi Φ. Hal tersebut terjadi karena proses transformasi dari *input space* kedalam *feature space*.

Langkah selanjutnya yaitu menentukan label dari data *microarray* dengan cara memasukkan data *input* dengan nilai $\vec{w}$ dan b yang telah dicari dengan menggunakan persamaan 2.29. Jika nilai $f(\vec{x})$ yang dihasilkan adalah $f(\vec{x}) > 0$, maka data tersebut akan terklasifikasi kedalam kelas positif (+1), sebaliknya, maka akan terklasifikasi kedalam kelas negatif (-1). Hasil yang didapatkan dari *learning* model ini yaitu berupa model persamaan *hyperplane* untuk setiap pasangan gen yang telah dipilih dan akurasi *training* untuk setiap pasangan gen.

2. Pengujian Model Klasifikasi

Pengujian model klasifikasi ini dilakukan untuk mengetahui akurasi yang dihasilkan oleh fungsi kernel linear, polynomial dan radial basis function (RBF) dengan masing – masing menggunakan keempat data *testing* yaitu leukimia, *colon tumor, lung cancer dan ovarian cancer*. Sebelum menjadi masukan data *testing* pada SVM, *dataset* tersebut akan dinormalisasi terlebih dahulu dan dicari nilai korelasi pada setiap gen. Dari ketiga kernel yang diuji, ada satu kernel yang menghasilkan akurasi terbaik.

IV. HASIL DAN PEMBAHASAN

Pada tahap ini akan dilakukan beberapa pengujian. Pengujian ini dilakukan untuk menganalisis seberapa besar pengaruh perhitungan korelasi yang dilakukan, pengaruh seleksi fitur menggunakan *Analysis of Variance* (ANOVA), pengaruh kernel – kernal yang digunakan untuk pengklasifikasian *dataset microarray*, serta pengaruh nilai – nilai parameter kernel SVM pada akurasi yang dihasilkan.

A. Skenario Pengujian

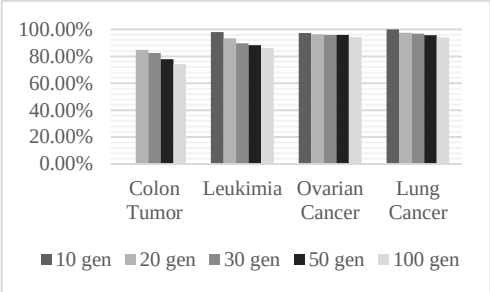
Pada pengujian ini akan dilakukan tiga skenario utama. Skenario pertama, akan dilakukan klasifikasi untuk setiap *dataset microarray* menggunakan SVM, tanpa menghitung korelasi dan seleksi fitur menggunakan ANOVA. Pada skenario ini, nilai parameter C, d dan σ akan diubah – ubah nilainya dan berlaku pada dua skenario lainnya. Tujuan dari skenario ini yaitu mencari akurasi tertinggi dari pengklasifikasian menggunakan SVM tanpa melakukan seleksi fitur terhadap *dataset*.

Skenario kedua, akan dilakukan pengklasifikasian *dataset microarray* dengan menghitung korelasi pada setiap gen terlebih dahulu. Hasil korelasi pada setiap gen akan diurutkan berdasarkan korelasi tertinggi ke korelasi terendah. Kemudian akan diambil sebanyak 10, 20, 30, 50 dan 100 gen dengan nilai korelasi tertinggi dan dijadikan sebagai masukan kedalam SVM. Tujuan dari skenario ini adalah mencari akurasi tertinggi yang dihasilkan berdasarkan perhitungan korelasi yang telah dirangking.

Skenario terakhir, akan dilakukan beberapa proses sebelum masuk kedalam klasifikasi, diantaranya menghitung nilai korelasi dengan mengambil 10, 20, 30, 50 dan 100 gen berdasarkan nilai korelasi tertinggi, melakukan seleksi fitur menggunakan ANOVA dan terakhir yaitu mengklasifikasikan pasangan gen pada setiap *dataset microarray* berdasarkan *keluaran* yang dihasilkan oleh ANOVA kedalam SVM. Tujuan dari pengujian ini yaitu mengetahui seberapa besar peranan korelasi dan ANOVA dalam melakukan seleksi fitur sehingga berpengaruh pada akurasi yang dihasilkan pada saat pengklasifikasian menggunakan SVM.

B. Pengaruh Jumlah Atribut Hasil Korelasi terhadap Akurasi

Perhitungan korelasi pada *dataset* sebelum melakukan seleksi fitur sangat berpengaruh pada akurasi yang dihasilkan SVM. Semakin tinggi nilai korelasi yang dihasilkan dari setiap gen, semakin kuat hubungan antara gen dan *class*nya (status kanker). Sehingga 10 gen pertama dengan nilai korelasi tertinggi akan selalu menghasilkan akurasi terbaik bila dibandingkan dengan 20, 30, 50 dan 100 gen yang telah dirangking lainnya.



Gambar. 2 Grafik Perbandingan Akurasi berdasarkan Rangking Korelasi

Grafik pada Gambar. 2 menunjukkan bahwa semakin banyak gen yang menjadi masukan belum tentu menghasilkan akurasi yang baik. Seperti pada *dataset* leukimia. Akurasi yang dihasilkan 10 gen dan 100 gen memiliki perbedaan yang cukup besar yaitu 11,75%. Semakin kecil korelasi yang dihasilkan semakin kecil akurasi yang didapatkan. Sehingga apabila jumlah gen yang berkorelasi lemah jumlahnya lebih banyak dibandingkan dengan gen yang berkorelasi kuat, akurasi yang dihasilkan akan tetap kecil.

C. Pengaruh Jumlah Atribut Hasil Korelasi terhadap Pasangan Gen yang Dihasilkan ANOVA

Seleksi gen menggunakan ANOVA sangat berpengaruh pada jumlah gen yang akan menjadi masukan pada proses klasifikasi. ANOVA memasangkan gen – gen yang dipilih berdasarkan perhitungan korelasi yang sebelumnya telah dirangking. Pemilihan pasangan gen yang informatif, dipilih melalui perhitungan ANOVA. Gen yang tidak informatif tidak akan memiliki pasangan dan tidak akan menjadi masukan dalam SVM.

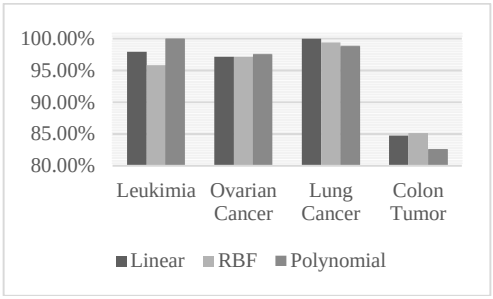
Tabel. 3. Perbandingan Jumlah Atribut dan *Running Time* menggunakan Korelasi

Dataset	**Jumlah Atribut**	**Jumlah Pasangan (pasang)**			***Running Time* (s)**		
		Tanpa Korelasi	**Dengan Korelasi**		**Tanpa Korelasi**	**Dengan Korelasi**	
Colon tumor	2000	338444	10 gen	-	2223,986	10 gen	0,719
			20 gen	17		20 gen	0,967
			30 gen	40		30 gen	1,247
			50 gen	123		50 gen	2,246
			100 gen	707		100 gen	6,333

Berdasarkan Tabel. 3 terlihat sangat jelas bahwa jumlah pasangan gen yang dihasilkan oleh ANOVA tanpa melakukan perhitungan korelasi sangat banyak. Walaupun itu semua merupakan pasangan gen yang informatif, namun belum tentu menghasilkan akurasi yang bagus apabila dibandingkan dengan pasangan gen yang telah di rangking berdasarkan nilai korelasi tertinggi. Juga waktu yang diperlukan untuk *learning model* pada SVM akan lama. Sehingga perhitungan korelasi pada ANOVA ini sangat membantu dalam proses seleksi gen.

D. Pengaruh Kernel yang Digunakan dalam SVM

Pada proses *learning* model menggunakan SVM, data pada *input space* ditransformasi kedalam *feature space* dengan menggunakan kernel *trick*. Kernel – kernel yang digunakan yaitu linear, polynomial dan RBF. Ketiga kernel tersebut memegang peranan penting dalam proses pengklasifikasian keempat *dataset* kanker. Dari ketiga kernel, akan ada satu kernel terbaik yang memisahkan kedua buah class pada masing – masing *dataset*. Berikut merupakan grafik yang dihasilkan dari setiap *dataset* dengan menggunakan kernel linear, polynomial dan RBF.



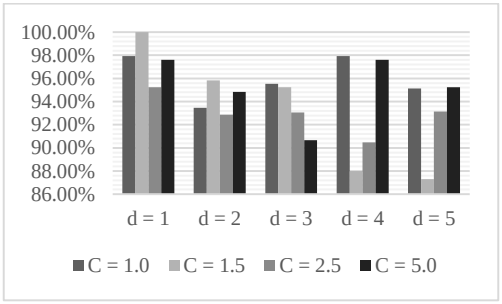
Gambar. 3 Grafik Perbandingan Akurasi berdasarkan *Kernel Trick* yang digunakan pada SVM

Akurasi yang ditampilkan pada grafik di Gambar. 3 merupakan akurasi terbesar dari masing – masing kernel yang dihasilkan oleh setiap *dataset* kanker. Berdasarkan Gambar. 3 terlihat bahwa masing – masing *dataset* memiliki akurasi terbaik yang dihasilkan oleh kernel yang berbeda.

Untuk *dataset lung cancer* akurasi terbesar diperoleh dari kernel linear yaitu sebesar 100% hal tersebut disebabkan karena *dataset lung cancer* dapat dipisahkan secara linear sehingga kernel linear menjadi akurasi tertinggi dibandingkan dengan kernel lainnya. Untuk *dataset* leukimia dan *ovarian cancer*, akurasi terbesar dihasilkan oleh kernel polynomial yaitu sebesar 100% dan 97,54%. Sedangkan untuk *dataset colon tumor* akurasi terbesar diperoleh dari kernel RBF sebesar 85,15%. Akurasi terbesar pada setiap *dataset* dengan kernel yang berbeda diakibatkan karakteristik *dataset* yang berbeda – beda.

E. Pengaruh Parameter d (degree) pada Kernel Polynomial

Pada setiap *dataset* kanker yang diuji, nilai parameter d pada kernel polynomial akan menemukan nilai optimal di setiap *dataset*nya. Berikut merupakan grafik perbandingan nilai parameter d pada *dataset* leukimia berdasarkan akurasi yang dihasilkan.



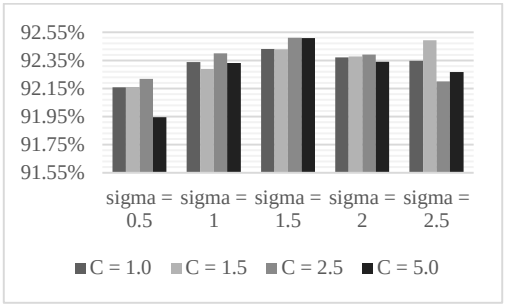
Gambar. 4 Grafik Perbandingan Nilai Parameter d (degree) pada *Dataset* Leukimia

Grafik pada Gambar. 4 menunjukkan bahwa semakin besar nilai parameter d yang digunakan pada kernel polynomial belum tentu menghasilkan akurasi terbaik dalam sebuah *dataset* seperti pada *dataset* leukimia ini. Nilai parameter d pada *dataset* leukimia sudah optimal pada $d=1$. Sehingga dapat dilihat bahwa akurasi terbesar yang dihasilkan yaitu 100% pada nilai parameter C = 1 dan $d=1$. Untuk ketiga *dataset* lainnya, kemungkinan nilai d yang optimal bisa berubah – ubah sesuai dengan karakteristik *dataset* itu sendiri.

Ketika sudah menemukan nilai d yang optimal, sebaiknya cukup sampai disana saja pengujiannya sebab semakin besar nilai parameter d yang digunakan semakin lama *running time* yang diperlukan, juga belum tentu akan menghasilkan akurasi yang lebih baik lagi.

F. Pengaruh Parameter σ (sigma) pada Kernel Radial Basis Function

Seperti halnya parameter d pada kernel polynomial, parameter σ pada kernel RBF pun akan menemukan nilai optimalnya dalam setiap *dataset*. Berikut di bawah ini merupakan grafik perbandingan nilai parameter σ pada *dataset Lung cancer*.

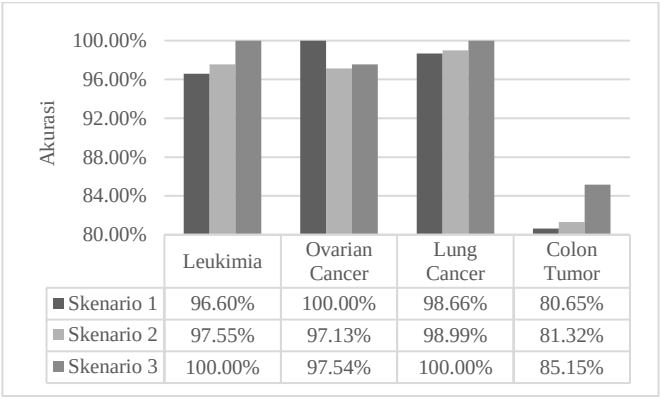


Gambar. 5 Grafik Perbandingan Nilai Parameter σ (sigma) pada *Dataset Lung Cancer*

Nilai σ pada kernel RBF akan optimal di satu nilai parameter seperti yang terlihat di grafik pada Gambar. 5. *Dataset lung cancer* memiliki akurasi tertinggi dengan nilai parameter $\sigma = 1.5$ yaitu sebesar 92,51%. Apabila nilai optimal dari parameter σ sudah ditemukan, akurasi terbesar akan dihasilkan dan ketika nilai parameter σ ditambah maka akurasi yang dihasilkan tidak stabil. Sebagian besar mengalami penurunan akurasi yang signifikan.

G. Skenario Terbaik berdasarkan Tiga Skenario Utama

Setelah melakukan pengujian berdasarkan tiga skenario utama, akan dilihat mana skenario yang menghasilkan akurasi terbaik untuk keempat *dataset*. Berikut merupakan grafik perbandingan yang dihasilkan oleh setiap *dataset* pada tiga skenario utama.



	Leukimia	Ovarian Cancer	Lung Cancer	Colon Tumor
Skenario 1	96.60%	100.00%	98.66%	80.65%
Skenario 2	97.55%	97.13%	98.99%	81.32%
Skenario 3	100.00%	97.54%	100.00%	85.15%

Gambar. 6 Grafik Perbandingan Nilai Parameter σ (sigma) pada *Dataset Lung Cancer*

Bila dilihat pada grafik yang ada pada Gambar. 6 akurasi yang dihasilkan untuk ketiga skenario berbeda – beda untuk setiap *dataset*nya. Namun berdasarkan akurasi yang dihasilkan, skenario tiga merupakan skenario terbaik untuk tiga *dataset* yaitu *colon tumor*, leukimia dan *lung cancer*. Hal ini dapat dilihat dari hasil akurasi yang semakin meningkat disetiap skenarionya. Kenaikan untuk *dataset* leukimia dari skenario satu ke skenario dua yaitu sebesar 0.95%, sedangkan kenaikan pada skenario satu ke skenario tiga sebersar 3.4% dan kenaikan pada skenario dua ke skenario tiga sebesar 2.45%. Untuk dataset *colon tumor* kenaikan pada skenario satu ke skenario dua yaitu sebesar 0.67%, sedangkan kenaikan pada skenario satu ke skenario tiga sebersar 4.5% dan kenaikan pada skenario dua ke skenario tiga sebesar 3.83%, serta untuk dataset *lung cancer* kenaikan dari skenario satu ke skenario dua yaitu sebesar 0.33%, sedangkan kenaikan pada skenario satu ke skenario tiga sebersar 1.01% dan kenaikan pada skenario dua ke skenario tiga sebesar 1.34%. Sehingga dengan adanya rincian kenaikan akurasi yang dihasilkan pada setiap skenario, terbukti bahwa skenario tiga menjadi skenario terbaik untuk ketiga *dataset.* Hal itu di karenakan adanya proses seleksi fitur sebelum pengklasifikasian yang sangat membantu dapat pemilihan gen yang tepat untuk masukan kedalam SVM, sehingga gen yang tidak berpengaruh pada status pasien (kanker / *non* kanker) akan dieleminasi.

V. KESIMPULAN

Berdasarkan pengujian yang telah dilakukan, didapatkan beberapa kesimpulan.

1. Akurasi yang dihasilkan skenario dua lebih besar dibandingkan skenario satu, seperti pada *dataset* leukimia, *lung cancer* dan *colon tumor*. Akurasi skenario dua meningkat masing – masing sebesar 0.95%, 0,33%, 0.67%. Peningkatan akurasi tersebut membuktikan bahwa pentingnya melakukan perhitungan korelasi untuk mengetahui seberapa kuat hubungan antara atribut (gen) dengan *class*nya (kanker atau *non* kanker).
2. Peningkatan akurasi yang dihasilkan oleh skenario tiga pada *dataset* leukimia, *lung cancer* dan *colon tumor* sebesar 2.45%, 1.01% dan 3.83%. Akurasi tersebut meningkat karena setelah melakukan perhitungan korelasi dilakukan proses seleksi fitur oleh ANOVA, sehingga pasangan gen informatif yang dihasilkan dari ANOVA menjadi pasangan terbaik untuk masukan kedalam SVM dan menghasilkan akurasi terbaik.
3. Karakteristik *dataset* kanker yang digunakan di dalam pengujian mempengaruhi akurasi setiap kernel yang diujikan. Sehingga tidak semua *dataset* akan memiliki akurasi tertinggi dari satu kernel yang sama.
4. Nilai parameter C, d dan σ yang diubah – ubah berdampak signifikan terhadap akurasi yang dihasilkan seperti pada *dataset* leukimia dan *lung cancer* yang mendapatkan akurasi sebesar 100% pada skenario tiga.
5. Akurasi yang dihasilkan data *colon tumor* merupakan akurasi terkecil jika dibandingkan dengan *dataset* lainnya, sehingga diperlukan penelitian kembali dengan menggunakan metode seleksi fitur atau metode klasifikasi yang berbeda untuk mendapatkan akurasi yang lebih baik. Kemudian penelitian dengan metode yang berbeda dapat dilakukan dengan menambah *dataset* kanker lain, selain empat *dataset* yang telah digunakan pada penelitian ini seperti *dataset breast cancer* dan *dataset* lainnya.

REFERENCES

[1] Kumar, Mukesh, Sandeep Singh, and Santanu Kumar Rath. "Classification of *Microarray* Data using Functional Link Neural Network." *Procedia Computer Science* 57 (2015): 727-737.

[2] Vanitha, C. Devi Arockia, D. Devaraj, and M. Venkatesulu. "Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection." *Procedia Computer Science* 47 (2015): 13-21.

[3] Bharathi, A., and A. M. Natarajan. "Cancer Classification of Bioinformatics data using ANOVA." *International journal of computer theory and engineering* 2.3 (2010): 369.

[4] Singh, Rabindra Kumar, and M. Sivabalakrishnan. "Feature Selection of Gene Expression Data for Cancer Classification: A Review." *Procedia Computer Science* 50 (2015): 52-57.

[5] Pusat Data dan Informasi. 2015. Situasi Penyakit Kanker. Kementrian Kesehatan RI.

[6] Díaz-Uriarte, Ramón, and Sara Alvarez De Andres. "Gene selection and classification of *microarray* data using random forest." *BMC bioinformatics*7.1 (2006): 1.

[7] Jain, Yogendra Kumar, and Santosh Kumar Bhandare. "Min max normalization based data perturbation method for privacy protection." *International Journal of Computer & Communication Technology (IJCCT)* 2.8 (2011): 45-50.