Prediksi Penyakit Menggunakan *Algoritma Differential Evolution* (DE) dan *Least Square Support Vector Machine* (LSSVM) Untuk Data Berdimensi Tinggi

Prediction Of Disease Using Differential Evolution (DE) and Least Square Support Vector Mchine (LSSVM) For High Dimensional Data

[1]Merry Sofhia Tambunan, [2]Fhira Nhita, [3]Danang Triantoro
Ilmu Komputasi Fakultas Informatika Universitas Telkom, Bandung
Ilmu Komputasi Fakultas Informatika Universitas Telkom, Bandung
Ilmu Komputasi Fakultas Informatika Universitas Telkom, Bandung
[1]merrysofhiatambunan@gmail.com, [2]fhiranhita@telkomuniversity.ac.id, [3]indwiarti@telkomuniversity.ac.id

ABSTRAK

Penyakit merupakan salah satu penyebab kematian tertinggi bagi masyarakat. Beberapa penyakit dapat dikategorikan sebagai penyakit mematikan. Penyakit *colon tumor* atau tumor usus dan leukimia merupakan beberapa penyakit yang berbahaya dan mematikan. Masyarakat terkadang tidak menyadari bahwa sedang mengidap beberapa penyakit berbahaya ini. Berdasarkan permasalahan tersebut diperlukan adanya suatu sistem prediksi penyakit *colon tumor* dan leukimia. Pada tugas akhir ini digunakan algoritma *Differential Evolution* (DE) dan *Least Square Support Vector Machine* (LSSVM) dalam memprediksi penyakit *colon tumor* dan leukemia. Data yang digunakan pada tugas akhir ini merupakan data penyakit dimensi tinggi, dimana akan dilakukan preprocessing data menggunakan PCA sehingga diperoleh data dengan dimensi baru yang lebih sedikit. Data yang sudah direduksi kemudian akan dimasukan ke dalam algoritma DE, dimana algoritma tersebut akan melakukan serangkaian proses evolusi. DE bertujuan untuk mencari parameter LSSVM yang optimal. Parameter tersebut kemudian akan digunakan pada metode klasifikasi LSSVM. Proses ini dilakukan untuk mendapatkan klasifikasi dari penyakit colon tumor dan leukimia. Dari hasil pengujian pada algoritma DE dan LSSVM diperoleh solusi optimal dengan akurasi 90.4762% untuk *colon tumor* dan 87.5 % untuk leukemia.

Kata kunci: data dimensi tinggi, PCA, *Differential Evolution* (DE), *Least Square Support Vector Machine* (LSSVM).

ABSTRACT

Disease is one of the highest death cause for society . Some diseases can be categorized as deadly disease. Colon tumor and leukemia are few examples of dangerous and deadly diseases. People sometimes don't reliaze that they're infected by these dangerous diseases. Based on this problem, it's a necessary to have a colon tumor and leukemia prediction system. In this final project, it uses differential evolution (DE) and Least Square Support Vector Machine (LSSVM) algorithm to predict colon tumor and leukemia diseases. The data that will be used in this final project is a high dimensional data of diseases, where it will be preprocessed with PCA so that it produce new data with smaller dimensions. The reducted data will be inserted into DE algorithm , where the algorithm will do series of evolution processes. DE intended to find optimal LSSVM parameters. These paramaters then will be used on LSSVM classification method. This process is done to produce classification for colon tumor and leukemia. From the testing results on DE and LSSVM algorithm it obtain the optimal solution that 90.4762% accuracy colon tumor and 87.5 % for leukemia

***Keywords**: High dimensional data, PCA, Differential Evolution* (DE), *Least Square Support Vector Machine* (LSSVM).

1. Pendahuluan

1.1 Latar Belakang

Data mining merupakan suatu proses penggalian informasi yang didapatkan dari database. Pada data mining terdapat beberapa tantangan yang dihadapi,salah satunya adalah permasalahan data dimensi tinggi. Data dimensi tinggi merupakan suatu data yang terdiri dari ratusan bahkan ribuan dimensi. Meningkatnya jumlah dimensi data akan mengakibatkan kurangnya performansi pada proses data mining. Permasalahan data dimensi tinggi ini disebut "*Curse of Dimensionality* "[5]. Salah satu contoh data dengan dimensi tinggi adalah data penyakit.

Dalam menyelesaikan permasalahan data dimensi tinggi, diperlukan adanya reduksi dimensi yang berguna untuk mendapatkan keakurasian data yang lebih baik. Reduksi dimensi dilakukan untuk mengurangi atribut-

atribut dari suatu data yang tidak diperlukan. Dalam permasalahan ini diperlukan suatu metode yang dapat membantu dalam melakukan reduksi dimensi.

Principal Component Analysis (PCA) merupakan salah satu teknik pengurangan dimensi yang bertujuan untuk mengurangi dimensi pada data *unsupervised* [13]. Pada Tugas akhir ini PCA digunakan sebagai teknik *preprocessing* data,dimana hasil preprocessing ini akan digunakan pada algoritma *Differential Evolution*.

DE (Differntial Evolution) merupakan salah satu algoritma yang termasuk dalam *EAs (Evolutionary Algorithms)* yang berguna sebagai salah satu teknik optimasi efektif yang biasanya digunakan pada permasalahan ilmiah maupun rekayasa. Pada tugas akhir ini algoritma DE digunakan untuk mendapatkan parameter yang akan digunakan pada *classifier* LS-SVM.

LS-SVM (Least Square Support Vector Machine), yang merupakan suatu tools pengklasifikasian pada suatu sampel tertentu dari *SVM (Support Vector Machine.* Kekurangan dari *LS-SVM* adalah sensitif pada perubahan nilai parameter. Algoritma *DE* digunakan untuk pengoptimasian parameter pada *LS-SVM* [1]. Penelitian dengan menggunakan DE dan LS-SVM sudah pernah dilakukan sebelumnya oleh Omar S.Sk oliman dan Eman AboElHamd dalam sebuah jurnal internasional berjudul *"Classification of Breast Cancer using Differntial Evolution and LeastSquares Support Vector Machine"* dengan tingkat akurasi sebesar 99,75% [1].

1.2 Perumusan Masalah

Berdasarkan latar belakang diatas, maka terdapat beberapa permasalahan yang akan diselesaikan dalam tugas akhir ini, permasalahan tersebut terdiri dari:

1. Bagaimana implementasi Algoritma Differential Evolution dalam mengoptimasi *Least Squares Support Vector Machine* (LS-SVM) pada data penyakit berdimensi tinggi?
2. Bagaimana cara kerja Algoritma *Differential Evolution* untuk menemukan parameter LS-SVM pada data penyakit berdimensi tinggi?
3. Bagaimana performansi yang didapatkan *Least Squares Support Vector Machine* (LS-SVM) pada data penyakit berdimensi tinggi ?

Adapun batasan masalah dari tugas akhir ini adalah sebagai berikut:

1. Data yang digunakan merupakan data penyakit leukimia dan colon tumor yang diambil dari *Biomedical Dataset* pada *Kent Ridge*.
2. Tidak dilakukan penanganan *outlier* pada dataset yang digunakan.
3. Dataset sudah memiliki label class.
4. Data yang digunakan merupakan data numerik.

1.3 Tujuan

Berdasarkan perumusan masalah diatas tujuan untuk menyelesaikan masalah tersebut adalah:

1. Mengimplementasikan Algoritma Differential Evolution dalam mengoptimasi *Least Squares Support Vector Machine (LS-SVM)* pada data penyakit berdimensi tinggi.
2. Mengetahui cara kerja Algoritma *Differential Evolution* untuk menemukan parameter LS-SVM pada data penyakit berdimensi tinggi.
3. Menganalisis performansi yang didapatkan *Least Squares Support Vector Machine (LS-SVM)* pada data penyakit berdimensi tinggi.

2. Landasan Teori

2.1 Data Berdimensi Tinggi

Pada data mining terdapat permasalahan-permasalahan yang dihadapi, data dimensi tinggi adalah salah satu pemasalahan yang terjadi pada data mining. Data berdimensi tinggi merupakan suatu data yang terditri dari ratusan maupun ribuan atribut. Data berdimensi tinggi menjadi masalah dalam proses data mining, karena banyaknya atribut yang ada. Padahal atribut-atribut yang terdapat pada data tersebut bisa saja merupakan atribut-atribut yang bisa dihilangkan atau tidak digunakan. Selain itu data dengan dimensi yang tinggi dapat mengurangi tingkat performansi dari suatu data karena banyaknya atribut. Permasalahan yang terdapat pada data dimensi tinggi disebut *"Curse Of Dimentionality"* [5].

2.2 *Principal Component Analysis (PCA)*

Dalam melakukan *preprocessing* data pada tugas akhir ini terdapat tiga jenis data penyakit berdimensi tinggi, dimana dalam pengolahannya akan membutuhkan waktu komputasi yang sangat lama. Karena itu dibutuhkan suatu teknik yang dapat menangani permasalahan tersebut. PCA merupakan suatu teknik yang dapat digunakan untuk mengekstrasi struktur dari suatu data yang berdimensi tinggi tanpa menghilangkan informasi signifikan pada keseluruhan data.

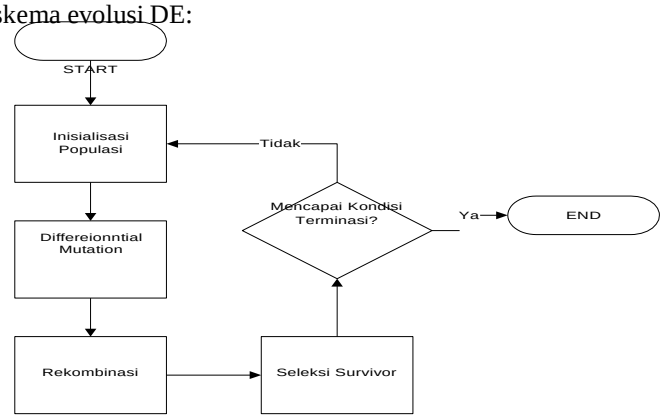
PCA merupakan teknik *multivariate* yaitu suatu teknik mentransformasikan variabel-variabel asal yang saling berkorelasi menjadi variabel baru yang tidak saling berkorelasi dengan cara mereduksi sejumlah variabel tersebut sehingga memilki dimensi yang lebih kecil namun dapat menerangkan sebagian besar variabel aslinya.

PCA akan menghasilkan dimensi baru yang disebut dengan *Principal Component* (PC).Dimana PC merupakan kombinansi linear dari dimensi asli pada data asli [14]. Selain itu semua PC saling ortogonal satu sama lain sehingga tidak ada informasi yang berulang [14].

2.3 *Differential Evolution (DE)*

EAs terdiri dari beberapa algoritma, salah satunya adalah *Diffential Evolution* (DE). *Differential Evoluton* (DE) adalah suatu metode optimasi dengan pendekatan heuristic untuk mencari nilai minimum dari fungsi ruang kontinyu yang nonlinier dan nondifferentiable [3]. DE dapat digunakan untuk menemukan minimum global dari fungsi multidimensional dan multimodal (fungsi yang terdiri dari satu nilai minimum) dengan probabilitas yang tinggi [3]. DE berbeda dengan metode optimasi lainnya, DE menggunakan *differential mutation* yang adalah mutasi semi terarah (*semi-directed*) atau bisa disebut operasi pra-seleksi khusus. Pada DE individu baru didapatkan dengan menggunakan perhitungan tertentu berbasis pada perbedaan jarak vektor antar individu orang tua [4]. Individu-individu yang berisi nilai real dianggap sebagai vektor. Dalam memilih orang tua DE tidak memperhatikan nilai *fitness*nya.

Berikut ini merupakan skema evolusi DE:



Gambar 1 : Skema Proses Evolusi Pada DE

Berikut ini merupakan penjelasan dari skema proses evolusi pada *Differential Evolution* (DE) pada gambar 2.3:

2.3.1 Representasi Individu

Pada DE representasi individu menggunakan bilangan *real*. Karena itu konsep dari DE adalah suatu individu yang berisi nilai *real* dapat dipandang sebagai suatu vektor, berdasarkan konsep tersebut maka untuk mencari perbedaan antara dua individu bias didapatkan dengan penrhitungan jarak antara dua vektor. Apabila suatu individu sudah terbentuk, individu-individu tersebut kemudian dikumpulkan menjadi suatu populasi. Berikut merupakan ilustrasi dari individu dalam suatu populasi:

2.3.2 Seleksi Orang Tua

Seleksi orang tua dilakukan dengan probabilitas yang sama untuk setiap individu tanpa memperhatikan nilai fitnessnya [3].

2.3.3 Differential Mutation

Differential mutation adalah sebuah proses pembangkitan vektor individu baru yang melibatkan individu orang tua [3]. Dalam membangkitkan vektor individu baru digunakan beberapa skema, diantaranya skema DE1 dan skema DE2.

2.3.4 Rekombinasi

Dalam meningkatkan keberagaman vektor-vektor parameter, vektor $\underline{v}$ direkombinasi dengan vektor sembarang dalam populasi, misal $\underline{Xi},G$. Proses ini menghasilkan vektor U berikut ini

$$U = (u1,u2,\ldots,uD)^{T} \quad \textbf{(2.1)}$$

untuk

$$u_{j,i,G+1} = \begin{cases} v_{j,i,G+1}\ \text{ jika } (rand_j[0,1) \le Cr)\ \text{atau}\ (j = j_{rand}),\ j = 1,2,..D \\ (x_{j,i,G}),\ \text{selainnya} \end{cases} \quad (2.2)$$

Keterangan :

u = individu hasil rekombinasi
D = jumlah individu pada suatu populasi
Cr = probabilitas crossover [0,1)
i =indeks individu dengan interval [1,D]
j = indeks gen dengan interval [1, D]
j_{rand} =bilangan bulat yang dipilih secara acak

berdasarkan rumus rekombinasi diatas suatu individu v akan direkombinasi (crossover) dengan suatu vektor sembarang dalam populasi $x_{i,G}$.

2.3.5 Seleksi Survivor

Metode ini berguna untuk mengetahui apakah suatu vektor adalah anggota generasi dari g+1, karena itu vektor tersebut diuji dengan $u_{i,G+1}$ yang dibandingkan dengan vektor target $v_{i,G+1}$. Apabila vektor u menghasilkan nilai yang lebih baik maka daripada $v_{i,G+1}$ maka rekombinasi u akan menggantikan $v_{i,G+1}$, tetapi jika hasilnya lebih buruk maka u tidak perlu menggantikan $v_{i,G+1}$. Jika nilai-nilai beberapa parameter dari percobaan baru yang dihasilkan vektor melebihi batas atas dan bawah yang sesuai, kita acak dan inisialisasi secara seragam dalam kisaran tertentu [9].

2.4 *Least Squares Support Vector Machine (LS-SVM)*

Least Squares Support Vector Machine (LS-SVM) adalah pengembangan dari metode SVM, tetapi kinerjanya lebih baik bila dibandingkan dengan SVM. *Least Squares Support Vector Machines* (LS-SVM) adalah formulasi ulang terhadap SVM standar yang mengarah untuk memecahkan sistem linear [7,8]. Perbedaan antara LS-SVM dan SVM adalah, LS-SVM menggunakan satu set persamaan linear untuk pelatihan, sementara SVM menggunakan masalah optimasi kuadrat [10]. LS-SVM dapat memproses data dalam jumlah besar tanpa harus menggunakan banyak memori maupun prosesor [12]. LS-SVM di formulasikan dengan sebuah fungsi constrain yang berupa persamaan. *Least Squares Support Vector Machines* (LS-SVM) adalah metode yang telah terbukti untuk klasifikasi dan fungsi pendekatan. Dibandingkan dengan Support Vector Machines standar (SVM) hanya membutuhkan memecahkan sistem linear [6]. Berikut ini merupakan persamaan dari LS-SVM [1] :

$$\frac{1}{2}w^2 + \frac{1}{2}C^2\sum_{i=1}^{n} e_i^2 \quad (2.3)$$

Dengan kendala,

$$y_i(w^T\varphi(x_i) + b) + e_i = 1,\ i = 1, \dots, n \quad (2.4)$$

Dimana untuk menemukan parameter w dan b perlu diubah menjadi optimasi tanpa pembatas [11]. Perubahan itu dilakukan dengan mengubah fungsi *Langrange* seperti dibawah ini

$$L(w,b,e,\alpha) = \frac{1}{2}||w||^2 + \frac{1}{2}C\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n}\alpha_i[y_i(w^T\varphi(x_i) + b) + e_i - 1] \quad (2.5)$$

Dalam hal ini α_i merupakan *Lagrange* multiplier, yang nilainya bias positif ataupun negatif.hal ini dikarenakan pembatas pada LS-SVM yang berupa persamaan. Sedangkan untuk kondisi optimalitas, maka disederhanakan menjadi :

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{n}\alpha_i y_i \varphi(x_i) \quad (2.6)$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \quad (2.7)$$

$$\frac{\partial}{\partial x} = 0 \rightarrow u = C \qquad (2)$$

$$\frac{\partial}{\partial} = 0 \rightarrow [\quad + \quad - 1 + \quad \qquad (\quad)$$

Pada LS-SVM masalah optimasi menjadi lebih sederhana, karena tidak perlu menyelesaikan *quadratic programing* seperti yang dilakukan SVM, tetapi hanya menggunakan solusi persamaan linier.

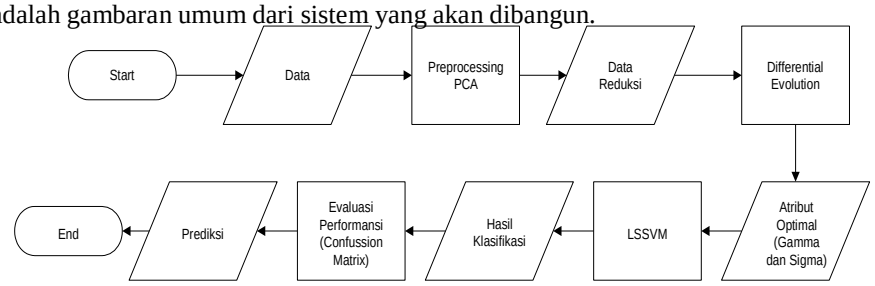
LS-SVM menggunakan fungsi kernel yang memungkinkan terjadinya operasi pada data berdimensi tinggi, dimana pada Tugas Akhir ini menggunakan jenis kernel RBF untuk mendapatkan hasil yang optimal [13].

3. Peranangan Sistem

Pada tugas akhir ini akan dirancang sistem untuk mengoptimalkan pemilihan atribut yang ada pada data penyakit dimensi tinggi.Dimana dalam pengklasifikasiannya digunakan *LS-SVM*. Sedangkan untuk untuk mengoptimasi pemilihan atribut pada data dimensi tinggi digunakan algoritma *Differential Evolution* (DE).

Data penyakit berdimensi tinggi yang digunakan merupakan *Biomedical Dataset* yang didapatkan dari Kent Ridge.Data penyakit berdimensi tinggi yang didapatkan selanjutnya akan dilakukan pengoptimasian dengan algoritma DE selanjutnya akan diklasifikasi dengan menggunakan LS-SVM.

Berikut adalah gambaran umum dari sistem yang akan dibangun.



Gambar 2 : Perancangan Sistem

3.1 Dataset

Data yang digunakan merupakan data penyakit berdimensi tinggi yang diperoleh dari *Kent Ridge Biomedical Data Set Repository*. Data ini terdiri dari *gene expression data*,*protein profiling data*,dan *genomic sequence data*, dimana data-data tersebut sudah pernah dipublikasikan dalam berbagai jurnal.Berikut ini adalah sampel data yang akan digunakan pada tugas akhir ini:

Tabel 1 menjelaskan data penyakit leukemia. Data tersebut terdiri dari 38 sampel *data training* dan 34 sampel *data testing* dengan 7129 atribut,yang terdiri dari 2 kelas yaitu ALL dan AML.

Tabel 1. Sampel Data Leukimia ALL-AML

No	Atribut 1	Atribut 2	Atribut 3	Atribut 4	Atribut 5	…	Atribut 7129	Kelas
1	-214	-153	-58	88	-295	…	-37	ALL
2	-139	-73	-1	283	-264	…	-14	ALL
3	-76	-49	-307	309	-376	…	-41	ALL
4	-135	-114	265	12	-419	…	-91	ALL
5	-106	-125	-76	168	-230	…	-25	ALL
…	…	…	…	…	…	…	…	…
38	-135	-186	-70	337	-407	…	-10	AML

Tabel 2 menjelaskan data penyakit colon tumor.Pada tersebut terdapat 62 record dengan 2000 atribut yang memiliki kelas positif (mengidap tumor usus) dan negatif (normal).

Tabel 2. Sampel Data Colon Tumor

No	Atribut 1	Atribut 2	Atribut 3	Atribut 4	Atribut 5	…	Atribut 2000	Kelas
1	8589.42	5468.24	4263.41	4064.94	1997.89	…	28.70	Negatif
2	9164.25	6719.53	4883.45	3718.16	2015.22	…	16.77	Positif
3	3825.71	6970.36	5369.97	4705.65	1166.55	…	15.16	Negatif
4	6246.45	7823.53	5955.84	3975.56	2002.61	…	16.09	Positif
5	3230.33	3694.45	3400.74	3463.59	2181.42	…	31.81	Negatif
…	…					…	…	…
62	7472.01	3653.93	2728.22	3494.48	2404.67	…	39.63	Positif

4. Pengujian dan Analisis

4.1 Strategi Pengujian

Strategi dari pengujian sistem ini adalah sebagai berikut:

a. Menentukan parameter yang akan digunakan
b. Melakukan preprocessing data,sebelum data digunakan. Preprocessing data dilakukan dengan PCA
c. Pembagian data menggunakan *percentage split 70,30* dan *K-Fold crossvalidation dengan k=3*.
d. Menentukan ukuran populasi (N) menjadi 50, 100, dan 200 populasi,dengan maksimum generasi 20,10 dan 5.
e. Menentukan probabilitas crossover (Cr) dan parameter pendukung pada DE (F) 0,6 dan 0,8.
f. Memasukan data *training* pada DE sehingga didapatkan parameter gamma dan sigma untuk LSSVM.
g. Memasukan gamma dan sigma pada LSSVM,yang kemudian diproses serta diuji dengan data *testing*.
h. Mendapatkan akurasi hasil klasifikasi LSSVM yang kemudian dihitung performansi dengan menggunakan LSSVM.

4.2 Analisis dan pengujian

4.2.1 Skenario 1

Pada skenario 1 akan dilakukan percobaan pada data penyakit *colon tumor* dan leukimia dengan jumlah populasi (N) masing-masing 50, 100, 200. Kombinasi F dan Cr masing-masing 0.6 dan 0.8. Pada skenario ini akan digunakan pembagian data menggunakan metode *percentage split* dengan proporsi 70% trainning dan 30% testing. Berikut ini merupakan hasil percobaan skenario 1 pada data *colon tumor* dan leukimia:

Tabel 3. Hasil Percobaan Skenario 1 Pada Data Colon Tumor

Kombinasi	Penyakit	Max Gen	N	F	Cr	Gamma	Sigma	Akurasi Training (%)	Akurasi Testing(%)
1	Colon	10	50	0.6	0.6	127.7098	40.7366	100	68.4211
2				0.6	0.8	19.6598	91.9528	100	78.9474
3				0.8	0.6	34.8363	47.8685	100	78.9474
4				0.8	0.8	94.631	29.14	100	78.9474
5		10	100	0.6	0.6	80.0877	21.9168	100	73.6842
6				0.6	0.8	53.8209	29.5632	100	78.9474
7				0.8	0.6	28.1254	46.9546	100	84.2105
8				0.8	0.8	15.2589	71.4617	100	63.1579
9		10	200	0.6	0.6	2.3675	104.341	100	78.9474
10				0.6	0.8	56.4113	23.1733	100	89.4737
11				0.8	0.6	61.607	23.0864	100	84.2105
12				0.8	0.8	5.0166	39.2933	100	73.6842

Berdasarkan tabel 3 diatas, akurasi terbaik adalah 89.4737 %. Akurasi ini diperoleh dari kombinasi F 0.6 dan Cr 0.8, dengan N 200. Dengan akurasi sebesar 89.4737%, maka diperoleh gamma dan sigma sebesar 55.4113 dan 23.1733 Gamma dan sigma tersebut merupakan parameter yang paling optimal. Selain itu berdasarkan tabel diatas terlihat bahwa semakin tingginya kombinasi F,Cr tidak menjamin akan menghasilkan akurasi yang tinggi juga. Misalnya, kombinasi 8 dan 12 pada jumlah populasi 100 akurasi yang diperoleh lebih kecil bila dibandingkan dengan populasi 200. Sedangkan pada kombinasi 4 akurasi yang diperoleh lebih besar bila dibandingkan dengan kombinasi 8 dan 12 yaitu sebesar 78.9474%. Pengambilan data yang dilakukan secara random juga dapat mempengaruhi parameter optimal yang didapat, misalnya gamma dan sigma pada kombinasi 5 menghasilkan nilai gamma dan sigma yang lebih besar dibandingkan dengan kombinasi 6. Tetapi akurasi yang diperoleh oleh kombinasi 5 jauh lebih kecil bila dibandingkan dengan akurasi pada kombinasi 6 yakni sebesar 73.6842%, hal ini membuktikan bahwa pengambilan data random dapat mempengaruhi hasil paramater optimal.

Tabel 4. Hasil Percobaan Skenario 1 Pada Data Leukimia

Kombinasi	Penyakit	Max Gen	N	F	Cr	Gamma	Sigma	Akurasi Training (%)	Akurasi Testing (%)
1	Leukemia	20	50	0.6	0.6	3.3919	95.0956	100	68.1818
2				0.6	0.8	13.4379	39.4101	100	77.2727
3				0.8	0.6	12.729	48.0364	100	77.2727
4				0.8	0.8	71.0902	60.1228	100	81.8182
5		10	100	0.6	0.6	182.114	10.4454	100	72.7273
6				0.6	0.8	78.4908	80.1659	100	68.1818
7				0.8	0.6	112.703	80.6486	100	68.1818
8				0.8	0.8	52.8114	32.8629	100	86.3636
9		5	200	0.6	0.6	35.5974	44.4453	100	81.8182
10				0.6	0.8	47.4133	7.3536	100	81.8182
11				0.8	0.6	16.3079	61.884	100	68.1818
12				0.8	0.8	16.0594	68.9796	100	81.8182

Berdasarkan Tabel 4 diatas, diperoleh gamma dan sigma paling optimal sebesar 52.8114 dan 32.8629. Parameter optimal tersebut diperoleh dari kombinasi 8. Akurasi yang diperoleh dari gamma dan sigma tersebut adalah 86.3636%. Akurasi ini merupakan akurasi terbesar pada data leukimia. Sama halnya dengan data *colon tumor* paramater gamma dan sigma tidak dapat menjamin tingginya akurasi yang akan diperoleh, hal ini bisa disebabkan kerena pangambilan data yang random.

4.2.2 Skenario 2

Pada skenario ini akan dilakukan percobaan pada data colon tumor dan leukemia, menggunakan pembagian data *k-fold crossvalidation* dengan k=3. Berikut merupakan hasil dan analisis skenario 2 :

Tabel 5. Percobaan Skenario 2 Pada Data Colon Tumor

Kombinasi	Penyakit	Max Gen	N	F	Cr	Gamma	Sigma	Akurasi Training (%)	Akurasi Testing (%)
1	Colon	20	50	0.6	0.6	39.1946	16.2949	100	80
2				0.6	0.8	20.677	19.4586	100	85.7143
3				0.8	0.6	107.168	23.2091	100	80
4				0.8	0.8	109.5948	28.3042	100	85.7143
5		10	100	0.6	0.6	51.2065	29.8364	100	85
6				0.6	0.8	40.7811	58.8074	100	80.9524
7				0.8	0.6	77.8788	22.4465	100	85
8				0.8	0.8	105.6745	30.2171	100	80.9524
9		5	200	0.6	0.6	101.0654	17.7081	100	85.7143
10				0.6	0.8	6.0188	24.0976	100	90.4762
11				0.8	0.6	94.982	76.0517	100	75
12				0.8	0.8	95.889	262413	100	85

Berdasarkan tabel 5 pada data colon tumor akurasi tertinggi diperoleh dari kombinasi 10 dengan nilai akurasi sebesar 90.4762. Dari akurasi tertinggi diperoleh nilai gamma dan sigma paling optimal sebesar 6.0188 dan 24.0976. Pada data colon tumor dengan menggunakan metode pembagian data *k-fold crossvalidasi* menghasilkan akurasi yang lebih tinggi bila dibandingkan dengan dengan pembagian data dengan *percentage split.* Hal ini terlihat dari akurasi tertinggi pada skenario 1 adalah 89.4737% sedangkan akurasi tertinggi pada skenario 2 adalah 90.4762%.

Tabel 6. Percobaan Skenario 2 Pada Data Leukimia

Kombinasi	Penyakit	Max Gen	N	F	Cr	Gamma	Sigma	Akurasi Training (%)	Akurasi Testing (%)
1	Leukemia	20	50	0.6	0.6	83.732	9.2346	100	87.5
2				0.6	0.8	69.1901	38.8928	100	79.1667
3				0.8	0.6	60.1704	4.0846	100	70.8333
4				0.8	0.8	17.7697	18.6268	100	79.1667
5		10	100	0.6	0.6	52.6401	35.3451	100	79.1667
6				0.6	0.8	103.013	14.3516	100	79.1667
7				0.8	0.6	97.5374	30.3643	100	75
8				0.8	0.8	42.8229	38.9035	100	83.3333
9		5	200	0.6	0.6	53.4774	1.8063	100	79.1667
10				0.6	0.8	37.0595	40.4285	100	83.3333
11				0.8	0.6	123.445	110.371	100	79.1667
12				0.8	0.8	74.4618	6.4686	100	79.1667

Berdasarkan hasil percobaan pada tabel 6 diperoleh akurasi tertinggi sebesar 87.5% pada kombinasi 1, dengan gamma dan sigma sebesar 83.732 dan 9.2346. Selain itu terdapat beberapa kombinasi yang memiliki akurasi yang sama seperti kombinasi 2, 4 , 5, 6, 9, 11, 12 dengan akurasi sebesar 79.1667 %. Hal ini menunjukkan bahwa pembagian data juga berpengaruh terhadap nilai gamma dan sigma yang didapat. Bila dibandingkan dengan skenario 1 dengan tertinggi 86.3636%, pada skenario 2 ini memiliki akurasi tertinggi 87.5%. Selisih dari keduanya adalah 1.1364%.Hal ini menunjukkan bahwa pada data leukemia akurasi tertinggi diperoleh dari pembagian data menggunakan *k-fold crossvalidasi*

5. Kesimpulan

Bab ini berisi kesimpulan dari tugas akhir dan saran –saran yang dapat digunakan untuk pengembangan lebih lanjut.

1. Berdasarkan percobaan yang dilakukan Algoritma DE dan LSSVM dapat digunakan untuk memprediksi penyakit pada data berdimensi tinggi. Algoritma DE akan menghasilkan parameter yang dibutuhkan oleh LSSVM yaitu gamma dan sigma, yang kemudian akan digunakan oleh LSSVM untuk menghasilkan prediksi penyakit pada data dimensi tinggi.
2. Berdasarkan percobaan yang telah dilakukan pada data colon tumor gamma dan sigma yang paling optimal diperoleh dari kombinasi 10 pada skenario 2 dengan menggunakan pembagian data k-fold crossvalidasi. Dimana percobaan ini mendapatkan akurasi sebesar 90.4762%. Sedangkan pada data leukemia, paramater gamma dan sigma paling optimal diperoleh dari kombinasi 1 dengan akurasi sebesar 87.5% dengan k-fold crossvalidasi
3. Performansi terbaik diperoleh dari pembagian data dengan menggunakan crossvalidasi. Hal ini dikarenakan crossvalidasi membagi data secara acak dan memilih data training dan testing berdasarkan akurasi tertinggi.

5.1 Saran

1. Penelitian ini juga dapat dilakukan menggunakan skema DE 1.
2. Menggunakan variasi nilai F dan Cr yang lebih besar untuk mendapatkan hasil yang lebih optimal.
3. Menggunakan *crossvalidasi* dengan pembagian data yang berbeda misalnya 80-20 atau 50-50 untuk mendapatkan hasil yang optimal.

Daftar Pustaka

[1] Omar S.Soliman, Eman AboElHamd . "Classification of Breast Cancer using Differential Evolution and LeastSquares Support Vector Machine," IJCSI International Journal of Computer Science Issue, Vol.3, Issue 2, Maret-April (2014).

[2] A. K. Qin, V. L. Huang, and P. N. Suganthan "Differential Evolution Algorithm With Strategy Adaptation for Global Numerical Optimization," IEEE TRANSACTION ON EVOLUTIONARY COMPUTATION, Vol.13, No.2, April (2009)

[3] Suyanto, Evolutionary Computation, Komputasi Berbasis "Evolusi" dan"Genetika",Bandung : INFORMATIKA, (2008).

[4] Suyanto, Soft Computing "Membangun Mesin Ber-IQ Tinggi," Bandung : INFORMATIKA, (2008).

[5] Michel Verleysen, "Learning high-dimensional data," IOS Press, pp. 141-162, (2003).

[6] L. Hoegaerts, J.A.K. Suykens, J. Vandewalle, B. De Moor, "A Comparison of Pruning Algorithms for Sparse Least Squares Support Vector Machines", Lecture Notes In Computer Science, January (2004).

[7] Nasser H. Sweilam, A.A. Tharwat, N.K. Abdel Moniem," Support vector machine for diagnosis cancer disease:A comparative study",Egyptian Informatics Journal,September (2010).

[8] Zhang Nian, Charles Williams, Esther Ososanya, Wagdy Mahmoud,"Streamflow Prediction Based on Least Squares Support Vector Machines" ,University of the District of Columbia Department of Electrical and Computer Engineering 4200 Connecticut Ave. NW Washington, DC, (2008).

[9] Ding, Chris dan Xiaofeng He, "K-means Clustering via Principal Component Analysis", Lawrence Berkeley National Laboratory, Berkeley (2004).

[10] Smith, Lindsay I."A tutorial on Principal Components Analysis",Februari 2002.

[11] Jackson, J.Edward. A User's Guide to Principal Components. New York, (1991).

[12] Prasetyo Eko, Data Mining, Mengolah Data Menjadi Informasi Menggunakan Matlab,Yogyakarta : ANDI, (2014).

[13] Dash Manoranjan, Huan Liu, Dimentionality Reduction, Tersedia : http://www.public.asu.edu/~huanliu/papers/dm07.pdf/.

[14] Li, J., 2009. http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html. [Online]