# Prediksi Harapan Hidup Pasca Operasi Toraks pada Pasien Penderita Kanker Paru-paru Menggunakan Metode *Genetic Algorithm* untuk *Feature Selection* dan *Naïve Bayes Classifier*

**Yuniar Agung Setyadi[1], Ibnu Asror[2], Yanuar Firdaus Arie Wibowo[3]**

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]yuniaragung@students.telkomuniversity.ac.id, [2]iasror@telkomuniversity.ac.id, [3]yanuar@telkomuniversity.ac.id

**Abstrak**

**Penanganan dini yang dilakukan untuk menekan tingkat kematian pasien kanker paru-paru pasca operasi toraks, dengan mengumpulkan data berupa informasi tentang pasien pasca operasi toraks menimbulkan masalah baru yaitu data berdimensi tinggi yang memiliki banyak atribut dan tidak bisa menghasilkan informasi yang akurat. Oleh karena itu, diperlukan skema komputasi yang dapat mereduksi dimensi pada data tersebut. Dalam hal ini, proses reduksi bertujuan untuk meringankan beban komputasi pada klasifikasi, proses reduksi yang digunakan yaitu seleksi fitur *Genetic Algorithm*. Kemudian metode klasifikasi *Naïve Bayes* digunakan untuk melakukan proses klasifikasi harapan hidup pasca operasi toraks. Adapun akurasi terbaik yang dihasilkan dari seleksi fitur *Genetic Algorithm* dan *Naïve Bayes Classifier* yaitu 85,319%.**

**Kata kunci : operasi toraks, *genetic algorithm, naïve bayes***

***Abstract***

***Early treatment is carried out to reduce the death rate of lung cancer patients after thoracic surgery by collecting data in the form of information about postoperative thoracic patients, creating a new problem that is high-dimensional data that has many attributes and cannot produce accurate information. Therefore, a computational scheme is needed to reduce the dimensions of the data. In this case, the reduction process aims to ease the computational burden on classification, the reduction process used is the Genetic Algorithm for Feature Selection. Then the Naïve Bayes Classifier method is used to carry out the classification process of life expectancy after thoracic surgery. The best accuracy resulting from the selection of Genetic Algorithm and Naïve Bayes Classifier features is 85,319%.***

***Keywords: thoracic surgery, genetic algorithm, naïve bayes***

## 1. Pendahuluan

Angka kematian operasi telah menjadi topik yang menarik di antara ahli bedah, pasien, pengacara, dan administrator kebijakan kesehatan di era sekarang. Komplikasi pernapasan pasca operasi dan semua jenis operasi toraks merupakan penyebab kematian yang paling umum yang sering terjadi.

Operasi toraks adalah cabang ilmu kedokteran yang mempelajari diagnosis dan tindakan untuk gangguan kesehatan yang disebabkan oleh penyakit atau cedera pada kerongkongan, paru-paru, dan organ tubuh lain yang ada di dada. Salah satu penyakit yang paling sering ditangani dengan operasi toraks adalah kanker paru-paru [1]. Kondisi kesehatan dan fungsi paru-paru pasien setelah menjalani operasi menjadi konsentrasi dalam mendiagnosis harapan hidup mereka.

Penanganan dini sebenarnya dapat dilakukan untuk menekan tingkat kematian pasca operasi toraks, salah satunya dengan mengumpulkan data yang berupa informasi tentang pasien kanker paru-paru pasca operasi toraks. Namun masalah yang timbul adalah data yang diperoleh merupakan data berdimensi tinggi yang memiliki banyak atribut dan tidak bisa menghasilkan informasi yang akurat.

Oleh karena itu perlu adanya *feature selection* atau seleksi pada atribut [2], yang bertujuan untuk mengurangi beban pada proses klasifikasi, salah satunya dengan metode *evolutionary data mining*. *Evolutionary Data Mining*, atau *Genetic Data Mining* adalah istilah umum untuk setiap data mining yang menggunakan *Evolutionary Algorithm*. Meskipun dapat digunakan untuk data mining dari urutan DNA, namun hal tersebut tidak terbatas pada konteks biologi dan dapat digunakan dalam setiap skenario prediksi berdasarkan klasifikasi [3]. *Evolutionary data mining* dapat menangani masalah dan memangkas waktu pengerjaan pada data berdimensi tinggi, sehingga informasi akan lebih cepat didapatkan. Proses reduksi pada atribut yang digunakan *Genetic Algorithm for Feature Selection*. Kemudian, proses klasifikasi bertujuan untuk mengklasifikasikan data yang diambil dari *website UCI Machine Learning Repository: Thoracic Surgery Data Set* [4] [5] dengan menggunakan metode *Naïve Bayes Classifier* [6], sehingga dapat menjawab apakah seseorang pengidap kanker paru-paru mimiliki harapan hidup pasca operasi toraks atau tidak berdasarkan data *training* yang didapat dan berapa nilai akurasi terbaik atas informasi data *testing* yang diolah. Studi yang pernah membahas harapan hidup pasca operasi toraks adalah Maciej Zieba et al. dalam "*Boosted SVM for extracting rules from imbalanced data in applicationto prediction of the post-*

*operative life expectancy in thelung cancer patients* dan *Dimensionality Reduction Using Genetic Algorithms* ", dalam studi tersebut dijelaskan bahwa hasil performansi dari metric *Gmean* yang digunakan menghasilkan nilai 65,73 [7]. *Geometric Mean* atau *Gmean* merupakan metrik yang mengukur keseimbangan antara kinerja klasifikasi pada kelas mayoritas dan minoritas. Nilai *Gmean* yang rendah adalah indikasi kinerja yang buruk dalam klasifikasi kasus positif bahkan jika kasus negatif diklasifikasikan dengan benar. Ukuran ini penting untuk menghindari kecocokan kelas negatif dan kecocokan kelas positif [8].

## 2. Studi Terkait

### 2.1 *Evolutionary Algorithms* (EAs)

*Evolutionary Algorithms* (EAs) merupakan algoritma-algoritma optimasi yang berbasis evolusi biologi yang ada di dunia nyata. Sehingga EAs merupakan algoritma yang mengimplementasikan abstraksi dari *Evolutionary Computation* (EC) [9] [10]. EAs atau algoritma evolusi bekerja dengan cara menciptakan serangkaian aturan acak untuk diperiksa terhadap data set pelatihan. Dengan kata lain EAs akan membangkitkan, menguji, dan berusaha memperbaiki sekumpulan kandidat solusi sampai ditemukannya satu solusi yang bisa diterima.

Pada awalnya, EAs dimulai dengan sekumpulan kandidat solusi (individu) yang disebut populasi. Populasi awal akan ber-evolusi menjadi populasi baru melalui serangkaian generasi (iterasi). Pada akhir generasi, EAs mengembalikan satu individu anggota populasi yang terbaik sebagai solusi untuk masalah yang dihadapi [3].

### 2.2 *Genetic Algorithm*

*Genetic Algorithm* (GA) adalah algoritma pencarian yang didasarkan pada mekanisme seleksi alamiah dan genetika alamiah. Pada awalnya, GA memang digunakan sebagai algoritma pencarian parameter-parameter optimal [11]. Tetapi, dalam perkembangannya, GA bisa diaplikasikan untuk berbagai masalah lain, seperti *learning*, peramalan, pemrograman otomatis, dan sebagainya. Terdapat lima komponen utama pada GA, yaitu : skema pengkodean atau representasi kromosom, nilai fitness, seleksi orang tua, pindah silang atau rekombinasi (*crossover*), dan Mutasi [12].

**Tabel 1 Rangkuman Spesifikasi Teknis GA [13]**

| | |
|---|---|
| Representasi | Pengodean (*binary encoding*) |
| Seleksi orang tua | Proporsional terhadap nilai *fitness* |
| Rekombinasi | *N-point* atau seragam (*uniform*) |
| Mutasi | Pembalikan bit dengan dengan probabilitas tetap dan bersifa bebas (*independent*) pada masing-masing bit. |
| Seleksi *survivor* | Semua individu baru menggantikan semua individu lama (*general replacement*) |
| Ciri khusus | Lebih menekankan pada rekombinasi |

Berikut adalah *pseudo-code* GA [13]:

```
Bangkitkan populasi awal, N kromosom
Loop sampai Kondisi Berhenti terpenuhi
    Dekodekan kromosom ke dalam individu
    Evaluasi individu
    Seleksi pasangan-pasangan orangtua
    Rekombinasi dengan probabilitas Pc
    Mutasi dengan probabilitas Pm
    Penggantian populasi
End
```

**Gambar 1 Pseudo-code GA**

### 2.3 *Naïve Bayes Classifier*

Klasifikasi bayes mengasumsikan bahwa suatu *feature* tidak berpengaruh dengan adanya *feature* lain. Hipotesis dalam teorema Bayes merupakan label kelas yang menjadi target dalam pemetaan dalam klasifikasi, dan bukti merupakan *features* yang menjadi masukan dalam model klasifikasi. Jika E adalah masukan yang berisi

*feature* dan H adalah label kelas, maka Naïve Bayes dituliskan dengan P(H|E).Notasi tersebut berarti probabilitas label kelas H didapatkan setelah *features* X diamati [14].

Berikut formula *naïve bayes* untuk klasifikasi :

$$P(H|E) = \frac{P(H)\Pi_{i=1}^{q} P(E_i|H)}{P(E)} \quad (1)$$

Dimana :

*P(E|H)* = Probabilitas data *feature E* pada kelas *H*

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

$P(H)\Pi_{i=1}^{q} P(E_i|H)$ = Probabilitas independen kelas *H* dari semua *feature* dalam *X*

Umumnya bayes mudah dihitung untuk *feature* bertipe kategorikal. Perlakuan untuk data numerik akan sedikit berbeda dengan data kategorikal. Salah satunya adalah dengan mengasumsikan bentuk tertentu dari distribusi menggunakan data training. Distribusi Gaussian biasanya dipilih untuk mempresentasikan *conditional probability feature continuous* pada sebuah kelas P ($X_i$ |Y). Distribusi Gaussian dikarakteristikan dengan dua parameter: rata rata (μ) dan variansi ($\sigma 2$), x adalah nilai *feature* pada data yang akan diprediksi [15]. Persamaan distribusi gauss adalah :

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} exp - \frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2} \quad [15](2)$$

Dimana :

μ = rata-rata

σ = standar deviasi

$x_i$ = data ke i

*Naïve bayes* melakukan klasifikasi terhadap data latih, sehingga diperoleh kelas prediksi yang selanjutnya dihitung akurasi klasifikasi menggunakan *confusion matrix*. Nilai *fitness* tertinggi dalam proses pelatihan data dipilih sebagai standar ukuran kualitas evaluasi individu berikutnya.

**Tabel 2 *Confusion Matrix* [8]**

| | ***Classified positive*** | ***Classified negative*** |
|---|---|---|
| ***Actual positive*** | TP | FN |
| ***Actual negative*** | FP | TN |

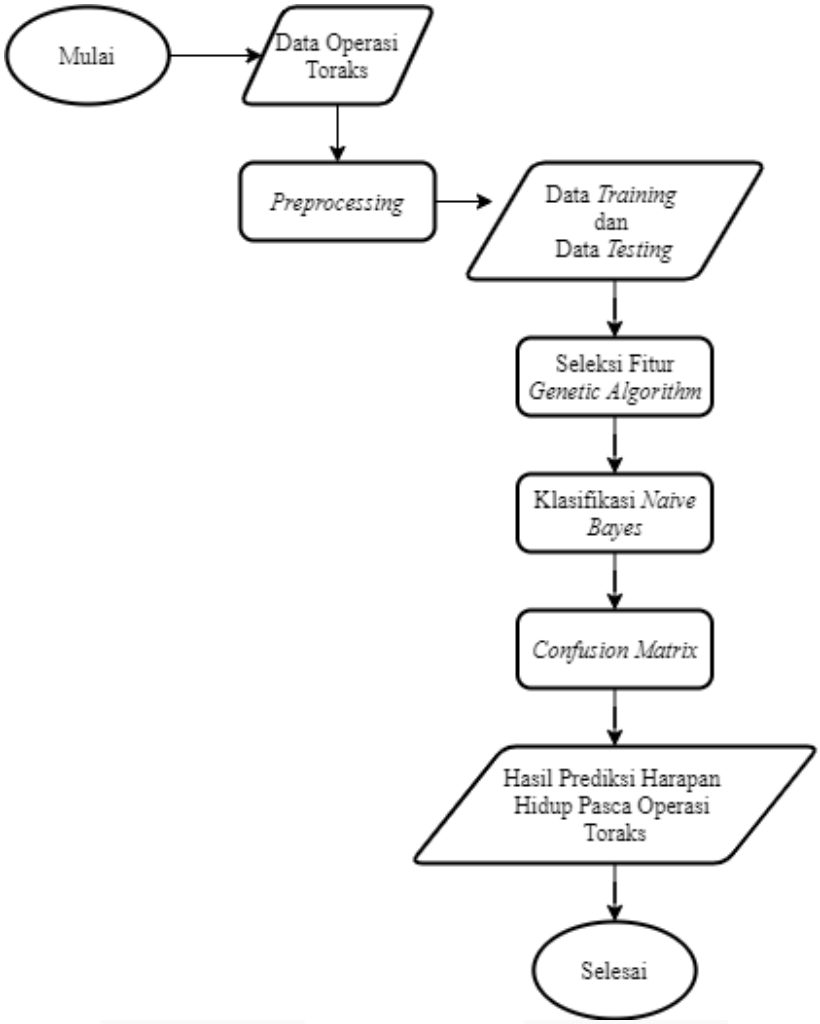
Berikut rumus untuk menghitung akurasi oleh *confusion matrix* :

$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad [16](3)$$

TP (*True Positive*) adalah jumlah kasus positif yang diklasifikasikan dengan benar. FN (*False Negative*) adalah jumlah kasus positif yang salah diklasifikasikan sebagai negatif. FP (*False Positive*) adalah jumlah kasus negatif yang salah diidentifikasi sebagai kasus positif dan TN (*True Negative*) adalah jumlah kasus negatif yang diklasifikasikan dengan benar [16].

## 3. Rancangan Sistem

Pada bagian ini akan dijelaskan tentang alur kerja sistem yang dibuat pada penelitian ini. Data operasi toraks diklasifikasi menggunakan *naïve bayes* kemudian GA akan melakukan seleksi pada fitur, sehingga mendapatkan hasil prediksi harapan hidup pasca operasi toraks pada pasien penderita kanker paru-paru.




**Gambar 2 Alur Deskripsi Sistem**

### 3.1 Data Set

Tahap pertama yang dilakukan pada penelitian tugas akhir ini adalah pengumpulan *dataset* yaitu data operasi toraks. *Dataset* tersebut diambil dari *website UCI Machine Learning Repository: Thoracic Surgery Data Set* [4]. Data tersebut yang digunakan untuk serangkaian proses klasifikasi harapan hidup pasca operasi toraks. Berikut spesifikasi data operasi toraks yang digunakan :

**Tabel 3 Spesifikasi *dataset* operasi toraks**

| Data set | Jumlah atribut | Jumlah kelas | Jumlah sampel | |
|---|---|---|---|---|
| *Thoracic Surgery* | 17 | 2 | 70 *death within one year after surgery* | 400 *survival* |

Terdapat 16 atribut yaitu DGN, PRE4, PRE5, PRE6, PRE7, PRE8, PRE9, PRE10, PRE11, PRE14, PRE17, PRE19, PRE25, PRE30, PRE32, dan AGE serta terdapat 1 atribut kelas yaitu Risk1Y. Berikut contoh data yang digunakan :

**Tabel 4 Data Set *Thoracic Surgery***

| No. | DGN | PRE4 | PRE5 | PRE6 | PRE7 | PRE8 | PRE9 | PRE10 | PRE11 | PRE14 | PRE17 | PRE19 | PRE25 | PRE30 | PRE32 | AGE | Risk1Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | DGN2 | 2,88 | 2,16 | PRZ1 | F | F | F | T | T | OC14 | F | F | F | T | F | 60 | F |
| 2. | DGN3 | 3,4 | 1,88 | PRZ0 | F | F | F | F | F | OC12 | F | F | F | T | F | 51 | F |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3. | DGN3 | 2,76 | 2,08 | PRZ1 | F | F | F | T | F | OC11 | F | F | F | T | F | 59 | F |
| 4. | DGN3 | 3,68 | 3,04 | PRZ0 | F | F | F | F | F | OC11 | F | F | F | F | F | 54 | F |
| 5. | DGN3 | 2,44 | 0,96 | PRZ2 | F | T | F | T | T | OC11 | F | F | F | T | F | 73 | T |
| … | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… |
| … | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… |
| 469. | DGN3 | 1,96 | 1,68 | PRZ1 | F | F | F | T | T | OC12 | F | F | F | T | F | 79 | F |
| 470. | DGN3 | 4,72 | 3,56 | PRZ0 | F | F | F | F | F | OC12 | F | F | F | T | F | 51 | F |

Deskripsi data :

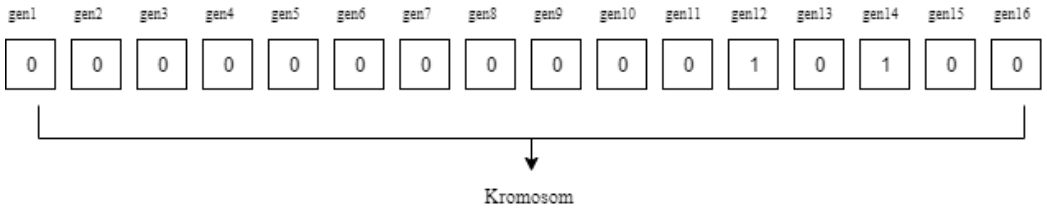
- DGN : Diagnosis - kombinasi spesifik kode ICD-10 untuk tumor primer dan sekunder serta multipel jika ada {DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1}
- PRE4 : FVC (*Forced Vital Capacity*) atau kapasitas vital yang di paksakan {*numeric*}
- PRE5 : Volume yang telah dihembuskan pada akhir detik pertama dari ekspirasi paksa - FEV1 {*numeric*}
- PRE6 : Status kinerja - skala Zubrod {PRZ2,PRZ1,PRZ0}
- PRE7 : Nyeri sebelum operasi {T,F}
- PRE8 : Hemoptisis sebelum operasi {T,F}
- PRE9 : Dispnea sebelum operasi {T,F}
- PRE10 : Batuk sebelum operasi {T,F}
- PRE11 : Kelemahan sebelum operasi {T,F}
- PRE14 : T dalam TNM klinis - ukuran tumor asli, dari OC11 (terkecil) hingga OC14 (terbesar) {OC11,OC14,OC12,OC13}
- PRE17 : DM tipe 2 - diabetes mellitus {T,F}
- PRE19 : MI hingga 6 bulan {T,F}
- PRE25 : PAD (*Peripheral Arterial Diseases*) atau penyakit arteri perifer {T,F}
- PRE30 : Perokok{T,F}
- PRE32 : Asma {T,F}
- AGE : Umur saat di operasi{*numeric*}
- Risk1Yr : 1 tahun periode bertahan hidup - (T) nilai nilai jika meninggal {T,F}

### 3.2 Seleksi Fitur *Genetic Algorithm*

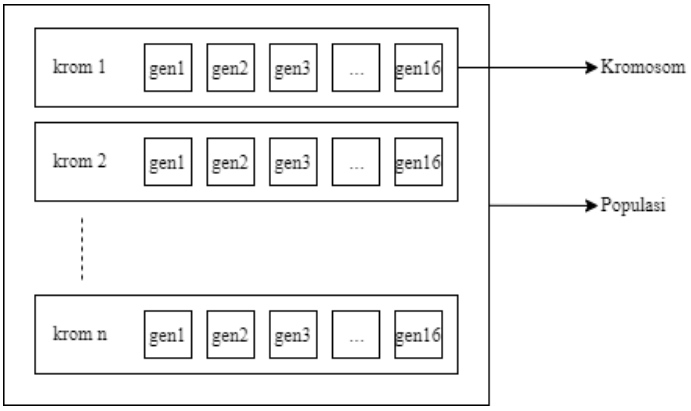
Setelah mendapatkan *fitness function* yang diperoleh dari nilai akurasi *naïve bayes*, kemudian GA melanjutkan untuk melakukan seleksi fitur.

1. Inisialisasi populasi

   Pada penelitian ini skema pengodean untuk setiap kromosom menggunakan *Binary Encoding*. Setiap kromosom dibangkitkan secara acak dengan nilai 0 dan 1. Setiap kromosom pada populasi merepresentasikan sebuah solusi kandidat terhadap masalah seleksi fitur. Jika sebuah bit sama dengan 0 artinya fitur tersebut tidak akan terpilih, sedangkan jika bit sama dengan 1 maka fitur tersebut terpilih. Pada penelitian ini terdapat 16 atribut dan 1 atribut kelas. Kolom atribut dijadikan kromosom dan kumpulan beberapa kromosom menjadi suatu populasi.




**Gambar 3 Contoh Kromosom pada Penelitian**




**Gambar 4 Contoh Populasi pada Penelitian**

2. Fungsi *fitness*

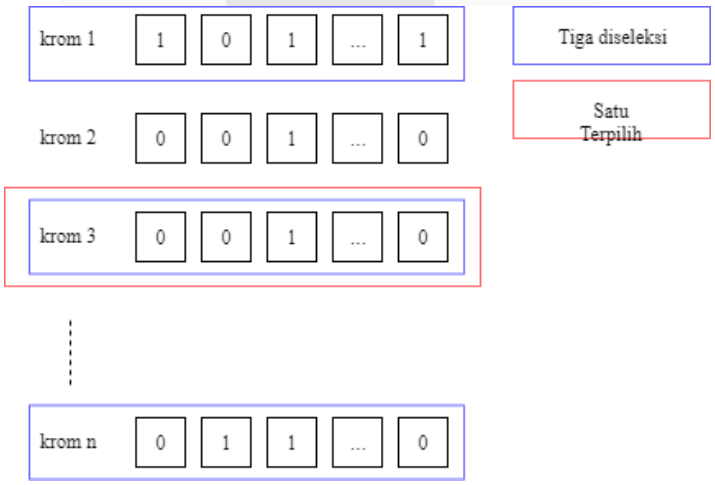
   Setiap proses pada GA memerlukan nilai *fitness*, suatu individu atau kromosom dievaluasi berdasarkan suatu fungsi tertentu sebagai ukuran nilai kualitasnya [17]. Fungsi tersebut dikenal sebagai fungsi *fitness*. Dalam penelitian ini evaluasi *fitness* diperoleh dari hasil perhitungan algoritma *naïve bayes* yang digunakan untuk mengevaluasi individu. *Naïve bayes* melakukan klasifikasi terhadap data latih, sehingga diperoleh kelas prediksi yang selanjutnya dihitung akurasi klasifikasi menggunakan *confusion matrix*. Nilai *fitness* tertinggi dalam proses pelatihan data dipilih sebagai standar ukuran kualitas evaluasi individu berikutnya. Pada penelitian ini nilai *fitness* memiliki *range* antara 0 sampai 1.

3. Seleksi orang tua

   Metode seleksi yang digunakan dalam penelitian ini adalah *tournament selection* dengan nilai *tournament* sama dengan 3. Nilai *tournament* merepresentasikan jumlah kromosom yang diambil secara acak dari populasi. Setiap kromosom memiliki pasangan untuk proses *crossover* [18].

Bangkitkan populasi awal, 20 kromosom
***Loop*** sampai Terpilih 20 kromosom
    Ambil 3 kromosom secara acak
    Pilih 1 kromosom dengan nilai *fitness* terbaik
    Penggantian populasi
***End***

**Gambar 5 *Pseudo-code* Proses Seleksi Orang Tua [19]**

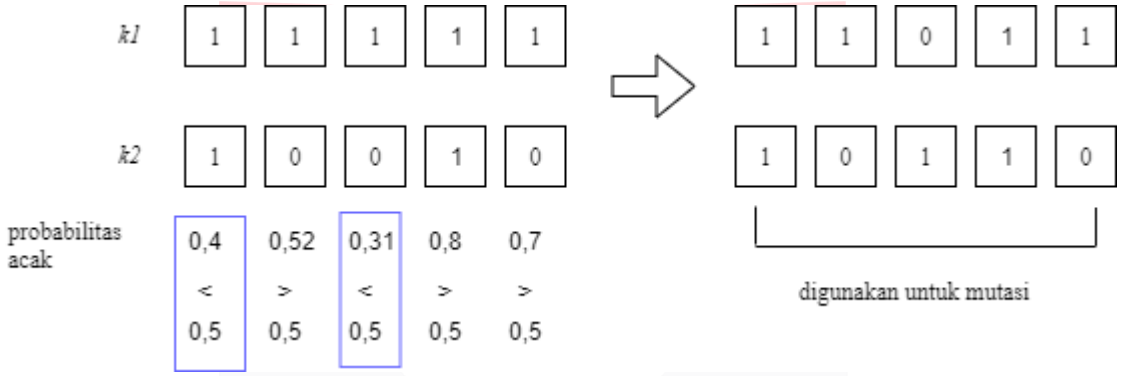



**Gambar 6 Contoh Seleksi *Tournament***

4. Rekombinasi (*crossover*)

Setelah mendapatkan populasi baru dari hasil seleksi orang tua, proses selanjutnya yaitu rekombinasi. Rekombinasi dilakukan pada setiap pasang kromosom *k1* dengan *k2*, *k3* dengan *k4*, … , *k19* dengan *k20*. Setiap pasangan kromosom dan pasangan *gen* akan diberi nilai acak dengan *range* 0 sampai 1, dimana nilai acak pasangan kromosom digunakan untuk menentukan apakah pasangan tersebut perlu dilakukan *crossover* atau tidak, sedangkan nilai acak yang diberikan pada setiap *gen* digunakan untuk *crossover*. *Crossover* akan dilakukan apabila nilai acak kurang dari nilai probabilitas *crossover*. Pada penelitian ini nilai probabilitas *crossover* yang ditentukan yaitu 0,5. Probabilitas *crossover independent* yang ditentukan yaitu 0,5 untuk penyilangan setiap pasang *gen* yang nilai acaknya kurang dari 0,5. Pada penelitian ini setiap pasang kromosom beserta gen diberi nilai random. Berikut adalah contoh proses *crossover* yang dilakukan :

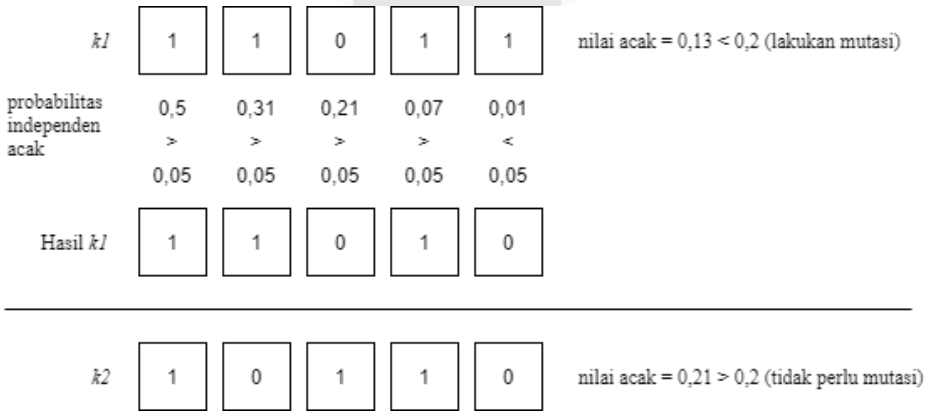
*k1* dan *k2* → nilai acak = 0,43
0,43 < 0,5 → lakukan *crossover*




**Gambar 7 Contoh Proses *Crossover***

5. Mutasi

Dalam dunia nyata mutasi didefinisikan sebagai pemutusan atau penggantian molekul DNA yang terdapat di dalam inti sel makhluk hidup dan berisi semua informasi genetis [20]. Pada GA mutasi bertujuan untuk menggantikan nilai gen pada kromosom [13]. Pada penelitian ini proses mutasi hampir sama dengan *crossover*. Setelah mendapatkan populasi baru yaitu kumpulan kromosom *offspring* (anak), setiap kromosom dan *gen* diberi nilai acak antara 0 sampai 1. Nilai acak kromosom tersebut digunakan untuk menetukan apakah kromosom tersebut perlu dimutasi atau tidak, sedangkan nilai acak setiap *gen* digunakan untuk melakukan mutasi pada *gen* tersebut. Apabila nilai acak yang diberikan kepada *k* kromosom kurang dari nilai probabilitas mutasi maka kromosom tersebut perlu dimutasi. Probabilitas mutasi yang ditentukan pada penelitian ini yaitu 0,2. Probabilitas *independent* mutasi yang ditentukan yaitu 0,05 untuk memutasi setiap *gen* yang nilai acaknya kurang dari 0,05.




**Gambar 8 Contoh Proses Mutasi**

6. Seleksi *survivor*

Generasi yang ditentukan pada penelitian ini yaitu 30, jadi terdapat perulangan sebanyak 30 kali dalam proses seleksi fitur GA. Poulasi lama (20 kromosom) digantikan dengan populasi baru dari hasil *tournament*, *crossover*, dan mutasi. Setiap perulangan dicatat *fitness* terbaik beserta kromosomnya.

## 4. Pengujian dan Analisis

Dalam pengujian ini seleksi fitur *genetic algorithm* akan memperhatikan jumlah individu atau kromosom, ukuran populasi, serta generasi pengujian pada masing-masing data. Sedangkan pengujian untuk klasifikasi akan memperhatikan parameter *cross validation*. Jumlah total individu pada penelitian ini adalah 470 dan nilai *k* pada *cross validation* yang ditentukan dalam pengujian adalah 5 dan 10. Sehingga pada *k=5* proporsi data *training* dan data *testing* adalah 376 data *training* dan 94 data *testing*. Sedangkan, pada *k=10* proporsinya adalah 423 data *training* dan 47 data *testing*. Terminasi yang dilakukan pada pengujian ini adalah apabila nilai akurasi tidak berubah sebanyak 4 iterasi, maka akan dilakukan terminasi.

### 4.1 *Genetic Algorithm for Feature Selection* dengan Klasifikasi *Naïve Bayes*

Pada bagian ini menampilkan hasil penelitian dari skenario seleksi fitur *genetic algorithm*. Skenario uji yang pertama yaitu jumlah populasi yang digunakan adalah 20, 40, 60, 80, dan 100 dengan merubah nilai *k* pada *cross validation*, masing-masing pada setiap *n*-populasi diberikan nilai *k*=5 dan *k*=10. Adapun parameter GA dan *naïve bayes* di-set pada nilai *default* yaitu nilai *crossover_proba*=0.5, *mutation_proba*=0.2, dan *tournament_size* (seleksi) = 3. Selanjutnya skenario uji yang dilakukan yaitu merubah parameter yang ada pada *genetic algorithm* yaitu nilai *crossover* dan mutasi berdasarkan hasil akurasi terbaik dari pengujian berdasarkan populasi. Nilai *crossover_proba* yang dipakai yaitu {0.3, 0.5, 0.7} dan *mutation_proba* yaitu {0.2, 0.4, 0.6}. Setelah itu dilakukan skenario uji acak dimana nilai semua parameter di ubah secara acak.

**Tabel 5 Hasil Seleksi Fitur *Genetic Algorithm* Skenario Pertama**

| Data Set | Ukuran Populasi | Akurasi (%) | |
|---|---|---|---|
| | | ***k = 5*** | ***k = 10*** |
| | | **Proporsi data 376:94** | **Proporsi data 423:47** |
| ***Thoraric Surgery*** | 20 | 82,340% | 83,617% |
| | 40 | 85,106% | 85,106% |
| | 60 | 85,106% | 85,106% |
| | 80 | 85,106% | 85,106% |
| | 100 | 85,106% | 85,106% |

Berdasarkan Tabel 5 hasil dari seleksi fitur *genetic algorithm* dengan Klasifikasi *naïve bayes* memberikan akurasi yang hampir sama di setiap pengujian yaitu 85,106%, hanya ada sedikit perbedaan ketika nilai populasi = 20 pada *k*=5 dan *k*=10 masing-masing akurasinya 82,340% dan 83,617%. Dari hasil pengujian seleksi fitur GA dan nilai akurasi yang diberikan pada skenario uji pertama dengan akurasi terbaik 85,106%.

**Tabel 6 Hasil Seleksi Fitur *Genetic Algorithm* Skenario Kedua (populasi=60 dan *k*=10)**

| | Akurasi (%) | | |
|---|---|---|---|
| | ***crossover_proba* = 0,3** | ***crossover_proba* = 0,5** | ***crossover_proba* = 0,7** |
| ***mutation_proba* = 0,2** | 85,106% | 85,106% | 85,106% |

| | | | |
|---|---|---|---|
| ***mutation_proba* = 0,4** | 85,106% | 85,106% | 85,106% |
| ***mutation_proba* = 0,6** | 85,106% | 85,106% | 85,106% |

Berdasarkan Tabel 6 akurasi dari parameter GA yang diubah didapatkan akurasinya sama yaitu 85,106%. Selanjutnya skenario uji ketiga yang dilakukan pada seleksi fitur GA dengan mengubah semua parameter secara acak.

**Tabel 7 Hasil Seleksi Fitur *Genetic Algorithm* Skenario Ketiga (parameter acak)**

| Dataset | Populasi | Akurasi (%) |
|---|---|---|
| | | ***k* = 7** |
| | | ***crossover_proba* = 0,3** |
| | | ***mutation_proba* = 0,6** |
| **Thoracic Surgery** | 60 | 85,319% |

Berdasarkan Tabel 7 akurasi yang didapatkan meningkat dan akurasi terbaik yaitu 85,319%. Dari hasil pengujian seleksi fitur GA dan nilai akurasi yang didapatkan pada setiap skenario uji, dapat disimpulkan bahwa nilai akurasi terbaik pada sistem yang dijalankan pada penelitian ini adalah 85,319% dan peningkatan nilai akurasi berdasarkan parameter yang diubah tidak lebih dari 5%. Perubahan yang terlihat jelas dari setiap parameter yang diubah hanya pada jumlah iterasi yang dilakukan pada setiap pengujian.

### 4.2 Klasifikasi *Naïve Bayes* Tanpa GA untuk Seleksi Fitur

Pada bagian ini menampilkan hasil pengujian dari skenario klasifikasi *naïve bayes.* Dalam penelitian ini, klasifikasi yang dilakukan yaitu menggunakan *Gaussian Naïve Bayes.*

**Tabel 8 Klasifikasi Naive Bayes Tanpa GA untuk Seleksi Fitur**

| Data set | Akurasi (%) |
|---|---|
| ***Thoraric Surgery*** | 17,872% |

Berdasarkan Tabel 8 untuk pengujian dengan hanya klasifikasi *naive bayes* tanpa GA untuk fitur seleksi, akurasi yang didapatkan hanya 17,872%.

### 4.3 Analisis Hasil Klasifikasi dengan Seleksi Fitur dan Tanpa Seleksi Fitur

Berdasarkan dari kedua pengujian yang dilakukan yaitu pengujian tanpa seleksi fitur dan pengujian seleksi fitur *genetic algorithm*, maka seleksi fitur GA mampu memberikan hasil yang jauh lebih baik dari pada klasifikasi *naïve bayes* dengan seluruh fitur yang dipakai. Hal ini disebabkan karena seleksi fitur GA hanya akan menggunakan fitur terpilih berdasarkan nilai acak yang diberikan ketika pembangkitan populasi. Berdasarkan analisis tersebut, berikut merupakan tabel perbandingan akurasi dengan seleksi fitur dan tanpa seleksi fitur dengan klasifikasi *naïve bayes* :

**Tabel 9 Perbandingan Klasifikasi Tanpa Seleksi Fitur dan Klasifikasi Dengan Seleksi Fitur**

| Data set | Akurasi (%) | |
|---|---|---|
| | **Klasifikasi *Naïve Bayes*** | **Seleksi Fitur *Genetic Algorithm* dengan Klasifikasi *Naïve Bayes*** |
| ***Thoracic Surgery*** | 17,872% | 85,319% |

Sesuai Tabel 9 maka klasifikasi *naïve bayes* hanya mampu memberikan nilai akurasi 17,872%. Sedangkan seleksi fitur *genetic algorithm* dengan *naïve bayes* mampu memberikan nilai akurasi 85,319%.

### 4.4 Analisis Hasil Klasifikasi setelah Penambahan dengan replikasi *Record* Data

Setelah melakukan semua pengujian dan analisis terhadap klasifikasi *naïve bayes* dan seleksi fitur GA dengan mengubah seluruh parameter, pada langkah terakhir pengujian penelitian ini dilakukan penambahan *record* data dengan replikasi menggunakan metode *sampling* untuk melihat hasil akurasi yang didapatkan jika record data ditambahkan, berikut tabel hasil penambahan *record* data pada data set *Thoracic Surgery* menggunakan metode *sampling*:

**Tabel 10 Hasil Akurasi setelah Penambahan Record Data (*Class Survival* = 400)**

| ***Original data set by class survival*** | **Akurasi (%)** | | |
|---|---|---|---|
| | ***Class death* = 100** | ***Class death* = 300** | ***Class death* = 500** |
| ***Class survival* = 400** | 82,001% | 65,719% | 65,775% |

**Tabel 11 Hasil Akurasi setelah Penambahan Record Data (*Class Death* = 70)**

| ***Original data set by class death*** | **Akurasi (%)** | | |
|---|---|---|---|
| | ***Class survival* = 500** | ***Class survival* = 700** | ***Class survival* = 900** |
| ***Class death* = 70** | 87,895% | 91,039% | 92,886% |

**Tabel 12 Hasil Akurasi setelah Penambahan *Record* Data**

| **Penambahan *record* data (Replikasi)** | **Akurasi (%)** | |
|---|---|---|
| | ***Class survival* = 900** | ***Class survival* = 100** |
| | ***Class death* = 100** | ***Class death* = 900** |
| | 90,100% | 90,000% |

Berdasarkan tiga tabel diatas menunjukan penambahan record data menggunakan metode sampling dapat menurunkan dan menaikan akurasi sebelumnya yaitu 85,319%, dimana Tabel 10 menunjukan penambahan jumlah data untuk kelas death semakin banyak data dapat menurunkan akurasi mencapai 65,719% pada jumlah kelas death = 300. Sedangkan pada Tabel 11 menunjukan semakin banyak jumlah data untuk kelas survival dapat menaikan akurasi mencapai 92,886% pada jumlah kelas survival = 900. Namun pada Tabel 12 menunjukan akurasi yang seimbang ketika data di set untuk survival atau death sama-sama pada proporsi 900:100 atau sebaliknya dengan akurasi 90.000%. Jika dilihat dari peningkatan akurasi setalah penambahan *record* data kenaikan akurasi terjadi diakibatkan oleh jumlah TP (*True Positive*) dan TN (*True Negative*) pada formula *confusion matrix* jumlah kasus positif dan negatif yang diklasifikasikan dengan benar memiliki nilai yang besar.

## 5. Kesimpulan dan Saran

### 5.1 Kesimpulan

Penanganan dini yang dilakukan untuk menekan tingkat kematian pasien kanker paru-paru pasca operasi toraks, dengan mengumpulkan data berupa informasi tentang pasien pasca operasi toraks menimbulkan masalah baru yaitu data berdimensi tinggi yang memiliki banyak atribut dan tidak bisa menghasilkan informasi yang akurat. Oleh karena itu, diperlukan skema komputasi yang dapat mereduksi dimensi pada data tersebut. Dalam hal

ini, proses reduksi bertujuan untuk meringankan beban komputasi pada klasifikasi, proses reduksi yang digunakan yaitu seleksi fitur *genetic algorithm*. Kemudian metode klasifikasi *naïve bayes* digunakan untuk melakukan proses klasifikasi harapan hidup pasca operasi toraks. Klasifikasi *naïve bayes* mampu memberikan hasil performansi untuk harapan hidup pasca operasi toraks 17,872%. Sedangkan seleksi fitur *genetic algorithm* mampu memberikan hasil 85,319%.

Berdasarkan hasil uji, dengan acuan 0 sampai 100% peningkatan akurasi dari klasifikasi *naïve bayes* kemudian menggunakan seleksi fitur *genetic algorithm* yaitu 67,447%. Dari hasil tersebut penggunaan GA untuk fitur seleksi mampu memberikan dampak yang besar, dikarenakan penggunaan atribut yang paling optimal ketika klasifikasi sehingga meningkatkan akurasi.

Penambahan *record* data dengan replikasi menggunakan metode *sampling* yang dilakukan pada pengujian mempengaruhi nilai akurasi, dimana pada uji yang dilakukan setelah penambahan *record* tersebut akurasi dapat meningkat jika jumlah kasus positif (*death*) dan kasus negatif (*survive*) yang diklasifikasikan dengan benar memiliki nilai yang besar. Jadi semakin besar jumlah TP + TN akan meningkatkan akurasi yang didapatkan. Namun, peningkatan akurasi tersebut tetap dipengaruhi oleh penggunaan GA untuk fitur seleksi. Akurasi terbaik yang didapatkan dari skenario uji ini adalah 92,886%.

### 5.2 Saran

Setelah proses prediksi harapan hidup pasca operasi toraks pada pasien penderita kanker paru paru ini, penulis menemukan saran yang dapat dilakukan, yaitu berdasarkan pola yang didapat dari proses prediksi pada data set *Thoraric Surgery* hasil akurasi mengalami kenaikan yang cukup besar. Untuk itu, saran yang dapat dilakukan untuk penelitian selanjutnya yaitu mencoba menggunakan data operasi toraks dengan jumlah atribut yang lebih banyak dari data set *Thoracic Surgery* yang digunakan pada penelitian ini, atau menggunakan metode klasifikasi lain.

## Daftar Pustaka


[1] C. E. Niluh Gede Yasmin, Keperawatan Medikal Bedah: Klien Dengan Gangguan Sistem Pernapasan, Jakarta: Penerbit Buku Kedokteran EGC, 2002.

[2] M. Taradeh, M. Mafarja, A. A. Heidari, H. Faris, I. Aljarah, S. Mirjalili and H. Fujita, "An evolutionary gravitational search-based feature selection," *Information Sciences*, vol. 497, pp. 219-239, 2019.

[3] Suyanto, Soft Computing: Membangun Mesin Ber-IQ Tinggi, Bandung: Informatika, 2008.

[4] L. Marek, P. Konrad, R. Adam and K. Jerzy, "Thoracic Surgery Data Data Set," National Science Foundation, 13 November 2013. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data. [Accessed 15 July 2020].

[5] A. J. M, L. D. B and C. Priti, Knowledge Discovery Using Associative Classification for Heart Disease Prediction, Berlin: Springer, Berlin, Heidelberg, 2013.

[6] S. Raschka and V. Mirjalili, Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2, Birmingham: Packt Publishing Ltd, 2019.

[7] e. a. Maciej Zieba, "Boosted SVM for extracting rules from imbalanced data in applicationto prediction of the post-operative life expectancy in thelung cancer patients," *Applied Soft Computing*, 2013.

[8] J. S Akosa, "Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data," in *Proceedings of the SAS Global Forum*, 2017.

[9] J. E. Jackson, A User's Guide to Principal Components, John Wiley & Sons, 2005.

[10] S. Mirjalili, "Genetic algorithm," in *Evolutionary algorithms and neural networks*, Springer, 2019, pp. 43-55.

[11] Suyanto, Artificial Intelligence, Bandung: Informatika, 2014.

[12] Ding, Z. Ye, B. Kui and Weihong, "Feature selection based on hybridization of genetic algorithm and competitive swarm optimizer," *Soft Computing*, pp. 1-10, 2020.

[13] Suyanto, Evolutionary Computation, Bandung: Informatika Bandung, 2008.

[14] J. H. &. M. Kamber, Data Mining: Concepts and Techniques, San Francisco: Elsevier, 2006.

[15] H. j. Ali and T. Mohammad, "A non-parametric mixture of Gaussian naive Bayes," in *Artificial Intelligence and Signal Processing Conference (AISP)*, 2017.

[16] Visa, R. Sofia, R. Brian, L. V. D. K. Anca and Esther, "Confusion Matrix-based Feature Selection.," *MAICS*, vol. 710, pp. 120-127, 2011.

[17] H. L. M. Dash, "Feature Selection for Classification," in *Intelligent Data Analysis 1 (1997) 131–156*, Singapore, 1997.

[18] W. Lee and H.-Y. Kim, "Genetic algorithm implementation in Python," in *Fourth Annual ACIS International Conference on Computer and Information Science (ICIS'05)*, IEEE, 2005, pp. 8-11.

[19] C. Manuel, "manuel-calzolari/sklearn-genetic: sklearn-genetic," April 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3348077.

[20] G. P. S. Luh, I. Made Adi Bhaskara and M. Sudarma, "Optimization of Feature Selection Using Genetic Algorithm with Naïve Bayes Classification for Home Improvement Recipients," *International Journal of Engineering and Emerging Technology*, vol. III, no. 1, 2018.