

Mendeteksi Potensi Terjangkitnya Penyakit Kanker pada Pasien dengan Memanfaatkan Microarray Data Ekspresi Gen Menggunakan Kombinasi Algoritma Genetika dan Support Vector Machine

Choiryaldi Setya Pratama^{1,a)}, Salamet Nur Himawan^{1,b)}, Triati Dewi Kencana Wungu^{2,c)} dan Suprijadi^{3,d)}

¹Program Magister Sains Komputasi,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung,
Jl. Ganesha no. 10 Bandung, Indonesia, 40132

²Kelompok Keilmuan Fisika Nuklir dan Biofisika,
Program Studi Fisika,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung,
Jl. Ganesha no. 10 Bandung, Indonesia, 40132

³Kelompok Keilmuan Fisika Teori Energi Tinggi dan Instrumentasi,
Program Studi Fisika,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung,
Jl. Ganesha no. 10 Bandung, Indonesia, 40132

^{a)} choiryaldisp8@gmail.com (corresponding author)

^{b)} salametnurhimawan@gmail.com

^{c)} triati@fi.itb.ac.id

^{d)} supri@fi.itb.ac.id

Abstrak

Pada penelitian ini, Algoritma Genetika (AG) dengan Support Vector Machine (SVM) digunakan dalam mengidentifikasi potensi terjangkitnya penyakit kanker melalui deteksi ekspresi gen. AG digunakan dalam proses seleksi pemilihan fitur dari ekspresi gen sedangkan SVM digunakan dalam proses klasifikasi potensi terjangkitnya suatu penyakit pada pasien. Terdapat dua proses pengolahan data yaitu pre-processing data dan pre-filtering data yang memudahkan dalam proses pemilihan fitur. Data yang digunakan pada tahap pengujian system menggunakan tiga set data ekspresi gen pasien dengan kelas yang berbeda yaitu kanker karsinoma, kolon tumor dan leukemia. Pada tahap pengujian menggunakan proses pre-filtering dan validasi silang k-fold menghasilkan akurasi sebesar 100% untuk data set leukemia, 90.40% untuk kolon tumor dan 100% untuk kanker karsinoma. Pengujian menggunakan pre-filtering memberikan hasil yang lebih baik dibandingkan tanpa menggunakan pre-filtering baik dari segi akurasi maupun jumlah fitur yang terpilih.

Kata-kata kunci: Ekspresi gen, microarray data, Algoritma Genetika, Support Vector Machine, kanker.

PENDAHULUAN

Diagnosa potensi penyakit pada pasien dapat dilakukan melalui ekspresi gen. Ekspresi gen merupakan rangkaian proses penerjemahan genetik dalam bentuk urutan basa pada DNA atau RNA. Banyak algoritma telah dikembangkan untuk mengklasifikasi potensi terjangkitnya suatu penyakit berdasarkan ekspresi gen

yang telah diseleksi dan secara signifikan dapat mengklasifikasikan penyakit dengan akurasi yang tinggi [1]. Data ekspresi gen berkaitan dengan data berdimensi tinggi sehingga pada proses klasifikasi akan melibatkan komputasi yang tinggi karena tidak menutup kemungkinan ada fitur yang sebenarnya tidak perlu diikutsertakan dalam proses klasifikasi tetapi diikutsertakan sehingga komputasi data menjadi lambat dan hasil klasifikasi pun kurang akurat. Sebagian besar algoritma klasifikasi dapat bekerja lebih baik jika jumlah fitur yang digunakan sedikit [2] maka dibutuhkan sebuah metode untuk memilih fitur yang tepat dalam mengklasifikasikan kelas terjangkit atau tidak terjangkitnya suatu penyakit pada pasien serta mengoptimasikan proses klasifikasi.

Beberapa metode klasifikasi yang telah dikembangkan menghasilkan akurasi yang baik dan mudah diimplementasikan pada kasus data berjenis *supervised* maupun *unsupervised*, seperti *Artificial Neural Network* (ANN), *Support Vector Machine* (SVM), *Naïve Bayes* dan metode *learning* lainnya [3]. Bagaimana pun, semakin besar dimensi yang digunakan dalam proses klasifikasi akan membuat komputasi didalamnya semakin tinggi. Diperlukan sebuah algoritma optimasi yang dapat mendukung dalam mengoptimalkan proses klasifikasi yang berfungsi sebagai *feature selection*. *Genetic Algorithm* (GA), *Simulated Annealing* (SA) dan algoritma optimasi lainnya telah banyak digunakan dalam permasalahan *feature selection* [4]. Mengkombinasikan dua metode atau lebih, secara statistik dapat menghasilkan akurasi yang baik, efisien dan efektif dalam memecahkan studi kasus. Pada studi kasus klasifikasi penyakit berdasarkan data ekspresi gen merupakan permasalahan yang dapat dipecahkan dengan kombinasi metode *learning* sebagai metode klasifikasi dan metode heuristik sebagai *feature selection*.

Pada makalah ini, kami mengklasifikasikan potensi pasien terjangkit penyakit berdasarkan data ekspresi gen dengan menggunakan *Support Vector Machine* (SVM) sebagai metode klasifikasi dan *Genetic Algorithm* (GA) sebagai *feature selection*. Pada sistem ini menghasilkan fitur-fitur yang telah terpilih, hasil prediksi serta tingkat akurasi dari sistem yang dikembangkan.

MATERIAL DATA DAN METODE PENELITIAN

Set Data

Penelitian ini menggunakan tiga data set *microarray* ekspresi gen dari pasien yaitu data set leukemia, data set kolon tumor dan data set kanker karsinoma. Setiap data set memiliki dimensi yang berbeda, pada data set leukemia memiliki jumlah data pasien yang tersimpan (*record*) 72 dengan jumlah fitur 7129, data set kolon tumor memiliki jumlah *record* 62 dengan jumlah fitur 2000 dan data set kanker karsinoma memiliki jumlah *record* 36 dengan jumlah fitur 7457. Tabel 1 menunjukkan contoh dari sampel data set leukemia yang mengandung informasi konsentrasi mRNA pada setiap pasien leukemia dan fiturnya. Setiap fitur menggambarkan kode gen yang dapat diakses dan diunduh pada database [5].

Tabel 10. Sampel data dari data set leukemia

	Fitur 1	Fitur 2	Fitur 3	Fitur 4	...	Fitur 7129
Pasien 1	-214	-153	-58	88	...	-295
Pasien 2	-139	-73	-1	283	...	-264
Pasien 3	-76	-49	-307	309	...	-376
Pasien 4	-106	-125	-76	168	...	-230
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Pasien 72	-176	-284	-81	9	...	-294

Setiap elemen dalam matriks *microarray* data merepresentasikan nilai tingkatan dari suatu ekspresi gen. Data ekspresi gen merupakan rangkaian *Deoxyribonucleic acid* (DNA). DNA merupakan molekul yang mengkodekan semua informasi yang diperlukan dalam perkembangan dan fungsional suatu organisme [6]. Kerja sel dapat dilihat dari protein yang dihasilkan dari kombinasi DNA ini, kemudian DNA akan mengalami transkripsi menjadi *Ribo-nucleic acid* (RNA). RNA merupakan rangkaian nukleotida seperti halnya DNA namun nukleotida *Thymine* diganti dengan *Uracil*. Pada proses transkripsi molekul DNA pada salah satu benang disalin menjadi pre-mRNA. Menurut *Central dogma of molecular biology*, setiap gen akan menghasilkan satu mRNA yang kemudian menghasilkan satu protein, maka level ekspresi gen satu gen dengan gen yang lain akan berbeda-beda. Teknologi *microarray* dapat mengukur ekspresi gen dalam jumlah besar secara simultan [7].

Data mentah pada eksperimen *microarray* berbentuk gambar, untuk mendapatkan data mengenai tingkat ekspresi gen, gambar tersebut harus dianalisa. Masing-masing titik diidentifikasi, diukur intensitasnya dan dibandingkan dengan latar belakangnya. Data tersebut kemudian dimasukkan kedalam sebuah hasil akhir berupa matriks ekspresi gen.

Pre-Filtering Data

Proses seleksi fitur diawali dengan adanya proses *pre-filtering* data. Proses *pre-filtering* data merupakan proses pembagian setiap fitur kedalam suatu nilai interval. Nilai interval dibuat berdasarkan nilai standar deviasi setiap fitur, dimana fitur yang memiliki nilai standar deviasi berada pada interval atas dan interval bawah tidak akan digunakan. Fitur yang digunakan adalah fitur yang memiliki nilai standar deviasi antara nilai interval atas dan interval bawah.

Genetic Algorithm (GA) Sebagai Metode Seleksi Fitur

Pada sistem ini seleksi fitur menggunakan algoritma genetika dimana pada representasi kromosom menggunakan bilangan biner. Bilangan biner dibuat secara acak dengan ukuran sebesar dimensi $Npop \times Nbit$, dimana $Npop$ merupakan jumlah populasi yang akan dimasukkan, sedangkan $Nbit$ berjumlah sama dengan jumlah fitur pada data. Jika nilai biner bernilai 1 maka diidentifikasi sebagai fitur yang terpilih sedangkan jika biner bernilai 0 maka fitur tidak terpilih. Pada algoritma genetika, secara umum terdapat 5 proses didalamnya, yaitu inisialisasi populasi, evaluasi *fitness*, seleksi orang tua, elitisme, pindah silang, mutasi dan seleksi *survivor* [8]. Setiap proses dalam algoritma genetika mengandung beberapa skema yang dapat diterapkan. Pada penelitian ini proses seleksi orang tua menggunakan skema *roulette wheel*, proses pindah silang menggunakan skema *uniformly crossover*, proses mutasi menggunakan skema mutasi tingkat gen, dan pada proses seleksi *survivor* menggunakan skema *general replacement*. Untuk proses evaluasi *fitness* dijelaskan di subbab SVM.

Algoritma genetika memiliki beberapa kelebihan yaitu algoritma ini hanya melakukan sedikit perhitungan matematis yang berhubungan dengan masalah yang ingin diselesaikan. Kemudian operator-operator evolusi yang digunakan membuat algoritma ini sangat efektif pada pencarian global, dan memiliki fleksibilitas yang tinggi untuk di-hybrid-kan dengan metode pencarian lainnya agar lebih efektif. Selain kelebihan, algoritma genetika juga memiliki kekurangan yaitu permasalahan dalam melakukan identifikasi fungsi *fitness*, mendefinisikan suatu permasalahan, terjadinya konvergensi yang prematur, banyaknya parameter yang akan digunakan seperti jumlah populasi (*population size*), tingkat mutasi (*mutation rate*), tingkat persilangan (*crossover rate*), metode seleksi, dan yang terakhir tidak dapat digunakan untuk permasalahan *local optima*.

Support Vector Machine (SVM) Sebagai Metode Klasifikasi

Proses klasifikasi menggunakan metode SVM untuk setiap data set dikelompokkan kedalam 2 kelas yaitu positif dan negatif. Kelebihan dari metode SVM adalah SVM dapat menentukan hyperplane atau bidang pemisah dengan memilih bidang dengan optimal margin maka generalisasi pada SVM dapat terjaga dengan sendirinya, tingkat generalisasi pada SVM tidak dipengaruhi oleh jumlah data latih, dengan menentukan parameter soft margin, noise dapat dikontrol sehingga makin besar parameter soft margin, makin besar pula pinalti yang dikenakan pada kesalahan pada klasifikasi sehingga proses pelatihan semakin ketat. Kekurangan SVM adalah SVM menggunakan fungsi pemisah yang memisahkan data ke dalam dua kelas, jika kelas yang ingin dipisahkan lebih dari dua maka dibutuhkan modifikasi, ketika data latih yang digunakan sangat besar, mempengaruhi waktu pelatihan dan ukuran memory yang dibutuhkan.

Dalam prosedur SVM, terdapat tiga jenis fungsi kernel yang dapat digunakan yaitu linier, polynomial dan radial basis fungsi (RBF)[4], masing-masing kernel memiliki persamaan seperti yang ditunjukkan pada persamaan (2.1) – (2.3).

- | | | |
|------------------------------|--|-------|
| a. Kernel Linear | $K(x_i, x) = x_i^T x$ | (2.1) |
| b. Kernel Polynomial | $K(x_i, x) = (\gamma \cdot x_i^T x + r)^p, \gamma > 0$ | (2.2) |
| c. Radial Basis Fungsi (RBF) | $K(x_i, x) = \exp(-\gamma x_i - x ^2), \gamma > 0$ | (2.3) |

Dari ketiga fungsi kernel tersebut maka akan diterapkan pada tiga data set (leukimia, kolon tumor dan kanker karsinoma). Hasil prediksi kelas menggunakan SVM akan dijadikan nilai *fitness* pada GA sehingga proses SVM menjadi fungsi objektif dari GA itu sendiri dan proses SVM ini akan dilakukan berulang selama proses seleksi fitur pada GA berlangsung.

Validasi Silang sebagai Teknik Validasi Tingkat Akurasi Sistem

Cross validation atau validasi silang merupakan teknik yang dapat digunakan sebagai pengujian sistem yang melibatkan data. Metode ini membagi data menjadi dua bagian, yaitu data pelatihan dan data pengujian. Kemudian, setelah data diujikan pada sistem lalu dilakukan proses silang dimana data pengujian dijadikan data pelatihan ataupun sebaliknya, data pelatihan sebelumnya dijadikan data pengujian. Pada pengujian sistem ini menggunakan jenis validasi silang *k-fold*, dimana variabel k menjadi porsi partisi dari data set. Pada penelitian ini variabel k yang digunakan adalah $k=3$.

HASIL PENGUJIAN SISTEM DAN ANALISA HASIL

Pada pengujian sistem ini akan dilakukan beberapa skenario pengujian yaitu terdiri dari pengaruh penggunaan ketiga kernel pada SVM, pengaruh proses pre-filtering dan pengaruh parameter P_c dan P_m terhadap tingkat konvergensi pada GA. Dikarenakan GA merupakan metode heuristik yang membangkitkan bilangan acak sehingga setiap simulasi pada program memberikan hasil yang dapat berbeda maka simulasi dilakukan dalam lima kali perulangan terhadap tiga data set. Dari hasil simulasi dengan menggunakan tiga fungsi Kernel (SVM) diperoleh data sebagai berikut (Tabel 2).

Tabel 11. Pengaruh fungsi Kernel (SVM) pada tiga data set terhadap akurasi prediksi kelas

Data Set	GA		SVM	Fitur Terpilih						K-fold (%)					
	Skema	Parameter*)		NonPre-filtering			Pre-filtering			NonPre-filtering			Pre-filtering		
				Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
Leukimia	<ul style="list-style-type: none"> • <i>Roulette Wheel</i> • <i>Uniform Crossover</i> • Mutasi tingkat Gen 	Pc = 0.5 Pm = 0.1 Npop = 20 Nbit = 7129 MaxGen = 50	Linear	3505	3549	3581	28	35	41	98.61	99.19	100	98.61	99.46	100
			Polynomial (4)	3517	3536	3565	24	31	35	68.28	68.28	68.28	88.70	91.34	94.43
			RBF (1)	3460	3504	3559	21	27	36	62.28	62.28	62.29	65.29	66.42	68.08
Kolon tumor	<ul style="list-style-type: none"> • <i>Roulette Wheel</i> • <i>Uniform Crossover</i> • Mutasi tingkat Gen 	Pc = 0.5 Pm = 0.1 Npop = 20 Nbit = 2000 MaxGen = 50	Linear	957	976	984	201	208	213	87.06	87.72	88.73	88.81	90.02	90.40
			Polynomial (4)	966	989	1010	206	232	235	64.52	65.80	67.73	66.03	67.72	69.44
			RBF (1)	973	1010	1016	216	226	239	64.52	64.53	64.55	64.52	64.53	64.55
Kanker karsinoma	<ul style="list-style-type: none"> • <i>Roulette Wheel</i> • <i>Uniform Crossover</i> • Mutasi tingkat Gen 	Pc = 0.5 Pm = 0.1 Npop = 20 Nbit = 7457 MaxGen = 50	Linear	3680	3730	3818	50	59	67	97.22	97.77	100	97.22	98.88	100
			Polynomial (4)	3684	3729	3773	49	54	60	50.00	50.00	50.00	69.44	72.22	75.00
			RBF (1)	3681	3707	3738	53	58	63	50.00	50.00	50.00	65.89	66.51	66.77

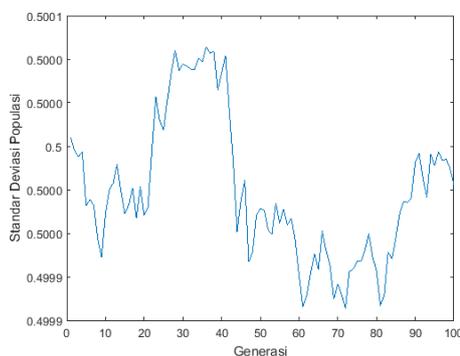
*)Pc = Probabilitas *Crossover*, Pm = Probabilitas Mutasi, Npop = Jumlah Individu dalam Setiap Populasi, Nbit = Jumlah Kromosom, MaxGen = Maksimum Iterasi/Generasi.

Berdasarkan Tabel 2 nilai akurasi tertinggi yaitu pada penggunaan fungsi linear Kernel. Fungsi linear kernel menghasilkan nilai maksimum akurasi lebih tinggi dibandingkan dengan fungsi polynomial Kernel dan RBF. Fungsi linear Kernel menghasilkan nilai maksimum akurasi mencapai 100% pada data set leukimia. Hasil akurasi 100% menunjukkan bahwa algoritma sistem dapat membedakan pasien mana yang terjangkit leukimia dan yang tidak terjangkit leukimia dari total *record* 72 pasien.

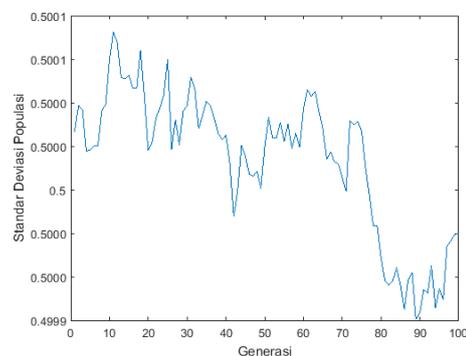
Pada penggunaan fungsi linear dapat memberikan hasil yang lebih baik dibandingkan fungsi kernel lainnya dikarenakan karakteristik data yang dapat dipisahkan menggunakan fungsi linear pada saat membuat *hyperplane* pemisah, dalam beberapa kasus yang bersifat non-linear tidak dapat dipecahkan menggunakan fungsi linear maka pada SVM terdapat fungsi lainnya seperti polynomial kernel, radial basis fungsi atau sigmoid kernel yang lebih dikenal dengan *kernel trick*. *Kernel trick* mentransformasikan data kedalam dimensi ruang fitur (*feature space*) sehingga dapat dipisahkan secara linear pada *feature space*. *Feature space* biasanya memiliki dimensi yang lebih tinggi dari vector input (*input space*). Hal ini mengakibatkan komputasi pada *feature space* mungkin sangat besar, karena ada kemungkinan *feature space* dapat memiliki jumlah *feature* yang tidak terhingga maka dapat menggunakan *kernel trick* dalam penyelesaian permasalahan ini. Syarat sebuah fungsi menjadi fungsi kernel adalah memenuhi teorema Mercer yang menyatakan bahwa matriks kernel yang dihasilkan harus bersifat *positive semi-definite*.

Pada tabel 2 juga menunjukkan bahwa proses *pre-filtering* dapat membantu proses pemilihan fitur yang optimal pada GA. Hal ini dikarenakan mempersempit ruang pencarian pada GA dalam mencari fitur yang optimal sehingga proses pencarian menjadi lebih cepat dan mengurangi kompleksitas dari komputasi. Selain itu, dengan melibatkan fitur yang banyak dalam proses *classifier* pada SVM dapat mengakibatkan *overfitting* terhadap proses pembuatan garis pemisah *hyperplane* sehingga mempengaruhi akurasi dari proses klasifikasi.

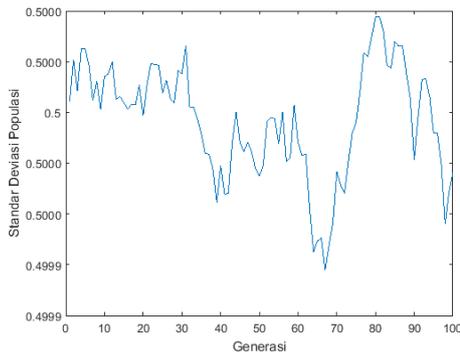
Berikutnya, pada pengujian parameter P_c dan P_m ini terdiri dari 9 pengujian dengan mengkombinasikan parameter P_c dan P_m yang terdiri dari $P_c = [0.5, 0.7, 0.9]$ dan $P_m = [0.01, 0.1, 0.2]$ serta menggunakan jumlah individu untuk setiap populasi ($N_{pop} = 100$) dan maksimum iterasi/generasi ($MaxGen = 100$). Pengujian ini bertujuan untuk melihat pengaruh perubahan parameter P_c dan P_m terhadap proses pencarian solusi optimal pada GA.



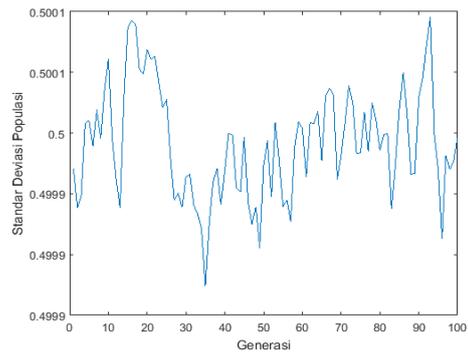
Gambar 8. $P_m = 0.01$ dan $P_c = 0.5$



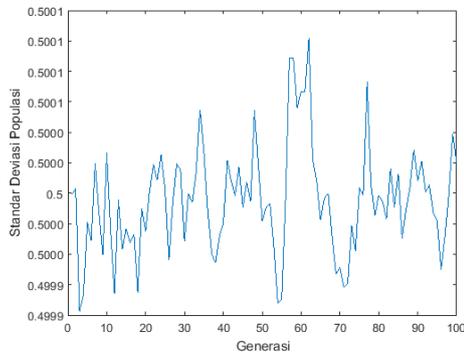
Gambar 9. $P_m = 0.01$ dan $P_c = 0.7$



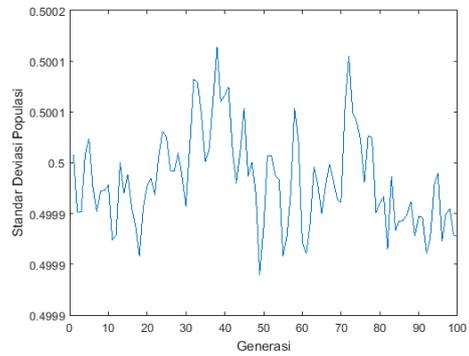
Gambar 10. $P_m = 0.01$ dan $P_c = 0.9$



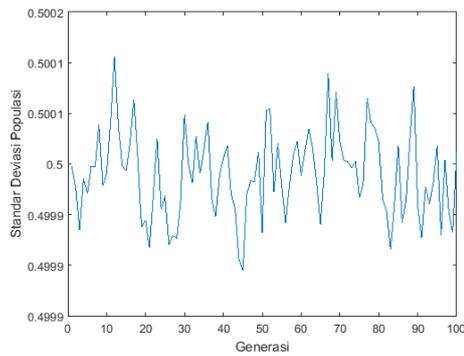
Gambar 11. $P_m = 0.1$ dan $P_c = 0.5$



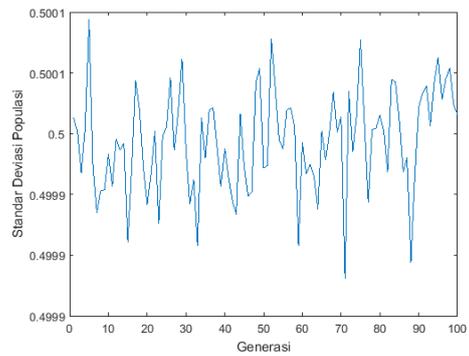
Gambar 12. $P_m = 0.1$ dan $P_c = 0.7$



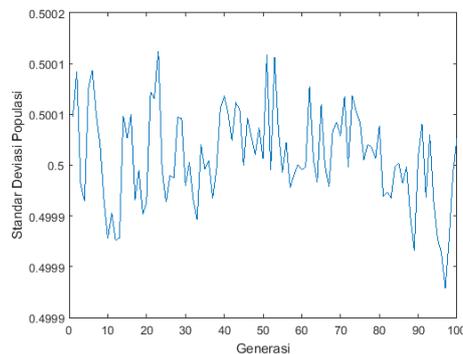
Gambar 13. $P_m = 0.1$ dan $P_c = 0.9$



Gambar 14. $P_m = 0.2$ dan $P_c = 0.5$



Gambar 15. $P_m = 0.2$ dan $P_c = 0.7$



Gambar 16. $P_m = 0.2$ dan $P_c = 0.9$

Pada grafik percobaan diatas yang ditunjukkan oleh Gambar 1 – Gambar 9 merupakan pergerakan nilai sebaran dari populasi yang terdiri dari 100 individu dalam tiap generasi pada GA. Dapat terlihat bahwa apabila nilai P_m semakin besar dan P_c yang semakin kecil maka kondisi sebaran individu pada populasi semakin variatif sedangkan nilai P_m semakin kecil dan P_c yang semakin besar sebaran individu pada populasi akan mendekati konvergen. Hal ini dapat dilihat pada perbedaan gambar diatas dimana semakin besar nilai P_m pergerakan grafik semakin fluktuatif yang mengakibatkan perubahan nilai sebaran lebih cenderung cepat terjadi pada setiap generasi dibandingkan dengan semakin besarnya nilai P_c .

Pada metode GA terdapat kekurangan yang sering dihadapi yaitu salah satunya adalah konvergensi prematur. Konvergensi prematur adalah dimana kondisi tiap individu pada populasi menjadi konvergen pada suatu titik dan mengakibatkan pencarian akan terjebak pada suatu domain. Proses konvergen dapat membuat proses pencarian menemukan solusi global optimum, akan tetapi proses konvergen ini pula dapat membuat proses pencarian terjebak pada solusi lokal optimum dan hal ini yang disebut dengan konvergensi prematur pada GA. Pada GA terdapat dua parameter yang dapat digunakan dalam mengontrol proses konvergen ini dengan mengatur nilai dari parameter P_c dan P_m .

Seperti yang sudah diketahui, metode GA merupakan pencarian heuristik dengan pendekatan probabilistik sehingga nilai parameter P_c dan P_m merupakan nilai peluang suatu individu pada populasi mengalami proses *crossover* atau mutasi. Proses *crossover* pada GA bertujuan untuk individu orang tua yang terbaik akan mewariskan gennya kepada anaknya dengan harapan menghasilkan individu baru yang lebih baik atau kurang lebih sama dengan orang tuanya, akan tetapi sebelum proses *crossover* ini terdapat proses seleksi orang tua dimana setiap individu dibangkitkan nilai probabilitas berdasarkan nilai *fitness* masing-masing, sehingga individu dengan nilai *fitness* yang tinggi akan lebih sering terpilih sebagai calon individu orang tua dan mengakibatkan populasi individu semakin cepat konvergen. Sebaliknya, pada proses mutasi adalah menciptakan gen pada populasi individu semakin variatif, akan tetapi semakin tingginya peluang mutasi yang terjadi pada tiap individu dapat mengakibatkan menurunkan kualitas dari individu tersebut. Hal ini dikarenakan semakin tingginya nilai P_m maka semakin besar nilai peluang tiap gen pada individu kembali dibangkitkan secara acak dan merusak susunan gen sebelumnya, sehingga pemilihan nilai parameter P_c dan P_m dapat mempengaruhi proses pencarian pada GA.

KESIMPULAN

Penelitian ini telah mengombinasikan metoda GA dan SVM sebagai teknik klasifikasi potensi penyakit pada pasien dan mengoptimasi pemilihan fitur untuk meminimumkan kompleksitas komputasi. Pada pengujian yang telah dilakukan dapat terlihat sistem yang dibangun memberikan hasil akurasi yang cukup baik dengan jumlah fitur yang lebih sedikit. Proses *pre-filtering* membantu dalam mempersempit ruang pencarian pada GA untuk mencari fitur yang optimal, sehingga fitur yang optimal dapat membantu metode SVM meningkatkan akurasi dalam mengklasifikasi potensi terkena kanker atau tidak pada pasien. Hal ini dikarenakan semakin banyaknya fitur yang digunakan dapat terjadi *overfitting* dalam pembuatan *hyperplane* atau garis pemisah antar kelas pada SVM.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada berbagai pihak yang telah membantu dalam penulisan makalah ini. Makalah ini didanai oleh Program Riset ITB 2018 dan Riset Pusat Unggulan Perguruan Tinggi (PUPT) Kemenristekdikti 2018.

REFERENSI

1. A. M. Bagirov, B. Ferguson, S. Ivkovic, G. Saunders, dan J. Yearwood, *New Algorithms for Multi-Class Cancer Diagnosis Using Tumor Gene Expression Signatures*. *Bioinformatics*, 19, 1800-1807 (2003).
2. P. Matthew dan R. Calvo, *Fast Dimensionality Reduction and Simple PCA*. *Intelligent Data Analysis*, 2, 203-214 (1998).
3. J. Khan, J. S. Wei, M. Ringner, L, et al, *Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks*. *Nat Med*, 7(6), 673-679 (2001).
4. S. Peng, Q. Xu, X. B. Ling, X. Peng, W. Du, dan L. Chen, *Molecular Classification of Cancer Types from Microarray Data Using the Combination of Genetic Algorithms and Support Vector Machines*. *FEBS Letters*, 555(2), 358-362 (2003).
5. <http://genomics-pubs.princeton.edu/oncology/>, 22 Maret 2018.
6. F. E. Streib dan M. Dehmer, *Analysis of Microarray Data: A Network-Based Approach*. Wiley-VCH (2008).
7. N. Evi, S. P. Yoga, *Pengklastran pasien Kanker Leukimia Berdasarkan Data Ekspresi Gen dengan Menggunakan Dekomposisi Nilai Singular*. *J Math and Its Appl*, 7, 13-25(2010).
8. Suyanto, *Artificial Intelligence: Searching - Reasoning - Planning - Learning*, Edisi Revisi, Bandung (2011).