Klasifikasi Data *Microarray* Menggunakan *Genetic Algorithm* (GA), *Naive Bayes* dan Regresi Logistik

Ergon Rizky Perdana Purba[1], Adiwijaya[2], Aniq Atiqi R[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung

[1]ergonpurba@student.telkomuniversity.ac.id,

[2]adiwijaya@telkomuniversity.ac.id, [3]aniqatiqi@telkomuniversity.ac.id

Abstrak

Kanker merupakan salah satu penyakit yang mematikan di dunia. Setiap tahunnya, penderita kanker terus meningkat dan banyak menelan korban jiwa. Hingga sampai saat ini, obat untuk penyakit yang mematikan ini masih sulit ditemukan. Dalam beberapa tahun terakhir, teknologi data *microarray* banyak digunakan untuk mendiagnosa kanker sejak dini, data DNA *microarray* adalah teknologi yang digunakan untuk melihat urutan sekuens asam nukleat yang berada pada lokasi tertentu pada struktur DNA yang dapat digunakan untuk menganalisa ribuan sampel pada waktu yang bersamaan sehingga nantinya dapat diklasifikasikan mana yang tergolong kanker dan bukan kanker. Oleh karena itu, data *microarray* adalah data yang memiliki ukuran dimensi data yang sangat besar. Data yang ukuran dimensinya sangat besar dapat mengakibatkan hasil perhitungan menjadi tidak optimal dan akurasi klasifikasi yang dihasilkan kecil. Untuk mengoptimalkan data dan meningkatkan nilai akurasi klasifikasi dari data yang dimensinya besar tersebut, dilakukan pengurangan dimensi dengan seleksi fitur *Genetic Algorithm* (GA). *Genetic Algorithm* biasanya mampu memberikan hasil yang baik dan tingkat akurasi yang cukup baik. Klasifikasi data *microarray* menggunakan metode *Naive Bayes* dan Regresi Logistik. Adapun akurasi terbaik dari *Genetic Algorithm* dan klasifikasi Regresi Logistik 100% pada data *colon tumor* dan *mll leukemia*. *Genetic Algorithm* dan klasifikasi *Naive Bayes* 57,7778% pada data MLL Leukemia. Dan Regresi Logistik 67% pada data *mll leukemia.*

Kata Kunci : kanker, seleksi fitur, klasifikasi, data microarray, genetic algorithm, naive bayes, regresi logistik

Abstrac

Cancer is one of the deadliest diseases in the world. Every year, cancer patients continue to increase and cause many casualties. Until now, the cure for this deadly disease is still hard to find. In recent years, microarray data technology has been widely used to diagnose cancer from an early age, DNA microarray data is a technology used to see sequences of nucleic acid sequences located at specific locations in DNA structures that can be used to analyze thousands of samples at the same time so that later can be classified which are classified as cancer and not cancer. Therefore, microarray data is data that has very large data dimensions. Data whose dimensions are very large can result in the calculation results being not optimal and the resulting classification accuracy is small. To optimize the data and increase the classification accuracy value of the large dimension data, dimensional reduction was carried out by selecting the Genetic Algorithm (GA) feature. Genetic Algorithm is usually able to provide good results and a fairly good level of accuracy. Classification of microarray data using the Naive Bayes method and Logistic Regression. The best accuracy of Genetic Algorithm and Logistic Regression classification in colon tumor data and mll leukemia are 100% and 57,7778% for Genetic Algorithm and Naive Bayes classification. And 67% accuracy get from Logistic Regression in the mll leukemia data.

Keywords : cancer, feature selection, classification, microarray data, genetic algorithm, naive bayes, logistic regression

1. Pendahuluan

Kanker adalah istilah umum untuk satu kelompok besar penyakit yang dapat mempengaruhi setiap bagian dari tubuh. Istilah lain yang digunakan adalah tumor ganas dan neoplasma. Sel-sel kanker akan berkembang dengan cepat, tidak terkendali, dan terus membelah diri. Selanjutnya menyusup ke jaringan sekitarnya *(invasive)* dan terus menyebar melalui jaringan ikat, darah, serta menyerang organ-organ penting dan saraf tulang belakang [1].

Kanker disebut sebagai penyakit yang dapat menyebabkan kematian. Pada umumnya penderita kanker baru menyadari kanker tersebut setelah mengalami stadium lanjut. Keterlambatan dalam penanganan penyakit kanker dapat berakibat fatal pada hidup penderita. Oleh karena itu sangat perlunya penanganan yang cepat untuk mengatasi kanker tersebut.

Menurut data *International Agency for Research on Cancer (IARC)*, terdapat 14,1 juta kasus baru kanker dengan sekitar 8,2 juta penderita meninggal akibat kanker dan 32,6 juta penderita kanker yang hidup dalam 5 tahun terakhir pada tahun 2012 di seluruh dunia. Sedangkan pada tahun 2030 diprediksikan angka kejadian kanker meningkat menjadi 21,7 juta penderita. Penyebab kematian kanker yang paling umum di dunia adalah kanker paru-paru, diperkirakan sekitar 1,59 juta kematian atau 19,4 % dari total kematian penyebab kanker di seluruh dunia [2].

Penderita penyakit kanker dapat dicegah dengan cara mengubah faktor resiko perilaku dan pola makan penyebab penyakit kanker [3]. Kanker yang diketahui sejak dini memiliki kemungkinan untuk mendapatkan penanganan lebih baik. Namun untuk melakukan deteksi dini tersebut bukanlah hal yang mudah. Di zaman teknologi sekarang ini banyak peneliti yang melakukan penelitian di bidang medis dan bioinformatika untuk penanganan kanker guna mencegah berkembangnya penyakit kanker. Peneliti-peneliti tersebut menggunakan suatu teknologi yang bernama *microarray* untuk mendeteksi dan menganalisis penyakit kanker.

Latar Belakang

Microarray adalah teknologi yang mampu menyimpan ribuan ekspresi gen yang diambil dari beberapa jaringan tertentu manusia sekaligus, di dalamnya memiliki potensi yang sangat besar untuk pengetahuan baru, yang mendasari kemajuan dalam fungsional genomik dan biologi molekuler [4]. Teknologi *microarray* ini biasa digunakan untuk klasifikasi penyakit termasuk penyakit kanker. Klasifikasi ini dilakukan dengan cara pengelompokkan sampel yang akan diuji sehingga nantinya bisa di analisis dan digolongkan apakah sampel yang telah diuji tersebut tergolong dalam jenis kanker atau bukan kanker. Klasifikasi data *microarray* membutuhkan proses yang bertahap. Dikarenakan dalam data *microarray* pasti mempunyai data yang berdimensi besar. Jika tiak dilakukan pengurangan dimensi maka hal ini akan mengakibatkan kurang efektif dan efisiennya pengolahan data pada saat dilakukan analisis sehingga hal ini menyebabkan beban komputasi menjadi tidak optimal atau tidak stabil [5].

Oleh karena itu, untuk mengolah data *Microarray* diperlukan seleksi fitur atau biasa disebut pengurangan dimensi untuk meringankan beban perhitungan sehingga nantinya didapat hasil yang stabil dan optimal pada saat dilakukan proses klasifikasi. Seleksi fitur yang digunakan adalah *Genetic Algorithm* (GA). Penulis menggunakan seleksi fitur ini karena dari beberapa penelitian yang sudah dilakukan, metode tersebut terbukti menghasilkan tingkat akurasi yang tinggi [6]. Setelah dilakukan seleksi fitur, maka dilakukan proses klasifikasi. Klasifikasi ini bertujuan untuk menentukan data kanker atu bukan kanker. Penulis menggunakan *classifier Naive Bayes* dan Regresi Logistik untuk proses klasifikasi. Menurut [6], terdapat peneliti yang sudah melakukan penelitian menggunakan *classifer Naive Bayes* dan Regresi Logistik dan didapatkan hasil akurasi yang baik daripada *classifer* lainnya.

Topik dan Batasannya

Berdasarkan masalah yang telah disampaikan di atas, maka dapat ditarik beberapa rumusan masalah, yaitu:

1. Bagaimana implementasi dari seleksi fitur atau pengurangan dimensi berdasarkan data *microarray* dengan menggunakan *Genetic Algorithm* (GA)?
2. Bagaimana proses klasifikasi data *microarray* dengan menggunakan *Naive Bayes* dan Regresi Logistik?
3. Bagaimana hasil performansi yang didapatkan dari hasil seleksi fitur dan metode klasifikasi yang dibangun berdasarkan data *microarray*?

Batasan-batasan masalah dalam pembuatan Tugas Akhir ini adalah sebagai berikut:
Data yang digunakan dalam Tugas Akhir ini adalah data *microarray* yang diambil dari *Kent Ridge Bio-medical Dataset* yang dapat diakses melalui link website http://datam.i2r.a-star.edu.sg/datasets/krbd/.

Tujuan

Adapun tujuan dari penelitian Tugas Akhir ini adalah sebagai berikut :

1. Mengimplementasikan proses seleksi fitur dengan menggunakan *Genetic Algorithm* (GA).
2. Mampu mengimplementasikan proses klasifikasi data *microarray* dengan menggunakan *Naive Bayes* dan Regresi Logistik.
3. Menganalisa performansi dari hasil seleksi fitur menggunakan *Genetic Algorithm* (GA) dan metode klasifikasi dengan *Naive Bayes* dan Regresi Logistik berdasarkan data *microarray*.

Organisasi Tulisan

Untuk memudahkan pemahaman pembaca pada penelitian Tugas Akhir ini, maka penulis menyusun Tugas Akhir ini dengan sistematika penulisan. Pada bagian pendahuluan menjelaskan latar belakang dipilihnya permasalahan penelitian, rumusan masalah, batasan masalah, dan tujuan penelitian. Dilanjutkan dengan studi terkait yang menjelaskan tentang teori pendukung yang digunakan dalam pengerjaan Tugas Akhir ini. Landasan teori yang digunakan dalam pengerjaan Tugas Akhir ini adalah penjelasan mengenai *microarray*, seleksi fitur, *Genetic Algorithm*, *Naive Bayes* dan Regresi Logistik. Selanjutnya membuat dan menjelaskan gambaran sistem yang akan dianalisis termasuk pada dataset spesifikasi, implementasi sistem dan tahap-tahap perancangan dan pengerjaan sistem yang akan dibangun dalam penelitian Tugas Akhir ini. Kemudian pada bagian evaluasi dijelaskan mengenai proses pengujian terhadap sistem yang dibuat dan hasil analisis dari pengujian metode yang telah dilakukan. Selanjutnya pada bagian kesimpulan berisikan kesimpulan yang dapat diambil dari hasil pengujian dan analisis yang telah dilakukan serta saran untuk pengembangan penelitian kedepannya.

2. Studi Terkait

Terdapat beberapa penelitian terdahulu yang menjadi acuan dalam pengerjaan Tugas Akhir ini. Kumar et al. (2015) melakukan penelitian pada data *microarray* dengan menggunakan metode T-statistik untuk seleksi fitur dan FLNN untuk klasifikasi. Pada penelitian tersebut, Kumar et al mencatat setiap akurasi yang didapat yaitu 96.15%, 97.78% dan 86.54%. Diaz-Uriarte et al. (2006) juga melakukan penelitian pada data *microarray* dengan menggunakan metode Random Forest pada seleksi fitur dan klasifikasinya. Penelitiannya mampu menghasilkan akurasi 95%. Pada penelitian yang dilakukan oleh C.D.A Vanita et al. (2015), menghasilkan akurasi di bawah 70% yang mana penelitian yang dia lakukan menggunakan metode *Mutual Information* untuk seleksi fitur dan beberapa metode untuk klasifikasinya yaitu KNN, ANN, SVM Linear, SVM RBF, SVM Quadratic, dan SVM Polynomial. Khare et al. (2016) melakukan penelitian dengan beberapa metode klasifikasi yaitu Bayesnet, SMO, *Simple Logistic*, One-R dan Zero-R sedangkan untuk seleksi fitur, peneliti memilih metode *Genetic Algorithm* dan masing-masing metode yang digunakan mampu mencapai akurasi di atas 75% [6]. Nugroho, Dwi (2016) melakukan penelitian mengenai prediksi penyakit menggunakan *Genetic Algorithm* (GA) dan *Naive Bayes.* Penelitian tersebut menghasilkan akurasi yang sangat baik sebesar 94.74% dan 100% [23]. Penerapan metode *Naive Bayes* juga sudah pernah diterapkan pada penelitian lain yaitu klasifikasi daun herbal. Hasil penelitian tersebut menunjukkan bahwa kinerja *Naive Bayes* mampu menghasilkan akurasi yang baik sebesar 75% [24]. Metode *Naive Bayes* juga pernah digunakan dalam mengklasifikasikan tebu. Pada penelitian tersebut *Naive Bayes* mampu menghasilkan persentase kinerja sebesar 73.3% [25]. Penelitian lain pada klasifikasi status angkatan kerja dengan menggunakan metode Regresi Logistik mampu menghasilkan akurasi sebesar 96.4% [26]. Klasifikasi dengan *Naive Bayes* dan Regresi Logistik juga pernah dilakukan oleh Rajagukguk, Nanci (2015) untuk mengklasifikasikan status pengguna KB di Kota Tegal pada Tahun 2014. Penelitian tersebut menghasilkan persentase kinerja yang cukup baik, yaitu 81.75% untuk *Naive Bayes* dan 83.33% untuk Regresi Logistik [27].

2.1 *Microarray*

Microarray adalah chip yang berukuran kecil yang terbuat dari lempengan kaca yang berisi ribuan bahkan puluhan ribu macam gen dalam bentuk fragmen DNA yang berasal dari penggandaaan 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 [7].

Tujuan dari klasifikasi pada data *microarray* adalah untuk menganalisis mana yang tergolong kanker dan mana yang bukan kanker. Dengan mengklasifikasikan data *microarray*, teknologi ini diharapkan mampu memberikan performa kinerja yang baik dan akurat dengan banyak metode pengklasifikasian sehingga nantinya metode yang kinerjanya mampu memberikan performa yang baik, dapat dikembangkan guna mendeteksi dini penyakit kanker.

2.2 *Feature Selection*

Feature Selection atau seleksi fitur 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 [8]. 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 pada saat klasifikasi. Masalah utama dalam seleksi fitur adalah pengurangan dimensi, dimana awalnya semua atribut diperlukan untuk memperoleh tingkat akurasi yang maksimal.

Tujuan utama dari seleksi fitur adalah untuk memilih fitur yang berpengaruh dan mengesampingkan fitur yang tidak berpengaruh. Ada begitu banyak metode yang dapat digunakan untuk seleksi fitur. Salah satu teknik yang digunakan penulis pada Tugas Akhir ini adalah *Genetic Algorithm* (GA).

2.3 Normalisasi Data

Data kanker yang akan digunakan dalam penelitian ini memiliki skala *(range)* yang berbeda-beda. Oleh karena itu tahap awal yang harus dilakukan terhadap data *microarray* adalah tahap *preprocessing. Preprocessing* yang dilakukan adalah normalisasi data pada data *microarray*. Untuk normalisasi data *microarray* digunakan teknik Min-Max Normalization dengan persamaan sebagai berikut.

$$Normalisasi\ (x) = \frac{x - Min(Xi)}{Max(Xi) - Min(Xi)} \quad (1)$$

Setelah dilakukan normalisasi data, maka akan diperoleh nilai masukan *(input)* data yang berada dalam jangkauan 0 sampai 1.

2.4 Seleksi Fitur *Genetic Algorithm* (GA)

Setelah melalui tahap normalisasi data *(preprocessing)*, maka tahap selanjutnya adalah melakukan seleksi fitur menggunakan *Genetic Algorithm* (GA). Seleksi fitur dilakukan untuk mereduksi data yang berdimensi tinggi agar menghindari masalah *curse of dimensionality*. Sebelum memasuki proses GA, jumlah data latih dan data uji harus sudah ditetapkan. Setelah itu dilakukan tahap-tahap GA untuk seleksi fitur [9]. Adapun langkah-langkah GA adalah sebagai berikut.

1) Membangkitkan Populasi Awal: Pada tahapan ini setiap individu direpresentasikan pada deretan bilangan biner 0 atau 1. Ukuran untuk populasi tergantung pada masalah yang akan diselesaikan. Setelah ukuran populasi diinisialisasi akan dilakukan pembangkitan populasi secara acak sebanyak jumlah fitur dan ukuran populasi. Jika jumlah bit sama dengan 0 artinya fitur tersebut tidak dipilih sebaliknya apabila bit sama dengan 1 maka fitur tersebut dipilih.

2) Evaluasi Nilai Fitness: Nilai *fitness* dijadikan sebagai acuan dalam mencapai nilai optimal dalam *Genetic Algorithm* (GA). Kelas klasifikasi yang dihasilkan akan dihitung dengan metode *Confussion Matrix* dalam persamaan 2 berikut.

$$Accuracy\ (x) = \frac{TP + TN}{TP + FN + FP + TN} * 100\% \quad (2)$$

Keterangan :

1) *True Positive* (TP) merupakan *positive class* yang terklasifikasikan dengan benar oleh sistem klasifikasi.

2) *True Negative* (TN) merupakan *negative class* yang terklasifikasikan dengan benar oleh sistem klasifikasi.

3) *False Positive* (FP) merupakan *negative class* yang terklasifikasikan oleh sistem klasifikasi sebagai *positive class.*

4) *False Negative* (FN) merupakan *positive class* yang terklasifikasikan oleh sistem klasifikasi sebagai *negative class.*

Tabel 1 *Confussion Matrix*

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

Berdasarkan Tabel 1, TP *(True Positive)* merupakan proporsi positif dalam data set yang diklasifikasikan positif. TN *(True Negative)* merupakan proporsi negatif dalam data set yang diklasifikasikan negatif. FN *(False Negative)* merupakan proporsi positif dalam data set yang diklasifikasikan negatif. FP *(False Positive)* merupakan proporsi negatif dalam data set yang diklasifikasikan positif [5]. Untuk mempertahankan agar individu terbaik yang memiliki nilai *fitness* tertinggi tidak rusak karena proses *crossover* ataupun hilang selama proses evolusi, maka perlu dibuat salinannya sebagai *Elitism.*

3) Seleksi Dengan Metode Roulette Wheel: Umumnya seleksi dengan metode ini disebut dengan seleksi orangtua. Masing-masing individu akan diseleksi menjadi orangtua dengan menempatkan setiap kromosom individu pada *Roulette Wheel* sesuai dengan proporsi nilai *fitness* yang dimiliki masing-masing individu. Semakin besar nilai *fitness* suatu kromosom maka proporsi yang dimilikinya dalam *Roulette Wheel* akan semakin besar pula sehingga peluang individu tersebut terpilih menjadi orangtua juga semakin besar.

4) Pindah Silang (Crossover): Kemudian dilakukan pindah silang atau *crossover* terhadap kromosom orangtua yang sudah terpilih sehingga nantinya menghasilkan kromosom *offspring* (anak). Proses *crossover* dilakukan pada setiap individu dengan probabilitas *crossover* yang sudah ditentukan. Pada penelitian ini dilakukan *Crossover Single Point* dimana panjang kromosom diseleksi secara acak dan terjadi penukaran variabel antar kromosom pada *single* atau titik tersebut untuk menghasilkan kromosom *offspring* (anak).

5) Mutasi: Mutasi dilakukan dengan membangkitkan kromosom *offspring* (anak) secara acak dalam bilangan biner dengan syarat probabilitas mutasi tertentu [5]. Kromosom anak dimutasi dengan menambahkan nilai acak yang sangat kecil dengan probabilitas yang rendah. Mutasi menggunakan probabilitas mutasi *(Pm). Pm* direpresentasikan sebagai presentasi dari jumlah total gen pada populasi yang mengalami mutasi.

6) Seleksi Survivor: Pada seleksi ini digunakan model populasi *Generational Replacement* yaitu mengganti seluruh kromosom lama dengan kromosom baru setelah proses evolusi atau hasil dari *crossover* dan mutasi, juga serta kromosom terbaik yang telah disimpan dalam *Elitism.* Terdapat juga kriteria terminasi pada metode *Genetic Algorithm* yaitu maksimum iterasi dalam *Genetic Algorithm* dalam melakukan *feature selection.*

Berikut akan disajikan parameter yang dibutuhkan dalam seleksi fitur dengan menggunakan metode *Genetic Algorithm* yang tersedia pada tabel 2.

Tabel 2 PARAMETER GA *(Genetic Algorithm)*

Parameter GA	Nilai
Jumlah Individu	300,500 dan 900,1500
Ukuran Populasi	60 dan 100
Maksimum Generasi	5 dan 15
Skema Pengkodean	*Binary Encoding*
Fungsi *Fitness*	Akurasi Naive Bayes
Crossover	*Single Point Crossover*
Peluang *Crossover*	0.8
Mutasi	*Flip Bit Mutation*
Peluang Mutasi	0.1
Mekanisme Seleksi	*Roulette Wheel*
Seleksi Survivor	*Generational Replacement*

2.5 Klasifikasi *Naive Bayes*

Selanjutnya akan dilakukan proses klasifikasi yang menjadi inti dari penyelesaian masalah pada penelitian ini dengan menggunakan metode *Naive Bayes* untuk mengklasifikasikan data *microarray* yang tergolong kanker dan bukan kanker. *Naive Bayes* merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. *Naive Bayes* didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara konditional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu [10]. *Naive Bayes* hanya membutuhkan jumlah data pelatihan *(Data Training)* yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Hai ini yang menjadi keuntungan dalam penggunaan metode *Naive Bayes.* Adapun persamaan *naive bayes* adalah sebagai berikut.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \tag{3}$$

Dimana, *P(H|X)* adalah probabilitas hipotesis *H* berdasarkan kondisi *X (posteriori probabilitas),* Sedangkan *P(X|H)* adalah probabilitas *X* berdasarkan kondisi pada hipotesis *H.* Selanjutnya, *P(H)* adalah probabilitas hipotesis *H* (prior probabilitas). Terakhir *P(X)* adalah probabilitas awal bukti *X* terjadi tanpa memandang hipotesis atau bukti yang lainnya.

Hubungan antara *Naive Bayes* dengan klasifikasi dan bukti klasifikasi adalah bahwa hipotesis dalam teorema *bayes* merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadi masukan dalam model klasifikasi. Jika *X* adalah vektor masukan yang berisi fitur dan Y adalah label kelas, *Naive Bayes* dituliskan dengan *P(X|Y).* Notasi tersebut berarti probabilitas label kelas *Y* didapatkan setelah fitur-fitur *X* diamati. Notasi ini disebut *posterior probability* atau probabilitas akhir untuk *Y,* sedangkan *P(Y)* disebut *prior probability* (probabilitas awal). Adapun persamaan *Naive Bayes* untuk klasifikasi antara lain sebagai berikut:

$$P(H|X) = \frac{\prod_{i=1}^{q} P(X_i|H).P(H)}{P(X)} \tag{4}$$

Keterangan :

P(H|X) = Probabilitas data dengan fitur *X* pada kelas *H.*

P(H) = Probabilitas awal dengan kelas *H.*

$\prod_{i=1}^{q} P(X_i|H)$ = Probabilitas independen kelas *A* dari semua fitur dalam *X.*

Pada umumnya *Naive Bayes* mudah dihitung dan diimplementasikan untuk fitur bertipe kategorikal. Namun berbeda untuk fitur dengan tipe numerik (kontinu) ada perlakuan khusus sebelum dimasukkan ke *Naive Bayes.* Caranya adalah dengan mengasumsikan bentuk tertentu dari distribusi probabilitas untuk fitur kontinu dan memperkirakan parameter distribusi dengan data pelatihan. Distribusi *Gaussian* biasanya dipilih untuk merepresentasikan probabilitas bersyarat dari fitur kontinu pada sebuah kelas *P(Xi|Y),* sedangkan distribusi *Gaussian* dikarakteristikkan dengan dua parameter yaitu mean *(μ)* dan varian *(σ^2)* [11]. Untuk setiap kelas *yj*, probabilitas bersyarat kelas *yj* untuk fitur *Xi* adalah :

$$g(x, \mu, \sigma^2) = P\left(X_i = x_i \middle| Y = y_j\right) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \quad (5)$$

Parameter μ bisa didapat dari *mean* sampel *Xi(x)* dari semua data latih yang menjadi milik kelas *yj*, sedangkan σ^2 dapat diperkirakan dari *varian* sampel (s^2) dari data latih.

2.6 Klasifikasi Regresi Logistik

Regresi Logistik adalah model regresi nonlinier yang digunakan untuk menganalisis pola hubungan antara sekumpulan variabel independen (X) dengan variabel dependen (Y) bertipe kategorik atau kualitatif. Klasifikasi dengan menggunakan regresi logistik bertujuan untuk mengetahui hubungan antara beberapa variabel dimana variabel responnya adalah bersifat kategorik, baik nominal maupun ordinal dengan variabel prediktornya dapat bersifat kategorik atau kontinu. Regresi logistik biner digunakan saat variabel respon merupakan variabel dikotomus (kategorik dengan dua macam kategori). Jika pada regresi logistik variabel responnya terdiri dari dua kategori misalnya *Y=1* hasilnya dinyatakan sebagai "sukses" dan *Y=0* menyatakan hasil yang diperoleh "gagal" maka regresi logistik tersebut menggunakan regresi logistik biner [12]. Berdasarkan [13] model regresi logistik adalah sebagai berikut.

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p} \quad (6)$$

Persamaan (1) tersebut kemudian ditransformasi yang dikenal dengan transformasi logit $\pi(x)$ untuk memperoleh fungsi g(x) yang linear dalam parameternya, sehingga mempermudah pendugaan parameter regresi yang dirumuskan sebagai berikut.

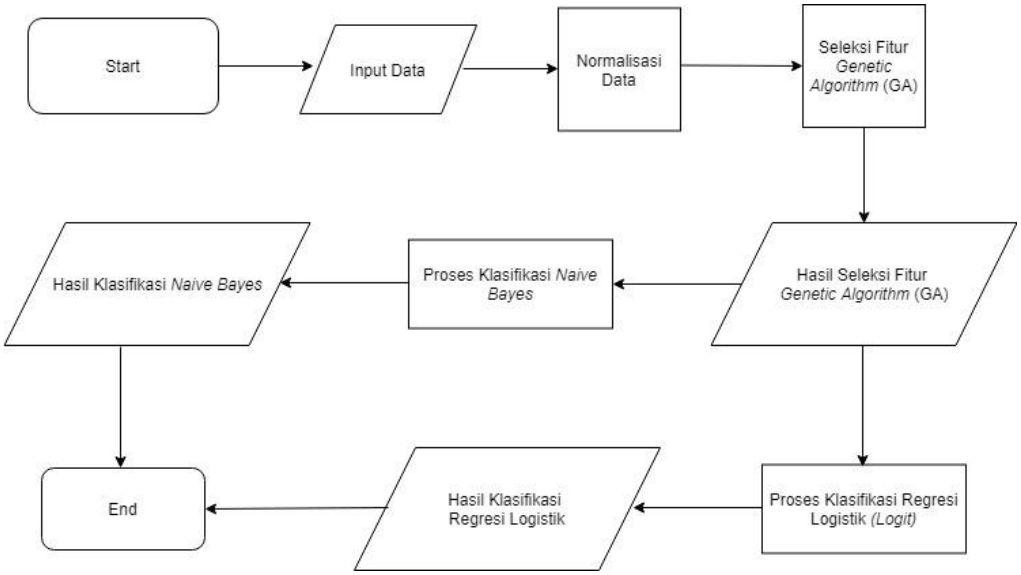
$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (7)$$

Metode *Maximum Likelihood Estimator* (MLE) adalah metode yang digunakan untuk menduga parameter-parameter yang terdapat dalam model regresi logistik. Metode ini menduga β dengan meterbesarkan fungsi *likelihood.*

Salah satu ukuran yang digunakan untuk menginterpretasi koefisien variabel prediktor adalah *Odds ratio* menunjukkan perbandingan peluang munculnya suatu kejadian dengan peluang tidak munculnya kejadian tersebut. Jika nilai *odds ratio* < 1, maka antara variabel prediktor dan variabel respon terdapat hubungan negatif setiap kali perubahan nilai variabel prediktor (X) dan jika nilai *0dds ratio* > 1, maka antara variabel prediktor dan variabel respon terdapat hubungan positif setiap kali perubahan nilai variabel prediktor (X).

3. Sistem yang Dibangun

Gambaran umum dari sistem klasifikasi dengan melibatkan seleksi fitur yang dibangun dalam penelitian ini dibuat dalam diagram alur *(flowchart).*



Gambar 1. Diagram Alur *(flowchart)*

Umumnya data *microarray* adalah data yang memiliki dimensi yang sangat besar. Hal ini sangat dapat mempengaruhi proses klasifikasi nantinya dan akan menyebabkan beban perhitungan menjadi tidak optimal sehingga akurasinya nanti menghsilkan nilai yang sangat kecil atau tidak optimal. Selain itu, data *microarray* juga memiliki skla *(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). Seperti yang sudah dijelaskan sebelumnya, seleksi fitur ini bertujuan untuk memilih fitur yang optimal dan mengesampingkan fitur yang tidak optimal atau tidak berpengaruh. Proses seleksi fitur ini dilakukan karena jumlah fitur ataupun atribut pada data *microarray* sangat banyak. Setelah seleksi fitur dilakukan, selanjutnya melakukan proses klasifikasi dengan *Naive Bayes* dan Regresi Logistik untuk menentukan data mana yang termasuk kelas kanker dan mana yang bukan kanker pada setiap sampel yang ditentukan.

3.1 Spesifikasi *Dataset*

Data yang digunakan dalam penelitian Tugas Akhir ini adalah data dari Kent Ridge. Dataset terdiri dari *Colon Cancer* dan *MLL Leukemia*. Berikut spesifikasi data yang digunakan dalam penelitian ini.

Tabel 3 Spesifikasi Data

Data	Jumlah Kelas	Jumlah Record	Jumlah Gen
Colon Cancer	2	62	2.000
MLL Leukemia	2	38	7.129

4. Evaluasi

Pada bagian ini akan dilakukan pengujian dan analisis untuk data *microarray* dalam klasifikasi kelas kanker dan bukan kanker. Pengujian ini dilakukan dengan menggunakan metode seleksi fitur *Genetic Algorithm* (GA), dan klasifikasi menggunakan metode *Naive Bayes* dan Regresi Logistik. Dataset yang digunakan pada pengujian ini yaitu *Colon cancer* dan *MLL leukemia*. Data yang digunakan bersumber dari Kent-Ridge Bio-medical Data Set Repository.

4.1 Pengujian Sistem

Adapun tujuan dari pengujian sistem adalah untuk mengetahui gen yang merupakan kanker dan bukan kanker dari dataset yang akan dianalisis. Pengujian ini dilakukan melalui seleksi fitur dan klasifikasi. Adapun tujuan dari seleksi fitur yaitu untuk mengurangi jumlah fitur yang tidak berpengaruh sehingga pada perhitungan dapat menghasilkan akurasi yang maksimal. Kemudian akan dilakukan proses klasifikasi dengan menggunakan metode *Naive Bayes* dan Regresi Logistik. Dengan memperhatikan jumlah atribut dan proporsi pengujian pada masing-masing data dari seleksi fitur dan klasifikasi. Kedua metode klasifikasi nantinya akan dibandingkan akurasi nya berdasarkan proporsi pengujian masing-masing data. Berikut adalah strategi yang digunakan dalam proses pengujian sistem melalui beberapa skenario, yaitu :

1. Skenario pertama akan dilakukan klasifikasi dari setiap *dataset microarray* dengan menggunakan metode *Naive Bayes* melalui seleksi fitur dengan menggunakan metode *Genetic Algorithm* (GA). Berikut adalah langkah-langkah skenario pertama:
   a. Menentukan parameter data input dari data *microarray* yang akan diuji. Kemudian mengolah data yang akan diuji tersebut dengan menormalisasi setiap data input dari *microarray* sehingga *range* masing-masing data berada pada *range* 0 sampai 1.
   b. Membagi *dataset microarray* menjadi data latih *(data train)* dan data uji *(data testing)* dengan proporsi pembagian data uji dan data latih 70:30 dan 80:20.
   c. Langkah selanjutnya, menentukan ukuran populasi yang akan dievaluasi. Dalam Tugas Akhir ini penulis menggunakan ukuran populasi 60 dan 100 untuk individu yang akan dievaluasi dengan maksimum generasi 5 dan 15 (iterasi), sehingga jumlah individu nantinya sebanyak 300 dan 500 serta 900 dan 1500 individu.
   d. Selanjutnya, dilakukan proses *training* untuk mencari nilai *mean* dan standar deviasi dari suatu individu. Setelah mendapatkan nilai *mean* dan standar deviasi tersebut, maka akan dilakukan proses *testing* dengan menggunakan *Naive Bayes.*
   e. Mencari nilai akurasi dari proses *testing* berdasarkan data yang akan di *testing* pada *naive bayes* dan kemudian nilai tersebut akan dijadikan sebagai nilai *fitness* pada *Genetic Algorithm.*
   f. Kemudian menentukan nilai *Pc* atau probabilitas *crossover* dan nilai *Pm* atau probabilitas mutasi. Pada Tugas Akhir ini penulis menggunakan nilai *Pc* sebesar 0,8 sedangkan untuk nilai *Pm* sebesar 0,1 dengan merujuk pada jurnal [14].
   g. Performansi yang sudah didapatkan dapat dianalisis dan dibandingkan dari proses testing pada masing-masing proporsi yang menghasilkan akurasi tertinggi.
2. Skenario kedua akan dilakukan klasifikasi dari setiap *dataset microarray* dengan menggunakan metode Regresi Logistik. Berikut adalah langkah-langkah dari skenario kedua:
   a. Menentukan data input dari data *microarray* yang akan dilakukan pengujian.
   b. Kemudian mengolah data yang sudah di input dengan cara membagi data tersebut menjadi data latih *(data train)* dan data uji *(data testing)* dengan proporsi masing-masing data latih dan data uji adalah 70:30, 80:20 dan 90:10.
   c. Langkah selanjutnya, memodelkan data latih dengan masing-masing proporsinya beserta dengan kelasnya dengan menggunakan transformasi logit.
   d. Selanjutnya, melakukan prediksi dari model yang sudah dibuat dengan logit dengan data uji (data yang bertipe response).
   e. Membuat aturan prediksi dengan hasil prediksi lebih besar dari 0.5,0 dan 1.

f. Kemudian menghitung *misClassificError* dengan cara menghitung nilai *mean* dari hasil prediksi dengan kelasnya.
g. Melakukan perhitungan akurasi dengan menggunakan nilai dari *misClassificError* yang sudah didapat.
h. Performansi yang sudah didapatkan dapat dianalisis dan dibandingkan dari proses regresi logistik pada masing-masing proporsi data yang menghasilkan akurasi tertinggi.

3. Skenario ketiga akan dilakukan klasifikasi dari setiap *dataset microarray* demgan menggunakan metode Regresi Logistik dengan menggunakan data hasil reduksi dimensi dari *Genetic Algorithm* (GA). Berikut adalah langkah-langkah dari skenario ketiga:
   a. Menentukan data input dari data *microarray* yang akan dilakukan pengujian.
   b. Kemudian memasukkan data hasil reduksi dimensi dengan menggunakan *Genetic Algorithm* untuk dilakukan pengujian dengan menggunakan data tersebut.
   c. Mengolah data yang sudah di input (data hasil reduksi dimensi di GA) dengan cara membagi data tersebut menjadi data latih *(data train)* dan data uji *(data testing)* dengan proporsi masing-masing data latih dan data uji adalah 70:30, 80:20 dan 90:10.
   d. Langkah selanjutnya, memodelkan data latih hasil reduksi dimensi dengan masing-masing proporsinya beserta dengan kelasnya dengan menggunakan transformasi logit.
   e. Selanjutnya, melakukan prediksi dari model yang sudah dibuat dengan logit dengan data uji (data yang bertipe response).
   f. Membuat prediksi dengan ketentuan hasil dari prediksi lebih besar dari 0.5, 0 dan 1.
   g. Kemudian menghitung *misClassificError* dengan cara menghitung nilai *mean* dari hasil prediksi dengan kelasnya.
   h. Melakukan perhitungan akurasi dengan menggunakan nilai dari *misClassificError* yang sudah didapat.
   i. Performansi yang sudah didapatkan dapat dianalisis dan dibandingkan dari proses regresi logistik dengan menggunakan data hasil reduksi dimensi *Genetic Algorithm* pada masing-masing proporsi data yang menghasilkan akurasi tertinggi.
4. Skenario keempat akan dilakukan klasifikasi dari setiap *dataset microarray* dengan menggunakan metode klasifikasi *Naive Bayes*. Berikut adalah langkah-langkah dari skenario keempat:
   a. Menentukan parameter data input dari data *microarray* yang akan diuji. Kemudian mengolah data yang akan diuji tersebut dengan menormalisasi setiap data input dari *microrray* sehingga *range* masing-masing data berada pada *range* 0 sampai 1.
   b. Melakukan pembagian data latih *(data train)* dan data uji *(data test)* dengan proporsi 70:30 dan 80:20.
   c. Melakukan proses training dengan menggunakan model klasifikasi *Naive Bayes* untuk mencari *mean* dan standar deviasi dari data yang akan diuji.
   d. Selanjutnya, melakukan proses testing dengan menggunakan parameter yang didapat pada saat melakukan proses training untuk memprediksi kelasnya.
   e. Performansi yang sudah didapatkan dapat dianalisis dan dibandingkan dari proses *naive bayes* pada masing-masing proporsi data yang menghasilkan akurasi tertinggi.
5. Skenario kelima membandingkan performansi antara skenario pertama, kedua, ketiga dan keempat.

4.2 Hasil dan Analisis

Pengujian yang telah dilakukan pada masing-masing skenario yang telah dibuat akan menghasilkan akurasi pada setiap skenarionya. Akurasi yang telah dihasilkan akan dilakukan analisis untuk mengetahui skenario mana yang terbaik yang dihasilkan pada pengujian masing-masing skenario yang telah dilakukan. Berikut akan dituliskan analisis dari setiap pengujian pada masing-masing skenario.

4.2.1 Hasil Seleksi Fitur *Genetic Algorithm* dengan Klasifikasi *Naive Bayes*

Pada bagian ini akan ditampilkan hasil penelitian dari skenario seleksi fitur *Genetic Algorithm* dan klasifikasi *Naive Bayes*. Individu yang akan dievaluasi pada penelitian Tugas Akhir ini adalah 60 dan 100 populasi dengan menggunakan maksimum generasi 5 dan 15, sehingga jumlah individu sebanyak 300 dan 500 serta 900 dan 1500 individu. Probabilitas *crossover pc* yang digunakan dalam pengujian adalah 0,8 sedangkan probabilitas mutasi *pm*

yang digunakan adalah 0,1. Berikut hasil seleksi fitur *Genetic Algorithm* dengan klasifikasi menggunakan metode *Naive Bayes*:

Tabel 4 Hasil Seleksi Fitur dengan *Genetic Algorithm* dan Klasifikasi *Naive Bayes*

Dataset	Jumlah Individu	Ukuran Populasi	Akurasi (%)		Rata – rata Akurasi (%)
			Proporsi 70:30	Proporsi 80:20	
Colon Tumor	300	60	41,1765	40	40,58825 %
	500	100	41,1765	40	40,58825 %
	900	60	41,1765	40	40,58825 %
	1500	100	41,1765	40	40,58825 %
MLL Leukemia	300	60	55,5556	60	57,7778 %
	500	100	55,5556	60	57,7778 %
	900	60	55,5556	60	57,7778 %
	1500	100	55,5556	60	57,7778 %

Berdasarkan Tabel 4 untuk proporsi pengujian 70:30 dengan jumlah individu 300 mampu memberikan hasil untuk *colon tumor* 41,1765 % dan *mll leukemia* 55,5556 %. Begitu juga dengan jumlah individu 900 memberikan hasil akurasi yang sama dengan jumlah individu 300. Sedangkan, untuk proporsi pengujian 70:30 dengan jumlah individu 500 mampu memberikan hasil untuk *colon tumor* 41,1765 % dan *mll leukemia* 55,5556 %. Dan pada jumlah individu 1500 juga memberikan hasil akurasi yang sama dengan jumlah individu 500. Kemudian, untuk proporsi pengujian 80:20 dengan jumlah individu 300 mampu memberikan hasil untuk *colon tumor* 40 % dan *mll leukemia* 60 %. Begitu juga dengan jumlah individu 900 memberikan hasil akurasi yang sama dengan jumlah individu 300. Sedangkan, untuk proporsi pengujian 80:20 dengan jumlah individu 500 mampu memberikan hasil untuk *colon tumor* 40 % dan *mll leukemia* 60 %. Dan pada jumlah individu 1500 juga memberikan hasil akurasi yang sama dengan jumlah individu 500. Dapat disimpulkan dari penelitian diatas jumlah maksimum generasi 5 dan 15 tidak berpengaruh pada hasil akurasi setiap proporsi masing-masing data (akurasinya sama).

Berdasarkan pengujian masing-masing data dengan proporsi yang sudah ditetapkan mampu menghasilkan rata rata akurasi untuk penyakit *colon tumor* 40,58825 % dan *mll leukemia* 57,7778 % dengan performansi tertinggi didapat dari proporsi 70:30 untuk data *colon tumor* dan *mll leukemia.* Sedangkan, proporsi pengujian 80:20 untuk data *mll leukemia.* Hal ini dapat dilihat dari kinerja yang menurun pada proporsi 80:20 untuk data *colon tumor* sebanyak 1,17% dan pada *mll leukemia* terjadi penurunan kinerja pada proporsi 70:30 sebanyak 4,4%. Sebaliknya untuk data *mll leukemia* proporsi 80:20 merupakan pengujian terbaik, karena mampu menghasilkan akurasi yang cukup baik. Dapat disimpulkan bahwa perlunya pemilihan jumlah atribut dan proporsi pengujian yang tepat akan memberikan hasil performansi yang baik. Berikut hasil *Confussion Matrix* dari seleksi fitur dengan *Genetic Algorithm* dan klasifikasi *Naive Bayes* berdasarkan proporsi datanya masing-masing:

Tabel 5 *Confussion Matrix* data *colon tumor* proporsi 70:30

Ukpop = 60		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 1	(FP) 0	1
	Negatives	(FN) 10	(TN) 6	16
Total		**11**	**6**	**17**
Ukpop = 100		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 1	(FP) 0	1
	Negatives	(FN) 10	(TN) 6	16
Total		**11**	**6**	**17**

Tabel 6 *Confussion Matrix* data *colon tumor* proporsi 80:20

Ukpop = 60		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 1	(FP) 0	1
	Negatives	(FN) 6	(TN) 3	9
Total		**7**	**3**	**10**
Ukpop = 100		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 1	(FP) 0	1
	Negatives	(FN) 6	(TN) 3	9
Total		**7**	**3**	**10**

Tabel 7 *Confussion Matrix* data *leukemia* proporsi 70:30

Ukpop = 60		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 4	(FP) 1	5
	Negatives	(FN) 3	(TN) 1	4
Total		**7**	**2**	**9**
Ukpop = 100		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 4	(FP) 1	5
	Negatives	(FN) 3	(TN) 1	4
Total		**7**	**2**	**9**

Tabel 8 *Confussion Matrix* data *leukemia* proporsi 80:20

Ukpop = 60		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 2	(FP) 0	2
	Negatives	(FN) 2	(TN) 1	3
Total		**4**	**1**	**5**
Ukpop = 100		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 2	(FP) 0	2
	Negatives	(FN) 2	(TN) 1	3
Total		**4**	**1**	**5**

4.2.2 Klasifikasi dengan Regresi Logistik

Pada bagian ini akan ditampilkan hasil penelitian dari skenario klasifikasi dengan Regresi Logistik. Hasil pengujian menggunakan proporsi 70:30, 80:20 dan 90:10 pada setiap data yang akan diuji. Berikut hasil klasifikasi dengan Regresi Logistik:

Tabel 9 Hasil Klasifikasi dengan Regresi Logistik

Dataset	**Akurasi (%)**			**Rata-rata Akurasi (%)**
	Proporsi 70:30	**Proporsi 80:20**	**Proporsi 90:10**	
Colon Tumor	56	50	66	57,3 %
MLL Leukemia	67	67	67	67%

Berdasarkan Tabel 4.6 untuk proporsi 70:30 mampu memberikan hasil untuk *colon tumor* sebesar 56% dan *mll leukemia* sebesar 67%. Sedangkan, untuk proporsi 80:20 mampu memberikan hasil 50% untuk *colon tumor* dan 67% untuk *mll leukemia.* Pada proporsi 90:10 mampu memberikan hasil untuk *colon tumor* 66% dan *mll leukemia* 67%.

Berdasarkan pengujian masing-masing data dengan proporsi yang sudah ditetapkan mampu menghasilkan rata-rata akurasi untuk penyakit *colon tumor* 57,3% dan *mll leukemia* 67%. Hal ini dapat dilihat dari kinerja yang mengalami penurunan pada proporsi 80:20 *colon tumor* sebesar 6%, kemudian mengalami peningkatan pada proporsi 90:10 sebesar 16%. Pada data *mll leukemia* masing-masing proporsi mempunyai akurasi yang sama. Data *mll leukemia* merupakan hasil pengujian terbaik karena mampu menghasilkan rata-rata akurasi yang cukup baik.

4.2.3 Klasifikasi dengan Regresi Logistik menggunakan Hasil Reduksi Dimensi *Genetic Algorithm*

Pada bagian ini akan ditampilkan hasil penelitian dari skenario klasifikasi dengan regresi Logistik dengan menggunakan inputan data hasil reduksi dimensi dengan *Genetic Algorithm.* Pengujian pada skenario ini menggunakan proporsi 70:30, 80:20 dan 90:10. Berikut hasil klasifikasi dengan Regresi Logistik dan Reduksi Dimensi *Genetic Algorithm:*

Tabel 10 Hasil Klasifikasi Regresi Logistik dengan Reduksi Dimensi *Genetic Algorithm*

Dataset	Akurasi (%)			Rata-rata Akurasi (%)
	Proporsi 70:30	Proporsi 80:20	Proporsi 90:10	
Colon Tumor	62	60	100	74%
MLL Leukemia	75	83	100	86%

Berdasarkan Tabel 10 untuk proporsi 70:30 mampu memberikan hasil untuk *colon tumor* sebesar 62% dan *mll leukemia* sebesar 75%. Sedangkan untuk proporsi 80:20 mampu memberikan hasil 60% untuk *colon tumor* dan 83% untuk *mll leukemia.* Pada proporsi 90:10 mampu memberikan hasil untuk *colon tumor* sebesar 100% dan *mll leukemia* sebesar 100%.

Berdasarkan pengujian masing-masing data dengan proporsi yang sudah ditetapkan mampu menghasilkan rata-rata akurasi untuk penyakit *colon tumor* sebesar 74% dan *mll leukemia* sebesar 86%. Hal ini bisa dilihat dari kinerja yang mengalami penurunan pada proporsi 80:20 *colon* tumor sebesar 2%, kemudian mengalami peningkatan pada proporsi 90:10 sebesar 40%. Dan pada data *mll leukemia* terjadi peningkatan 8% pada proporsi 80:20 dan meningkat sebesar 17% pada proporsi 90:10. Data *mll leukemia* merupakan pengujian terbaik karena mampu menghasilkan rata-rata akurasi yang sangat baik dan juga terjadi peningkatan pada setiap proporsi pengujiannya.

4.2.4 Klasifikasi dengan *Naive Bayes*

Pada bagian ini akan ditampilkan hasil penelitian dari skenario klasifikasi dengan *Naive Bayes.* Hasil pengujian menggunakan proporsi 70:30 dan 80:20. Berikut hasil klasifikasi dengan *Naive Bayes*:

Tabel 11 Hasil Klasifikasi dengan *Naive Bayes*

Dataset	Akurasi (%)		Rata-rata Akurasi (%)
	Proporsi 70:30	Proporsi 80:20	
Colon Tumor	58	40	49 %
MLL Leukemia	100	100	100 %

Berdasarkan Tabel 11 untuk proporsi 70:30 mampu memberikan hasil untuk *colon tumor* sebesar 58% dan *mll leukemia* sebesar 100%. Sedangkan untuk proporsi 80:20 mampu memberikan hasil 40% untuk *colon tumor* dan 100% untuk *mll leukemia.*

Berdasarkan pengujian masing-masing data dengan proporsi yang sudah diteteapkan mampu menghasilkan rata-rata akurasi untuk penyakit *colon tumor* sebesar 49% dan *mll leukemia* sebesar 100%. Hal ini bisa dilihat dari kinerja yang mengalami penurunan pada proporsi 80:20 data *colon tumor* sebesar 18%. Pada data *mll leukemia* proporsi 70:30 dan 80:20 akurasi yang dihasilkan stabil sebesar 100%. Data *mll leukemia* merupakan pengujian terbaik karena mampu menghasilkan rata-rata akurasi yang sangat baik dan kinerja yang dihasilkan pada masing-masing proporsi data stabil sebesar 100%.

4.2.5 Perbandingan Performansi *Genetic Algorithm* pada Klasifikasi *Naive Bayes* dan Regresi Logistik

Berdasarkan keempat pengujian yang sudah dilakukan, maka dapat disimpulkan masing-masing metode menghasilkan performansi yang baik. Adapun perbandingan performansi pada masing-masing metode yang digunakan dalam pengujian Tugas Akhir ini adalah sebagai berikut:

Tabel 12 Perbandingan Performansi *Genetic Algorithm* pada klasifikasi *Naive Bayes* dan Regresi Logistik

Dataset	Reduksi dan Klasifikasi	Rata-rata Akurasi (%)
Colon Tumor	*Genetic Algorithm* dengan *Naive Bayes*	40,58825%
	Genetic Algorithm dengan Regresi Logistik	74%
	Regresi Logistik	57,3%
	Naive Bayes	49 %
MLL Leukemia	*Genetic Algorithm* dengan *Naive Bayes*	57,7778%
	Genetic Algorithm dengan Regresi Logistik	86%%
	Regresi Logistik	67%
	Naive Bayes	100 %

Berdasarkan Tabel 12 maka dapat disimpulkan seleksi fitur *Genetic Algorithm* dengan klasifikasi Regresi Logistik mampu memberikan hasil yang lebih baik pada data *colon tumor*. Sedangkan klasifikasi dengan Regresi Logistik (tanpa GA) mampu memberikan hasil yang baik jika dibandingkan dengan seleksi fitur *Genetic Algorithm* dengan klasifikasi *Naive Bayes.*

Pada data *mll leukemia*, klasifikasi *Naive Bayes* dan seleksi fitur *Genetic Algorithm* dengan klasifikasi Regresi Logistik mampu memberikan hasil yang sangat baik jika dibandingkan dengan seleksi fitur *Genetic Algorithm* dengan klasifikasi *Naive Bayes* dan Regresi Logistik yang tanpa GA.

Sesuai dengan pengujian yang telah dilakukan, maka performansi terbaik dari seleksi fitur *Genetic Algorithm* dan klasifikasi *Naive Bayes* mampu memberikan hasil untuk penyakit *colon tumor* 40,58825% dan *mll leukemia* 57,7778%. Seleksi fitur *Genetic Algorithm* dengan klasifikasi Regresi Logistik mampu memberikan hasil untuk penyakit *colon tumor* 74% dan *mll leukemia* 86%. Klasifikasi dengan Regresi Logistik (tanpa GA) mampu memberikan hasil untuk penyakit *colon tumor* 57,3% dan *mll leukemia* 67%. Sedangkan klasifikasi dengan *Naive Bayes* (tanpa seleksi fitur) mampu memberikan hasil untuk penyakit *colon tumor* 49% dan *mll leukemia* 100%.

5. Kesimpulan dan Saran

Berdasarkan hasil dan analisis pada Tugas Akhir ini, diperoleh kesimpulan bahwa jika dilihat berdasarkan rata-rata keseluruhan data latih *(train)* dan data uji *(test)* maka akurasi tertinggi adalah klasifikasi dengan *Naive Bayes* dan metode seleksi fitur dengan *Genetic Algorithm* dan Klasifikasi dengan Regresi Logistik sebesar 100% dan 86%. Kemudian jika dilihat berdasarkan data maka akurasi tertinggi terdapat pada data *colon tumor* dan *mll leukemia* sebesar 100% dengan metode seleksi fitur *Genetic Algorithm* dan klasifikasi Regresi Logistik pada proporsi 90:10. Begitu juga klasifikasi dengan *Naive Bayes* (tanpa seleksi fitur) pada data *mll leukemia* menghasilkan akurasi tertinggi sebesar 100% pada proporsi 70:30 dan 80:20.

Setelah melakukan proses klasifikasi pada data *microarray*, penulis memiliki saran dalam pengerjaan penelitian ini untuk proses pengembangan kedepannya, yaitu memperhatikan setiap atribut data yang digunakan untuk proses klasifikasi, mempelajari terlebih dahulu metode yang ingin digunakan dalam proses seleksi fitur dan

klasifikasi, begitu juga dengan spesifikasi laptop atau komputer yang digunakan harus diperhatikan sehingga proses penelitian dan pengujian dapat berjalan dengan lancar dan tidak mengalami hambatan.

Daftar Pustaka

[1] Rahman, Ahmad Abdul. 2014. *Kanker*. Global Bioscience. http://www.CancerHelps.com. (Diakses pada 1 November 2018).

[2] Pangesti, Agnes Widhiya. 2016. *Identifikasi faktor resiko kanker serviks pada mahasiswi Universitas Muhhamadiyah Yogyakarta.* Universitas Muhhamadiyah Yogyakarta. 3(2):1. http://repository.umy.ac.id/bitstream/handle/123456789/2681/Agnes%20Widhiya%20Pangesti_20120320101_5_BAB%201.pdf?sequence=5&isAllowed=y. (Diakses pada 31 Oktober 2018).

[3] Kementerian Kesehatan 2015. Situasi Penyakit Kanker. http://www.depkes.go.id/resources/download/pusdatin/infodatin/infodatin-kanker.pdf. (Diakses pada 31 November 2018).

[4] Vanitha, C.D.A., Devaraj, D. And Venkatesulu, M., 2015. *Gene Espression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection.* Procedia Computer Science, 47, pp.13-21.

[5] Mukesh Kumar, Sandeep Singh, and Santanu Kumar Rath. *Classification of microarray data using functional link neural network.* Procedia Computer Science, 57:727-737, 2015.

[6] Ramadhani, Putri Tsatsabila. 2017. *Deteksi Kanker berdasarkan Klasifikasi Data Microarray menggunakan Functional Link Neural Netwrok dengan Seleksi Fitur Genetic Algorithm.* Indo-JC, Vol.2:13.

[7] Generasi Biologi. 2016. *Microarray:Biologi di Era Pascagenomik.* http://www.generasibiologi.com/2012/08/microarray-biologi-di-era-pascagenomik.hrml. (Diakses pada 1 November 2018).

[8] Nugroho, Mohamad Fajarianditya., Wibowo, Setyoningsih. 2017. *Fitur Seleksi Forward Selection Untuk Menentukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naive Bayes.* Jurnal Informatika Upgris, Vol.3, No.7.

[9] Sarmilah, Milah. 2018. *Analisis Seleksi Fitur Genetic Algorithm dan Ekstraksi Fitur Wavelet Pada Klasifikasi Microarray Data Menggunakan Naive Bayes.* Bandung, Open Library Telkom University.

[10] Saleh, Alfa. 2015. *Klasifikasi Metode Naive Bayes Dalam Data Mining Untuk Menentukan Konsentrasi Siswa.* KeTIK, ISBN : 979-458-766-4

[11] Prasetyo, Eko. 2012. Data Mining Konsep dan Aplikasi Menggunakan MATLAB. Yogyakarta:Andi

[12] Ramli, 2013. *Perbandingan Metode Klasifikasi Regresi Logistik Dengan Jaringan Saraf Tiruan*. Jurnal EKSPONENSIAL Vol.4, Nomor 1.

[13] Hosmer, D. W., dan Lemeshow, S. (2000). *Applied Logistic Regression.* New York: John Wiley & Sons, Inc.

[14] O Babatunde, Leisa Armstrong, Jinsong Leng, and Dean Diepeveen. A genetic Algorithm-based feature selection. International Journal of Electronics Communication and Computer Engineering, 5(4):889-905, 2014.

[15] 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.

[16] Husna Aydadenta, Adiwijaya, (2018). *A Clustering Approach for Feature Selection in Microarray Data Classification using Random Forest*, Journal of Information Processing System 14(5).

[17] 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).

[18] Pratiwi, M. S., Aditsania, A., Adiwijaya. 2018. *Cancer Detection Based on Microarray Data Classification using Genetic Bee Colony (GBC) and Conjugate Gradient Backpropagation with Modified Polak Ribiere (MBP-CGP).* In 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA) (pp. 163-168). IEEE.

[19] Astuti, W., dan Adiwijaya, A. (2019). *Principal Component Analysis Sebagai Ekstraksi Fitur Data Microarray Untuk Deteksi Kanker Berbasis Linear Discriminant Analysis.* JURNAL MEDIA INFORMATIKA BUDIDARMA, 3(2), 72-77.

[20] Adiwijaya, A. (2018). *Deteksi Kanker Berdasarkan Klasifikasi Microarray Data*. JURNAL MEDIA INFORMATIKA BUDIDARMA, 2(4), 181-186.

[21] Ma'ruf, F. A., Adiwijaya and Wisesty, U. N. (2019, March). *Analysis of the influence of Minimum Redundancy Maximum Relevance as dimensionality reduction method on cancer classification based on microarray data using Support Vector Machine classifier.* In Journal of Physics: Conference Series (Vol. 1192, No. 1, p.012011). IOP Publishing.

[22] Purbolaksono, M. D., Widiastuti, K. C., Adiwijaya, Mubarok, M. S., and Ma'ruf, F. A. (2018, March). *Implementation of mutual information and bayes theorem for classification microarray data.* In Journal of Physics: Conference Series (Vol. 971, No. 1, p. 012011). IOP Publishing.

[23] Nugroho, Dwi. 2016. *Prediksi Penyakit Menggunakan Genetic Algorithm (GA) dan Naive Bayes untuk Data Berdimensi Tinggi.* Bandung: Telkom University.

[24] Liontoni, F. dan Manik, F. Y. 2016. *Klasifikasi Daun Herbal Menggunakan Metode Naive Bayes Classifier dan K-nearest Neighbor.* Jurnal Simantec.

[25] Anandita, E. R. 2014. *Klasifikasi Tebu Dengan Menggunakan Algoritma Naive Bayes Classification Pada Dinas Kehutanan dan Perkebunan Pati*. Skripsi, Jurusan Sistem Informasi, Universitas Dian Nuswantoro, Semarang.

[26] Juwita, Puspa., Sugiman, Hendikawati, Putriaji. (2018). *Ketepatan Klasifikasi Metode Regresi Logistik dan CHAID dengan Pembobotan Sampel*. PRISMA, Prosiding Seminar Nasional Matematika.

[27] Rajagukguk, Nanci., Ispriyanti, Dwi., Wilandari, Yuciana. 2015. *Perbandingan Metode Klasifikasi Regresi Logistik Biner Dan Naive Bayes Pada Status Pengguna KB Di Kota Tegal Tahun 2014.* JURNAL GAUSSIAN, Vol. 4, No. 2, 2015, 365-374.

[28] Sari, P, K., and Purwadinata, A. (2019). *Analysis Characteristics of Car Sales In E-Commerce Data Using Clustering Model.* Journal of Data Science and Its Applications, 2(1), 68-77.

[29] Pusponegoro, N. H., Djuraidah, A., Fitrianto, A., and Sumertajaya, I. M. (2019). *Geo-additive Models in Small Area Estimation of Poverty.* Journal of Data Science and Its Applications, 2(1), 59-67.

[30] Manik, A., Adiwijaya, A., and Utama, D. Q. (2019). *Classification of Electrocardiogram Signals using Principal Component Analysis and Levenberg Marquardt Backpropagation for Detection Ventricular Tachyarrhythmia.* Journal of Data Science and Its Applications, 2(1), 78-87.

[31] Naf'an, M. Z., Bimantara, A. A., Larasati, A., Risondang, E. M., and Nugraha, N. A. S. (2019). *Sentiment Analysis of Cyberbullying on Instagram User Comments.* Journal of Data Science and Its Applications, 2(1), 88-98.

Lampiran 1 Data Sampel

Tabel 1 Sampel Data *Colon Tumor*

No.	Fitur 1	Fitur 2	Fitur 3	...	Fitur 2000	Kelas
1	8589.416	5468.241	4263.408	...	28.70125	1
2	9164.254	6719.53	4883.449	...	16.77375	0
3	3825.705	6970.361	5369.969	...	15.15625	1
4	6246.449	7823.534	5955.835	...	16.085	0
5	3230.329	3694.45	3400.74	...	31.8125	1
...	...	...	...	...	...	...
62	7472.01	3653.934	2728.216	...	39.63125	0

Tabel 2 Sampel Data *MLL Leukemia*

No.	Fitur 1	Fitur 2	Fitur 3	...	Fitur 7129	Kelas
1	0.53144	0.566775	0.435315	...	0.438462	1
2	0.68357	0.827362	0.534965	...	0.615385	1
3	0.811359	0.905537	0	...	0.407692	1
4	0.691684	0.693811	1	...	0.023077	1
5	0.750507	0.65798	0.403846	...	0.530769	1
...	...	...	...	...	...	...
38	0.691684	0.459283	0.414336	...	0.646154	0

Lampiran 2 Hasil *Confussion Matrix Genetic Algorithm* dengan *Naive Bayes*

Tabel 3 *Confussion Matrix* data *colon tumor* proporsi 70:30

Ukpop = 60		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 1	(FP) 0	1
	Negatives	(FN) 10	(TN) 6	16
Total		**11**	**6**	**17**
Ukpop = 100		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 1	(FP) 0	1
	Negatives	(FN) 10	(TN) 6	16
Total		**11**	**6**	**17**

Tabel 4 *Confussion Matrix* data *colon tumor* proporsi 80:20

Ukpop = 60		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 1	(FP) 0	1
	Negatives	(FN)	(TN)	9

		6	3	
Total		**7**	**3**	**10**
Ukpop = 100		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 1	(FP) 0	1
	Negatives	(FN) 6	(TN) 3	9
Total		**7**	**3**	**10**

Tabel 5 *Confussion Matrix* data *leukemia* proporsi 70:30

Ukpop = 60		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 4	(FP) 1	5
	Negatives	(FN) 3	(TN) 1	4
Total		**7**	**2**	**9**
Ukpop = 100		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 4	(FP) 1	5
	Negatives	(FN) 3	(TN) 1	4
Total		**7**	**2**	**9**

Tabel 6 *Confussion Matrix* data *leukemia* proporsi 80:20

Ukpop = 60		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 2	(FP) 0	2
	Negatives	(FN) 2	(TN) 1	3
Total		**4**	**1**	**5**
Ukpop = 100		True Class		Total
		True	False	
Predicted Class	Positives	(TP) 2	(FP) 0	2
	Negatives	(FN) 2	(TN) 1	3
Total		**4**	**1**	**5**