Klasifikasi *K-Nearest Neighbor* untuk Data Microarray dengan Seleksi *Genetic Algorithm*

Shuni'atul Ma'wa[1], Adiwijaya[2], Aniq A. Rohmawati[3]

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

Abstrak

Microarray adalah teknik modern yang memfasilitasi analisis simulasi dari sejumlah data yang menggambarkan ekspresi gen yang diperlukan untuk memecahkan masalah biologis yang kompleks, seperti deteksi suatu penyakit tertentu. Data microarray memiliki karakteristik berdimensi besar, dimana banyaknya variabel respon lebih kecil dibandingkan variabel prediktor. Oleh karena itu, diperlukan skema yang didalamnya terdapat proses reduksi dimensi dan proses klasifikasi. Dalam hal ini, proses reduksi dimensi bertujuan untuk meringankan beban komputasi serta menghindari *overfitting* pada klasifikasi. Proses reduksi yang digunakan pada penelitian ini yaitu seleksi fitur *Genetic Algorithm* (GA). Kemudian, proses klasifikasi yang bertujuan untuk mendeteksi penyakit kanker atau bukan kanker dilakukan dengan menggunakan metode klasifikasi *K-Nearest Neighbor* (KNN). Adapun akurasi dari metode GA-KNN pada data tumor usus, kanker paru-paru, dan kanker darah memiliki rata rata akurasi sebesar 95,01%.

Kata Kunci: *K-Nearest Neighbor*, *Genetic Algorithm*, Deteksi Kanker, Data Microarray

Abstract

Microarray is a modern technique that facilitates simulation analysis of a number of data that describe the expression of genes needed to solve complex biological problems, such as the detection of a particular disease. Microarray data has large dimension characteristics, where the number of response variables is smaller than the predictor variable. Therefore, a scheme is needed in which there is a dimension reduction process and classification process. In this case, the dimension reduction process aims to ease the computational burden and avoid overfitting the classification. The reduction process used in this study is the selection feature of the Genetic Algorithm (GA). Then, the classification process that aims to detect cancer or non-cancer is carried out using the K-Nearest Neighbor (KNN) classification method. The accuracy of the GA-KNN method in data on colon tumor, lung cancer, and leukimia has an average accuracy of 95.01%.

Keywords: K-Nearest Neighbor, Genetic Algorithm, Cancer Detection, Microarray Data

1. Pendahuluan

Bagian ini akan menjelaskan latar belakang dipilihnya permasalahan penelitian, perumusan masalah, batasan masalah, tujuan dan metodologi penyelesaian masalah dan sistematika penulisan.

Latar Belakang

. Kanker adalah penyakit yang disebabkan oleh pertumbuhan sel-sel abnormal yang tidak terkendali, yang menyebabkan jaringan tubuh normal rusak dan penyebab kematian tertinggi di dunia. Pada tahun 2018, terdapat 18,1 juta jiwa dengan banyaknya kematian akibat kanker sebesar 9,6 juta jiwa [1]. Oleh karena itu, diperlukan teknologi untuk mendeteksi penyakit kanker sejak dini agar mendapat penanganan lebih awal dengan hasil analisis yang akurat. Namun, medeteksi kanker bukan suatu hal yang mudah, sehingga untuk menganalisis penyakit kanker akan dilakukan dengan teknik microarray. Microarray adalah teknologi modern yang digunakan untuk membaca ekspresi gen pada tubuh makhluk hidup khususnya manusia. Teknologi microarray digunakan untuk mengklasifikasikan ekspresi gen kaitannya dengan penyakit tertentu, termasuk penyakit kanker. Klasifikasi ini dilakukan dengan cara menganalisis dan mengelompokan sampel terhadap kelas kanker atau tidak kanker.

K-Nearest Neighbor (KNN) adalah algoritma yang berfungsi untuk melakukan klasifikasi suatu data berdasarkan data latih (*training datasets*), yang diambil dari *K* tetangga terdekatnya (*nearest neighbors*). Dalam penentuan tetangga terdekat dilibatkan perhitungan korelasi antar atribut dan nilai fitness sebagai langkah awal dan pertimbangan pembobotan. KNN merupakan metode yang banyak digunakan karena memiliki kelebihan

diantaranya untuk data yang berdimensi besar dan MKNN memperbaiki kekurangan KNN dalam memilih atribut yang relevan terhadap observasi dengan melibatkan bobot pada hasil fitness. *Genetic Algorithm* (GA) adalah teknik pencarian dalam bidang komputasi untuk menemukan solusi dengan melibatkan pendekatan optimasi dalam pencariannya. Teknik dalam GA didasarkan pada biologi evolusioner seperti pewarisan, mutasi, seleksi dan *crossover*. Keunggulan menggunakan GA adalah kemampuannya untuk memilih atribut-atribut yang relevan pada data berdimensi tinggi berdasarkan proses mutasi pada komrosom individu yang merepresentasikan atribut pada data. GA pernah digunakan untuk meningkatkan kinerja KNN dan Naïve Bayes dalam deteksi diabetes untuk pemilihan fitur oleh R. N. Patil dan S. C. Tamane (2018) [2]. Untuk deteksi tumor otak oleh M. Poornima, C. Kuyin, M. Revathy (2018) [3]. Kemudian, GA pernah digunakan untuk *minimizing the cost of two-tier cellular network with queuing handoff calls in microcell using genetic algorithm* oleh P. Goel dan D. K. Lobiyal (2018) [4].

Topik dan Batasannya

Berdasarkan masalah yang telah disampaikan maka rumusan masalah panda penelitian ini adalah untuk mengetahui bagaimana mengimplementasikan proses seleksi atribut dengan *Genetic Algorithm* (GA) pada data microarray dan bagaimana hasil akurasi klasifikasi microarray dengan *K-Nearest Neighbor*.

Adapun batasan masalah dalam pengerjaan Tugas Akhir yaitu Data yang digunakan dalam penelitian ini adalah antara lain data *colon tumor*, *lung cancer* dan *leukimia* yang didapat dari Kent-Ridge Bio-medical Data Set Repository (http://leo.ugr.es/elvira/DBCRepository/).

Tujuan

Tujuan dari penelitian ini adalah untuk mengetahui bagaimana mengimplementasikan proses seleksi atribut dengan *Genetic Algorithm* (GA) pada data microarray dan bagaimana hasil akurasi klasifikasi microarray dengan *K-Nearest Neighbor*.

Tabel 1. Keterkaitan antara tujuan, pengujian dan kesimpulan

No	Tujuan	Pengujian	Kesimpulan
1	Bagaimana mengimplementasikan proses seleksi atribut dengan GA pada data microarray	Pengujian menggunakan matlab	Tujuan tercapai
2	Bagaimana hasil akurasi klasifikasi microarray dengan KNN	Pengujian menggunakan matlab	Tujuan tercapai

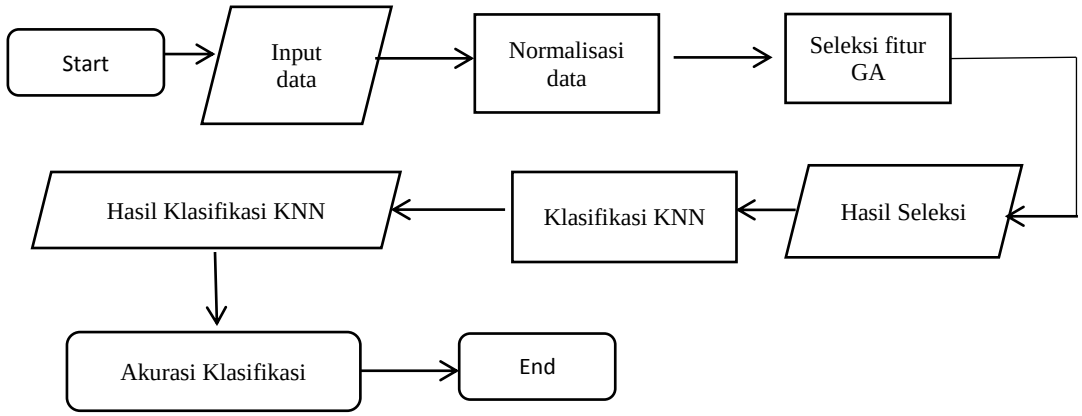
Organisasi Tulisan

Laporan tugas akhir ini disusun dengan sistematika penulisan sebagai berikut:

1. Pendahuluan
   Bab ini berisi penjelasan latar belakang dipilihnya permasalahan penelitian, topik dan batasannya, tujuan dan organisasi tulisan.
2. Studi Terkait
   Bab ini berisi landasan teori yang mendukung penelitian pada tugas akhir ini. Landasan teori ini meliputi penjelasan mengenai microarray, reduksi dimensi, *genetic algorithm*¸ *K-Nearest Neighbor*.
3. Sistem yang Dibangun
   Bab ini berisi dataset spesifikasi, gambaran umum sistem, evaluasi dan implementasi sistem.
4. Evaluasi
   Bab ini menguraikan hasil pengujian terhadap sistem yang dibuat beserta hasil analisis dari pengujian yang telah dilakukan.
5. Kesimpulan
   Bab ini berisi kesimpulan dari hasil pengujian dan skenario yang telah dilakukan.
6. Lampiran
   Bab ini berisi lampiran *screenshot* tampilan sistem

2. Studi Terkait

Gambaran umum dari sistem klasifikasi yang dibangun dalam penelitian ini dibuat dalam diagram alur (flowchart).



Gambar 2.1 Flowchart Sistem

Umumnya data microarray adalah data yang memiliki dimensi yang besar. Hal ini mempengaruhi proses klasifikasi nantinya dan akan menyebabkan beban perhitungan menjadi tidak optimal sehingga akurasi nya nanti menghasilkan nilai yang sangat kecil atau tidak optimal. Selain itu data microarray juga memiliki skala (range) yang memiliki perbedaan pada setiap fiturnya. Untuk mengatasi masalah ini, maka perlu dilakukan normalisasi data (pada tahap *preprocessing)* yang membuat range nilai pada setiap fitur (atribut) data microarray berada pada range nilai 0 sampai 1 (bernilai biner). Setelah normalisasi data selesai, maka dilakukan seleksi fitur dengan *Genetic Algorithm* (GA).

1. Microarray

Micoarray adalah sebuah cara yang digunakan untuk menjelaskan fungsi dan ekspresi secara simultan dan dalam satu kali percobaan. Teknologi ini menggunakan chip yang berukuran sangat kecil yang terbuat dari lempengan kaca yang berisi ribuan bahkan puluhan ribu macam gen dalam bentuk fragmen DNA yang berasal dari penggandaan cDNA. Fragmen DNA yang memuat gen tersebut dapat mengenali gen dalam suatu sampel jaringan yang dianalisis. Pola ekspresi suatu gen dalam jaringan yang berbeda pun juga dapat diamati dengan menggunakan teknik ini. Prinsip kerja dari microarray adalah mengukur jumlah hibridisasi mRNA pada cDNA dalam chip tersebut. Pada umumnya analisis dengan menggunakan microarray dilakukan dengan menggunakan dua sampel yang berbeda, misalnya sel kulit normal dengan sel kanker kulit. Kedua sampel tersebut diisolasi mRNA-nya dan kemudian diletakkan dalam chip microarray. Kemudian chip tersebut diberi penanda radioaktif untuk menghasilkan warna fluorosens setelah dilakukan scanner yang terhubung dengan komputer. Kemudian komputer akan menganalisis kedua sampel tersebut berdasarkan pola warna yang ada [9]. Untuk efektifitas dan efisiensi deteksi kanker maka dilakukan simulasi klasifikasi data kanker dengan teknik komputasi menggunakan algoritma seleksi dan klasifikasi.

2. Seleksi Fitur

Seleksi Fitur atau *Feature Selection* adalah bagian dari metode reduksi dimensi yang mana merupakan sebuah proses yang biasa digunakan pada *Machine Learning* dimana sekumpulan dari fitur yang dimiliki oleh data digunakan untuk pembelajaran algoritma [10]. Seleksi fitur adalah suatu proses yang paling penting karena dapat mempengaruhi tingkat akurasi klasifikasi karena jika dataset berisi sejumlah fitur, dimensi dataset akan menjadi besar hal ini membuat rendahnya nilai akurasi klasifikasi. Masalah utama dalam seleksi fitur adalah pengurangan dimensi, dimana awalnya semua atribut diperlukan untuk memperoleh tingkat akurasi yang maksimal.

Tujuan utama dari *feature selection* adalah untuk memilih fitur yang berpengaruh (fitur optimal) dan mengesampingkan fitur yang tidak berpengaruh. Ada begitu banyak metode yang dapat digunakan untuk *feature seelction*. Pada Tugas Akhir ini penulis menggunakan teknik *Genetic Algorithm* (GA).

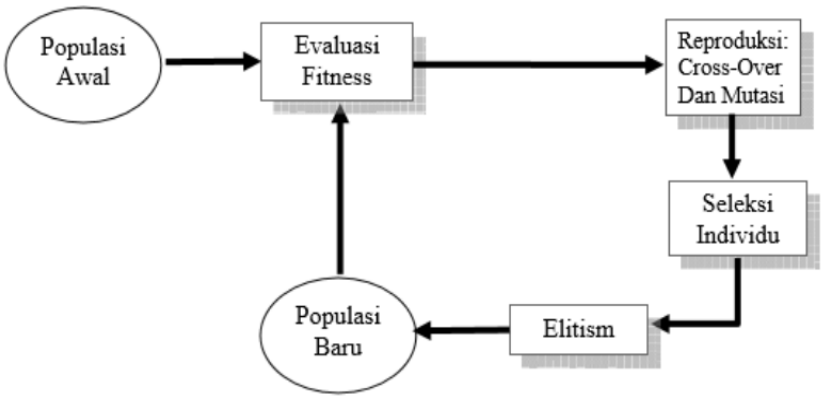
3. *Genetic Algorithm*

Genetic Algorithm (GA) salah satu stokastik modern yang umum digunakan. Seperti diketahui GA, pada dasarnya terinspirasi oleh evolusi alam dan seleksi [11]. Dimana GA dibangkitkan sebuah populasi yang terdiri beberapa individu dimana setiap individu mempresentasikan sebuah solusi. Setiap individu berisi beberapa kromosom, kromosom-kromosom tersebut terdiri dari beberapa gen.

Banyak penelitian yang telah ada menggunakan GA untuk menyelesaikan studi kasus yang dikerjakan. GA sendiri memiliki bentuk sederhana yang disebut dengan *Simple Genetic Algorithm* (SGA). SGA digunakan untuk menyelesaikan masalah optimasi diskret, ciri utamanya adalah tidak terlalu cepat dalam menemukan solusi optimalnya, tetapi memiliki *heuristic* yang baik untuk masalah kombinatorialnya[12].

3.1 Siklus *Genetic Algorithm*

Berikut adalah proses *Genetic Algorithm* yang telah diperbarui oleh Michalewicz dengan menambahkan operator elitism dan membalik proses seleksi setelah proses reproduksi



Gambar 2.2 Alur Genetic Algorithm oleh Michalewicz

3.2 Populasi Awal

Membangkitkan populasi awal adalah proses membangkitkan sejumlah individu secara acak atau melalui prosedur tertentu. Ukuran untuk populasi tergantung pada masalah yang akan diselesaikan dan jenis operator genetika yang akan diimplementasikan. Setelah ukuran populasi ditentukan, kemudian dilakukan pembangkitan populasi awal. Teknik dalam pembangkitan populasi awal dalam penelitian ini yaitu dengan cara random generator. Random generator merupakan cara yang melibatkan pembangkit bilangan random untuk nilai setiap gen dengan representasikan kromosom yang digunakan.

3.3 Nilai Fitnes

Nilai fitness adalah nilai yang menyatakan baik tidaknya suatu solusi (individu). Nilai fitness ini yang dijadikan acuan dalam mencapai nilai optimal dalam GA. GA bertujuan mencari individu dengan nilai fitness yang paling tinggi [13].

Kemudian untuk mempertahankan agar individu terbaik yang memiliki nilai fitness tertinggi tidak rusak karena proses crossover ataupun hilang selama proses evolusi, perlu dibuat satu atau dua salinannya, atau yang disebuit sebagai elitism. Prosedur ini hanya digunakan pada GA yang berjenis Generational Replacement. Semua individu yang telah dievaluasi akan diurutkan dari nilai yang paling tinggi dan selanjutnya disimpan sebagai calon solusi untuk generasi berikutnya. Sehingga, populasi baru yang dihasilkan selalu memiliki salah satu individu yang kualitasnya sama atau lebih baik dibandingkan individu pada generasi sebelumnya [13].

3.4 Seleksi

Seleksi digunakan untuk memilih individu-individu mana saja yang akan dipilih untuk proses kawin silang dan mutasi. Seleksi digunakan untuk mendapatkan calon induk yang baik. "induk yang baik akan menghasilkan keturunan yang baik". Semakin tinggi nilai fitness suatu individu semakin besar kemungkinannya untuk terpilih. Langkah pertama yang dilakukan dalam seleksi adalah nilai fitness. Nilai fitness ini yang nantinya akan digunakan pada tahap-tahap seleksi berikutnya. Masing-masing individu dalam wadah seleksi akan menerima probabilitas reproduksi yang tergantung pada nilai obyektif dirinya sendiri terhadap nilai obyektif dari semua individu dalam wadah seleksi tersebut.

3.5 Pindah Silang(crossover)

Prinsip dari pindah silang ini adalah melakukan operasi (pertukaran, aritmetika) pada gen-gen yang bersesuaian dari dua induk untuk menghasilkan individu baru. Proses crossover dilakukan pada setiap individu dengan probabilitas crossover yang ditentukan. Operator crossover ini bergantung pada representasi kromosom yang dilakukan, terdapat beberapa model crossover yaitu crossover satu titik, banyak titik, dan aritmatika [13].

3.6 Mutasi

Mutasi pada *Genetic Algorithm* (GA) berperan untuk menggantikan gen yang hilang dari populasi akibat proses seleksi yang memungkinkan munculnya kembali gen yang tidak muncul pada inisialisasi populasi. Kromosom anak dimutasi dengan menambahkan nilai random yang sangat kecil (ukuran langkah mutasi), dengan probabilitas yang rendah. Mutasi ini menggunakan probabilitas mutasi (Pm) dimana Pm didefinisikan sebagai presentasi dari jumlah total gen pada populasi yang mengalami mutasi. Mutasi memiliki beberapa metode yang sesuai dengan representasi kromosom yang digunakan. Salah satunya adalah mutasi biner, yaitu mengganti satu atau beberapa nilai gen dari kromosom [3].

Langkah-langkah pada proses mutasi yaitu:

1. Hitung jumlah gen pada populasi.
2. Pilih secara acak gen yang akan dimutasi.
3. Tentukan kromosom dari gen yang terpilih untuk dimutasi
4. Ganti nilai gen (0 ke 1, 1 ke 0) dari kromosom yang akan dimutasi tersebut.

3.7 Seleksi Survivor

Seleksi Survivor dapat juga disebut sebagai pengganti populasi. Terdapat dua model populasi yang bisa digunakan, yaitu *generational replacement* dan *steady state*. *Generational replacement* yaitu mengganti seluruh kromosom lama dengan kromosom baru setelah proses evolusi, sedangkan *steady state* hanya mengganti kromosom tertentu [12].

4. *K-Nearest Neighbor*

K-Nearest Neighbor (KNN) merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran. Nilai k yang terbaik untuk algoritma ini tergantung pada data, secara umumnya, nilai k yang tinggi akan mengurangi efek noise pada klasifikasi, tetapi membuat batasan antar setiap klasifikasi menjadi lebih kabur. KNN adalah metode klasifikasi non-parametrik yang sederhana tetapi efektif dalam banyak kasus. Untuk menerapkan KNN kita perlu memilih nilai k yang sesuai dan keberhasilan klasifikasi bergantung pada nilai ini. Ada banyak cara untuk memilih nilai k, dengan menjalankan banyak algoritma kali dengan nilai k berbeda dan pilih satu dengan kinerja terbaik. Banyak peneliti yang melakukan pengembangan tentang perbaikan KNN, baik dalam memperbaiki nilai akurasi KNN maupun dalam hal optimasi nilai k pada KNN.

5. Pengukuran Performansi

Confusion matrix merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya *confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya. Adapun tabel *confussion matrix* antara lain:

Tabel 3.3 Confussion Matrix [8]

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative (TN)	False Positif (FP)
	Positive	False Negative (FN)	True Positif (TP)

TP adalah *True Positive*, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem. TN adalah *True Negative*, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem. FN adalah *False Negative*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem. FP adalah *False Positive*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem. Akurasi dapat diukur dengan persamaan sebagai berikut:

$$\text{Akurasi} = \frac{1}{\text{fitness}}$$

3. Sistem yang Dibangun

Berikut penjelasan proses tahapan dari rancangan sistem yang akan dibangun, yaitu:

3.1.1 Normalisasi

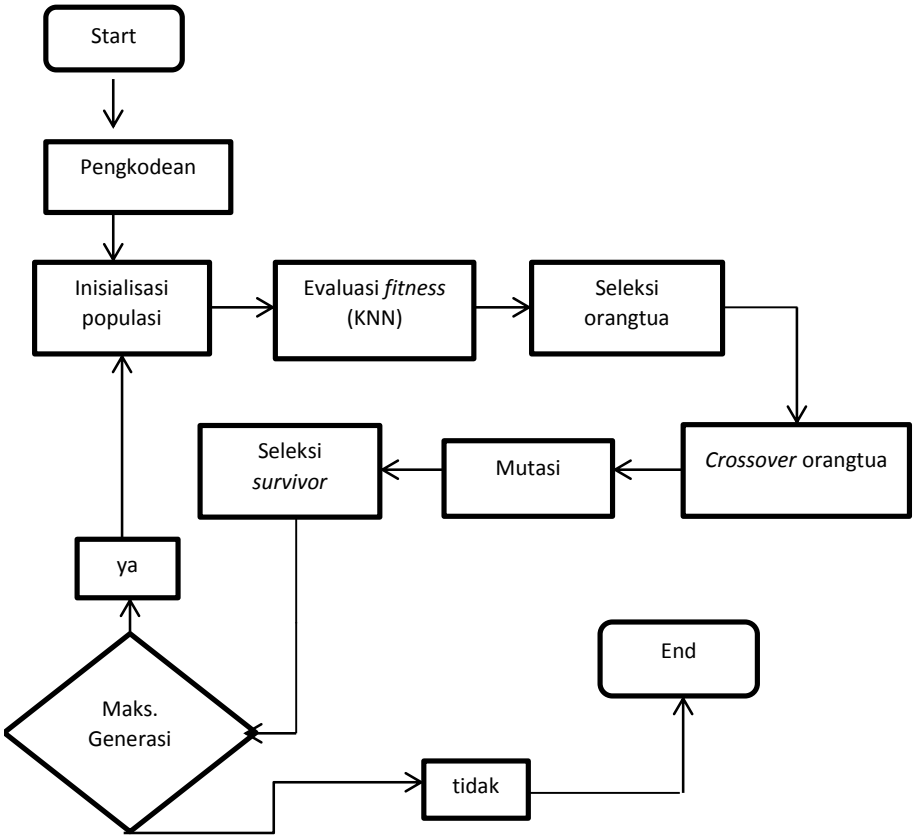
Setiap data kanker yang akan digunakan dalam penelitian ini memiliki perbedaan spesifikasi range nilai yang cukup signifikan. Oleh karena itu, diperlukan normalisasi data sehingga skala (range) nilai pada setiap data kanker berada pada range 0 sampai 1. Dibawah ini merupakan rumus umum untuk normalisasi data,

$$\text{Normalisasi} = \frac{data - \min(data)}{\max(data) - \min(data)} \quad (3.1)$$

Proses normalisasi tersebut dapat membuat range nilai pada dataset memiliki range nilai pada dataset memiliki range nilai yang seragam antara 0 sampai 1, sehingga kompleksitas data pada saat data digunakan sebagai masukan (input) akan berkurang.

3.1.2 Seleksi Fitur *Genetic Algorithm* (GA)

Setelah tahap normalisasi data dilakukan, selanjutnya tahap berikutnya adalah melakukan seleksi fitur. Adapun langkah-langkah dan gambaran skema umum dari proses seleksi fitur dengan *Genetic Algorithm* (GA) adalah sebagai berikut (Gambar 3.2)



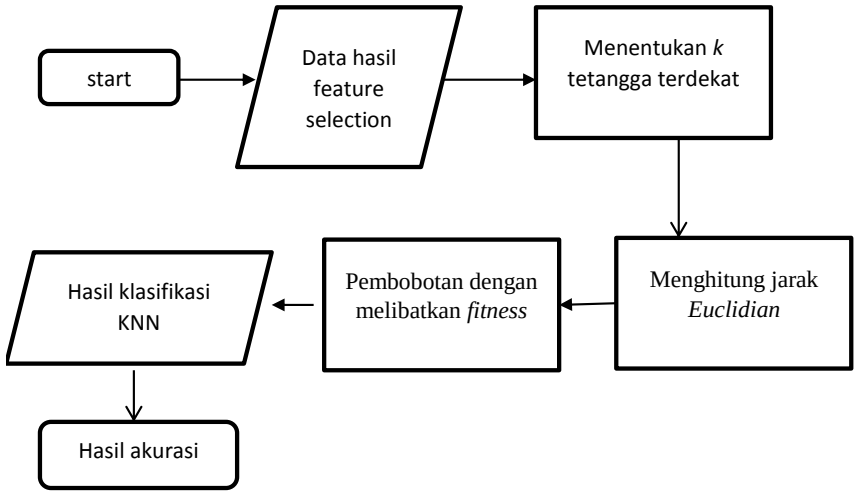
Gambar 3.2 Alur seleksi fitur dengan seleksi *Genetic Algorithm* (GA)

Proses GA berhenti jika kriteria terminasi yang ditetapkan di awal sudah terpenuhi. Iterasi GA akan terus berjalan jika kriteria terminasi belum terpenuhi. Berikut akan disajikan parameter dalam mengimplementasikan GA sebagai seleksi fitur dengan 10 dan 100 individu yang akan dianalisis terdapat pada Tabel 3.2 [17].

Tabel 3.2 Parameter GA Proses Seleksi

Parameter GA	**Nilai**
Ukuran Populasi	10 dan 100
Maksimum Generasi	3
Skema Pengkodean	*Binary Encoding*
Fungsi *Fitness*	Akurasi KNN
Crossover	*Single Point Crossover*
Peluang *Crossover*	0.8
Mutasi	*Flip Bit Mutation*
Peluang Mutasi	0.1
Mekanisme Seleksi	*Genetic Algorithm*
Seleksi Survivor	*Generational Replacement*

3.1.4 Klasifikasi KNN



Gambar 3.3 Alur Klasifikasi KNN

Setelah melakukan reduksi dimensi dengan *feature selection* pada data *microarray*, selanjutnya adalah melakukan proses klasifikasi. Klasifikasi ini nantinya akan menghasilkan data mana yang termasuk kanker dan mana yang bukan kanker berdasarkan data microarray. Dalam tahap klasifikasi penulis menggunakan *K Nearest Neighbor* (KNN). Adapun tahapan langkah algoritma KNN sebagai berikut:

1. Menentukan parameter *k* tetangga terdekat.
2. Menentukan nilai korelasi dari *k* tetangga terdekat
3. Menghitung kuadrat jarak *Eucliden* objek terhadap data training, dengan X dan Y masing-masing adalah data uji dan data latih.

$$d(X,Y) = \sqrt{\sum_{i=I}^{n}(X_i - Y_i)^2} \quad (3.2)$$

4. Menghitung bobot dengan mempertimbangkan *fitness* proses seleksi, dengan *U* adalah koefisien pemulusan (*smoothing*)

$$W(i) = \textit{fitness}\ (i)\ x\ \frac{1}{d(i)+U} \quad (3.3)$$

5. Nilai bobot terbesar adalah prediksi kelas dari data uji.

4. Evaluasi

Berikut adalah hasil dan analisis yang dilakukan pada penelitian ini

Pada proses seleksi dengan GA digunakan skenario testing-training menggunakan "HoldOut" dengan memilih secara acak data training sebesar 50% dari dimensi data dan max generasi 5 . Sehingga untuk data training colon tumor sebesar 31 data, leukimia sebesar 19 data dan lung sebesar 90 data. Selanjutnya parameter yang digunakan pada algoritma GA disesuaikan dengan Tabel 3.2 . Diperoleh hasil akurasi untuk masing-masing data sebagai berikut

Tabel 4.1 Akurasi GA dan KNN

Data	Akurasi
Colon Tumor	93,02%
Leukimia	96%
Lung	96,03%
Rata-Rata	95,01%

Hasil akurasi terbesar dengan skenario diatas adalah 96,03% dengan rata-rata akurasi keseluruhan adalah 95,01%.

Selanjutnya proses seleksi dengan GA digunakan skenario testing-training menggunakan "HoldOut" dengan memilih secara acak data training sebesar 70% dari dimensi data dan max generasi 3. Sehingga untuk data training colon tumor sebesar 43 data, leukimia sebesar 27 data dan lung sebesar 127 data. Diperoleh hasil akurasi untuk masing-masing data sebagai berikut

Tabel 4.2 Akurasi GA dan KNN

Data	Akurasi
Colon Tumor	90,7%
Leukimia	92%
Lung	97,62%
Rata-Rata	93,44%

Hasil akurasi terbesar dengan skenario diatas adalah 97,62% dengan rata-rata akurasi data keseluruhan adalah 93,44%.

5. Kesimpulan

Microarray adalah teknik modern yang memfasilitasi analisis simulasi dari sejumlah data yang menggambarkan ekspresi gen yang diperlukan untuk memecahkan masalah biologis yang kompleks, seperti deteksi suatu penyakit tertentu. Data microarray memiliki karakteristik berdimensi besar, dimana banyaknya variabel respon lebih kecil dibandingkan variabel prediktor. Oleh karena itu, diperlukan skema yang didalamnya terdapat proses reduksi dimensi dan proses klasifikasi. Dalam hal ini, proses reduksi dimensi bertujuan untuk meringankan beban komputasi serta menghindari overfitting pada klasifikasi. Proses reduksi yang digunakan pada penelitian ini yaitu seleksi fitur *Genetic Algorithm* (GA). Kemudian, proses klasifikasi yang bertujuan untuk mendeteksi penyakit kanker atau bukan kanker dilakukan dengan menggunakan metode klasifikasi K-Nearest Neighbor (KNN). Adapun akurasi dari metode GA-KNN pada data tumor usus sebesar 93,02%, data kanker paru-paru sebesar 96,03% dan kanker darah sebesar 96%, dengan rata-rata akurasi data keseluruhan adalah 95,01%.

Daftar Pustaka

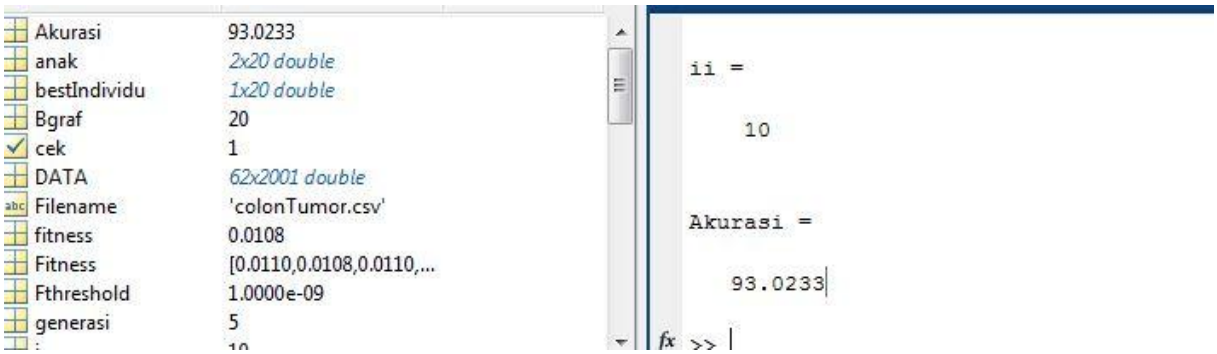
[1] *International Agency for Reasearch on Cancer*. WHO. https://www.who.int/cancer/PRGlobocanFinal.pdf. 2018.
[2] Ratna Nitin Patil, Dr Sharvari Chandrasekhar Tamane,. 2018. *Upgrading the Performance* of KNN *and Naïve Bayes in Diabetes Detection with Genetic Algorithm for Feature Selection. International Journal of Scientific Research in Computer Science, Engineering and Information Technology*.
[3] M Poornima, C Kuyin, M Revathy, 2018. *Brain Tumor Detection Using Genetic Algorithm, International Journal of Scientific Research in Computer Science, Engineering and Information Technology*.
[4] Pankaj Goel, D. K. Lobiyal, 2018 *Minimizing the cost of two-tier cellular network with queuing handoff calls in microcell using genetic algorithm. Malaya Journal of Matematik*, Vol. S, No. 1, 14-21, 2018
[5] Liton C. P., Abdulla A. S., Nahid Sultan. 2013. *Methodological Analysis of Principal Component Analysis (PCA) Method. IJCEM International Journal of Computational Engineering & Management*, Vol. 16 Issue 2, March 2013 ISSN (Online): 2230-7893 www.IJCEM.org
[6] Cristinel Constantin, 2014. *Principal Component Analysis - A Powerfull Tool In Computing Marketing Information.* Bulletin of the Transilvania University of Braşov Series V: Economic Sciences • Vol. 7 (56) No. 2
[7] Zhiliang W., Yalin S., Peng Li, 2014. *Functional Principal Components Analysis of Shanghai Stock Exchange 50 Index. Hindawi Publishing Corporation Discrete Dynamics in Nature and Society* Volume 2014, Article ID 365204, 7 pages http://dx.doi.org/10.1155/2014/365204
[8] Ramadhani, P., T., 2017. Deteksi Kanker berdasarkan Klasifikasi Data Microarray menggunakan *Functional Link Neural Network* dengan Seleksi Fitur Genetic Algorithm. Indo-JC, Vol.2:13.
[9] Generasi Biologi, 2016. Microarray:Biologi di Era Pascagenomik. http://www.generasibiologi.com/2012/08/microarray-biologi-di-era-pascagenomik.html.
[11] J. H. Holland, *Adaptation in natural artificial systems*. 2nd edition, MIT Press (1992).
[12] Suyanto, S. M. (2008*). Evolutionary Computing*. Bandung: Informatika
[13] Etin, "Kecerdasan Buatan: Bab & Algoritma Genetika," [Online]. Available:http://lecturer.eepisits.edu/~entin/Kecerdasan%20Buatan/Buku/Bab%207 %20 Algoritma%20Genetika.pdf.
[14] Galih Hendro M, T. B. Adji, N. A. Setiawan,2012, penggunaan metodologi analisa komponen utama(PCA) untuk mereduksi faktor-faktor yang mempengaruhi penyakit jantung koroner. Seminar nasional "*Science, Engineering and Technology*".
[15] Tyang Luhtu, 2013, langkah umum *principal component analysis* , https://tyangluhtu.wordpress.com/2013/04/19/langkah-umum-principal-component-analysis/

[16] Siti Mutrofin, Abidatul Izzah, Arrie Kurniawardhani, Mukhamad Masrur, 2014, optimasi trknik klasifikasi *modified K Nearest Neighbor* menggunakan algoritma genetika, sistem informatika, Universitas Darul Ulum.
[17] Milah Sarmilah, 2018, Analisis Seleksi Fitur *Genethic Algorithm* Dan Ekstraksi Fitur *Wavelet* Pada Klasifikasi Microarray Data Menggunakan Naïve Bayes, Universitas Telkom.
[18] Nurfalah, A. Adiwijaya, and Suryani, A.A., (2016). Cancer Detection Based On Microarray Data Classification Using PCA And Modified Back Propagation. Far East Journal of Electronics and Communications, 16(2), p.269.
[19] Husna Aydadenta, Adiwijaya, (2018), A Clustering Approach for Feature Selection in Microarray Data Classification using Random Forest, Journal of Information Processing System 14(5)
[20] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, D. S. Kusumo, (2018). Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification, Journal of Computer Science 14(11)
[21] Astuti, Widi, and Adiwijaya Adiwijaya. "Principal Component Analysis Sebagai Ekstraksi Fitur Data Microarray Untuk Deteksi Kanker Berbasis Linear Discriminant Analysis." JURNAL MEDIA INFORMATIKA BUDIDARMA 3.2 (2019): 72-77.
[22] Adiwijaya, A. (2018). Deteksi Kanker Berdasarkan Klasifikasi Microarray Data. JURNAL MEDIA INFORMATIKA BUDIDARMA, 2(4), 181-186.

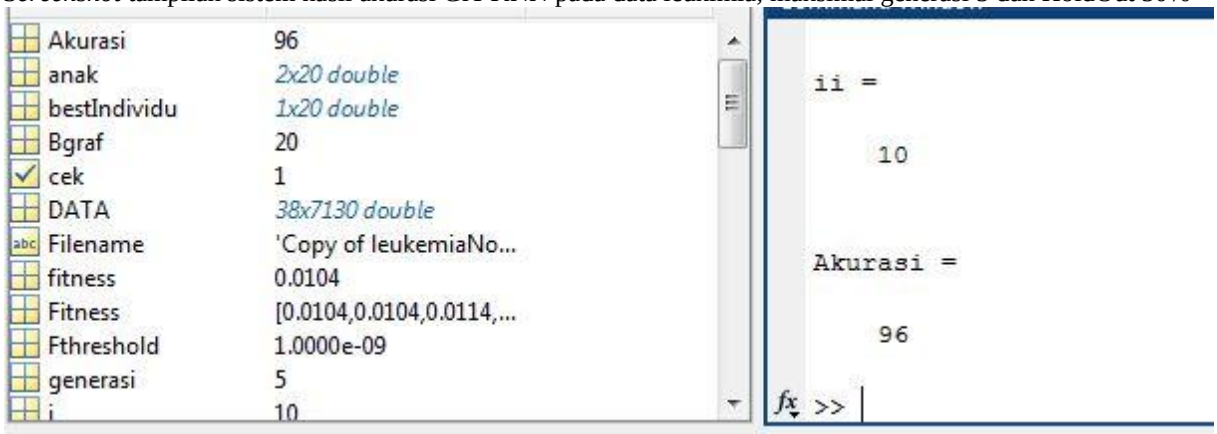
Lampiran

Berikut adalah beberapa lampiran hasil pengujian yang dilakukan pada penelitian ini

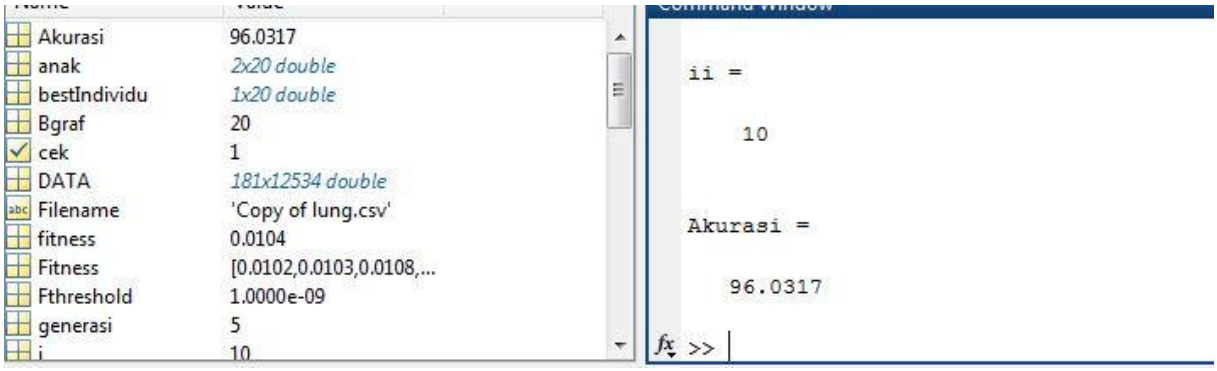
Lampiran 1
Screenshot tampilan sistem hasil akurasi GA-KNN pada data colon tumor, maksimal generasi 5 dan HoldOut 50%



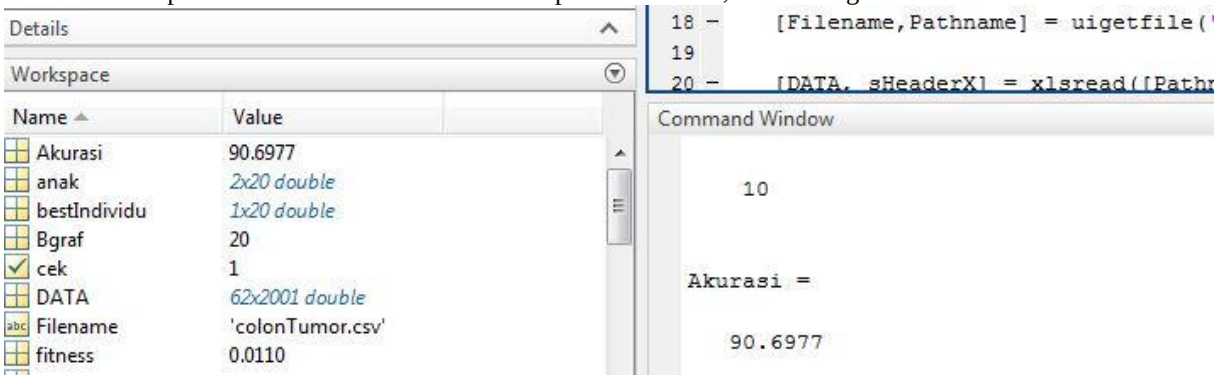
Lampiran 2
Screenshot tampilan sistem hasil akurasi GA-KNN pada data leukimia, maksimal generasi 5 dan HoldOut 50%



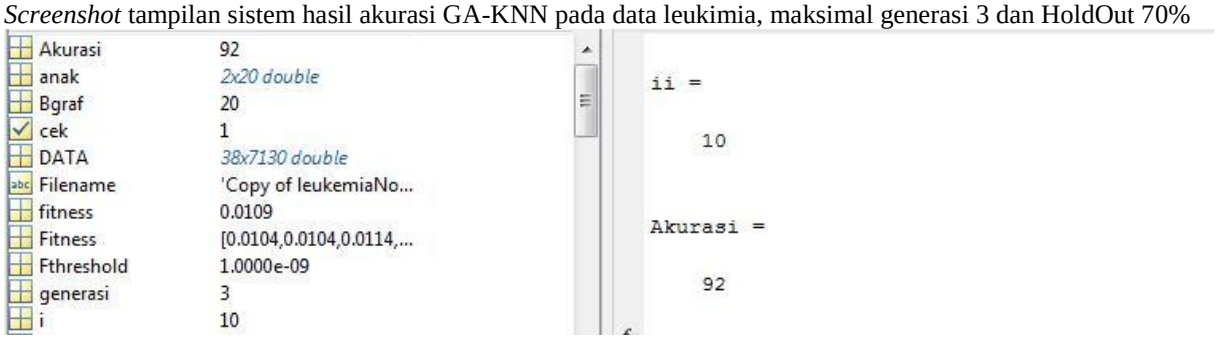
Lampiran 3
Screenshot tampilan sistem hasil akurasi GA-KNN pada data lung, maksimal generasi 5 dan HoldOut 50%



Lampiran 4
Screenshot tampilan sistem hasil akurasi GA-KNN pada data colon, maksimal generasi 3 dan HoldOut 70%



Lampiran 5
Screenshot tampilan sistem hasil akurasi GA-KNN pada data leukimia, maksimal generasi 3 dan HoldOut 70%



Lampiran 6
Screenshot tampilan sistem hasil akurasi GA-KNN pada data lung, maksimal generasi 3 dan HoldOut 70%