# Membandingkan Seleksi Variabel Pada Data *Microarray* Menggunakan *Important Variable Value* dan *Genetic Algorithm* (Studi Kasus *Lung Cancer Dataset* dan *Prostate Cancer Dataset*)

Diana Nurlaily [(1)], Farida Nur Hayati [(2)], Elly Pusporani [(3)]
Institut Teknologi Kalimantan
Jl, Soekarno-Hatta Km. 15 Telp./Fax. (0542) 8530801 Balikpapan 76127
e-mail: diana.nurlaily@lecturer.itk.ac.id, farida.nur@lecturer.itk.ac.id dan
elly.pusporani@ lecturer.itk.ac.id

**ABSTRAK**

Teknologi DNA microarray menarik minat yang luar biasa di kalangan komunitas ilmiah maupun kalangan industri. Ukuran sampel kecil dengan dimensi tinggi adalah tantangan utama untuk analisis data microarray. Banyak penelitian yang berkaitan dengan data microarray misalnya menyelidiki mekanisme genetik kanker, mengklasifikasikan jenis kanker atau membedakan antara jaringan kanker dan non-kanker. Penelitian tersebut bertujuan untuk menghasilkan kesimpulan dan interpretasi yang bermanfaat dari kumpulan data yang kompleks. Dalam penelitian ini, data yang digunakan adalah data *lung cancer* sebanyak 24257 variabel dan data *prostate cancer* sebanyak 12626 variabel. Data tersebut dianalisis dengan dua metode seleksi variabel yaitu *important variable value* dan *genetic algorithm* untuk meningkatkan akurasi klasifikasi data. *Important variable value* merupakan metode seleksi variabel dimana variabel dipilih berdasarkan tingkat kepentingan variabel tehadap model klasifikasi. Sedangkan metode *genetic algoritm* memilih variabel berdasarkan *fitness value*, kombinasi dan *cross over* sehinga menghasilkan kombinasi kumpulan variabel baru. Berdasarkan hasil analisis seleksi variabel pada data *lung cancer*, didapatkan jumlah variabel terpilih sebanyak 112 variabel dengan metode *important variable value*. Sedangkan metode *genetic algorithm* didapatkan jumlah variabel terpilihnya sebanyak 12266 variabel. Pada data *prostate cancer*, didapatkan jumlah variabel terpilih sebanyak 299 variabel dengan metode important variable important. Sedangkan metode *genetic algorithm* didapatkan jumlah variabel terpilihnya sebanyak 6359 variabel.
***Kata kunci :*** *Microarray, important variable value, genetic algorithm, seleksi variabel*

***ABSTRACT***

*DNA microarray technology is attracting tremendous interest in both the scientific community and industry. Small sample size with high dimensions is a significant challenge for microarray data analysis. Many studies are related to microarray data, for example, investigating the genetic mechanism of cancer, classifying cancer types, or distinguishing between cancerous and non-cancerous tissues. The study aims to generate valuable conclusions and interpretations from complex data sets. In this study, the data used were lung cancer data as many as 24257 variables and prostate cancer data as many as 12626 variables. The data were analyzed using two variable selection methods, namely important variable value and genetic algorithm, to increase the accuracy of data classification. Important variable value is a variable selection method where the variable is selected based on the level of importance of the variable to the classification model. In comparison, the genetic algorithm method selects variables based on fitness values, combinations, and cross-overs to produce new combinations of variables. Based on the variable selection analysis results on lung cancer data, the number of selected variables was 112 variables using the important variable value method. In comparison, the genetic algorithm method obtained the number of selected variables as many as 12266 variables. In prostate cancer data, the number of selected variables is 299 variables using the important variable important method. In contrast, the genetic algorithm method obtained the number of selected variables as many as 6359 variables.*
***Keywords :*** *Feature selection, microarray, important variable value, genetic algorithm*

## 1. PENDAHULUAN

Teknologi DNA *microarray* menarik minat yang luar biasa baik di kalangan komunitas ilmiah maupun kalangan industri. Kemampuan mengukur secara bersamaan aktivitas dan interaksi ribuan gen, menjadikannya sebagai pengetahuan baru mengenai mekanisme sistem kehidupan. Data microarray sendiri mempunyai ciri-ciri yaitu sampel kecil, dimensi tinggi, noise tinggi, redudansi tinggi dan distribusi kelas tidak seimbang (Wang & Simon, 2011). Apabila data tersebut dianalisis dengan metode yang tepat, maka akan dihasilkan pengetahuan baru yang bermanfaat dibidang tertentu misalnya biologi dan kedokteran. Banyak penelitian yang telah dirancang berkaitan dengan data microarray misalnya untuk menyelidiki mekanisme genetik kanker, dan untuk mengklasifikasikan berbagai jenis kanker atau membedakan antara jaringan kanker dan non-kanker. Semua penelitian ini bertujuan untuk menghasilkan kesimpulan dan interpretasi yang bermanfaat dari kumpulan data yang kompleks (Hira & Gillies, 2015).

Ukuran sampel kecil dengan dimensi tinggi adalah tantangan utama analisis menggunakan data *microarray*. Dimensi data *microarray* yang dapat mencapai 450.000 variabel membuat analisis menjadi kurang efektif untuk diproses dalam program komputer. Berdasarkan permasalahan tersebut maka perlu dilakukan analisis dengan mengurangi dimensi/variabel. Hal yang biasanya dilakukan untuk mengurangi dimensi suatu data *microarray* adalah menggunakan seleksi variabel. Seleksi variabel bekerja dengan cara menghilangkan variabel yang berlebih dan kurang relevan. Beberapa metode seleksi variabel yang dapat digunakan antara lain adalah *important value*, *Filter method* dan *Genetic algoritm*.

Penelitian tentang seleksi variabel data *microarray* telah banyak dilakukan antara lain oleh oleh Hambali, Oladele, & Adewole (2020) tentang seleksi variabel pada data kanker *microarray*, menghasilan kesimpulan bahwa teknik seleksi variabel *hybrid* dapat meningkatkan performa akurasi yang lebih baik dari model klasifikasi penyakit kanker. Selain itu Rasmita Dash (2020) meneliti seleksi variabel dan klasifikasi data microarray berdasarkan ranking variabel.

Dalam penelitian ini, data yang digunakan adalah data kanker paru-paru sebanyak 24257 Variabel dan data kanker prostat sebanyak 12626 Variabel. Data tersebut kemudian akan dianalisis dengan dua metode seleksi variabel yaitu *important variable value* dan *genetic algorithm.* untuk memilih dimensi atau variabel data sehingga dapat meningkatkan akurasi klasifikasi data. Metode important variable value termasuk dalam kategori embedded seleksi variabel dan genetic algoritm termasuk wraper seleksi variabel. Pada tahun 2021 Wang, dkk melakukan penelitian terkait seleksi variabel menggunakan *important variable value*, yaitu memilih variabel berdasarkan kepentingannya. *Important variable value* merupakan salah satu fitur seleksi variabel yang ada di metode klasifikasi mengguankan *decision tree* atau *random forest.* Sedangkan Sayed dkk pada tahun 2019 meneliti seleksi variabel menggunakan genetic algorithm, dimana didapatkan seleksi variabel menggunakan KNN dan *random forest*.

Sehingga tujuan dari penelitian ini adalah Mengkaji metode seleksi variabel untuk mengklasifikasikan data *microarray,* dan Membandingkan metode *important variable value* dan *genetic algorithm* untuk seleksi variabel data *microarray*.

## 2. METODE PENELITIAN

### 2.1 Important Variable Value

*Important variable value* merupakan salah satu metode untuk memilih variabel. Metode *important variabl*e memilih variabel berdasarkan tingkat kepentingannya. Pada tahun 1984 Brieman, dkk mengusulkan *important variable* dengan menggunakan pemisahan untuk mengurangi resiko bahwa variabel yang penting disamarkan. *Important variable* mewakili signifikasin dari setiap variabel dalam data pengaruhnya terhadap model yang dihasilkan. Setiap variabel akan diberi peringkat sesui dengan kontribusinya pada model. Sehingga dengan metode ini dapat menghilangkan variabel tertetu yang tidak memberikan kontribusi terhadap model. Nilai dari *important variable* dihitung dari penjumlahan penurunan kesalahan saat setiap variabel dipisahkan. Nilai *important variabel* dibatasi antara 0 sampai 1.

Salah satu cara yang paling sering digunakan untuk dalam menentukan tingkat kepentingan variabel yaitu menggunakan permutation importance. Pengukuran Important variable dilakukan pada metode klasifikasi menggunakan Random Forest, dimana pada Random Forest memperkirakan kepentingan covariat dengan cara permutasi setiap variabel dan mengklasifikasikan ulang berdasarkan variabel yang telah dipermutasi (Rad, Koohkan, Fanaei, & Rad, 2015). Setiap Variabel diukur tingkat kepentingannya dengan cara semua variabel dipermutasi dan ukuran tingkat kepentingannya sebagai perbedaan dalam ukurasi dari prediksi yang disebabkan oleh permutasi variabel tersebut (Hjerpe, 2016).

### 2.2 Genetic Algorithm

*Genetic Algorithm* (GA) adalah metode yang digunakan untuk menemukan solusi pada permasalahan optimasi dan pencarian secara komputasi. Metode GA terinspirasi dari dari evolusi biologi yang meliputi pewarisan, mutasi, seleksi, dan cross over. Metode GA memungkinkan untuk menemukan solusi dari masalah yang tidak dapat ditangani oleh metode optimasi lain yang disebabkan oleh kurangnya kontinuitas, turunan, linieritas dan variable lainnya. Metode GA dapat menyelesaikan permasalahan data yang mempunyai dimensi tinggi dalam menyeleksi variable. Variabel diseleksi menggunakan nilai *fitness* dan kemudian dikombinasikan dengan *cross over* dan mutasi sehingga menghasilkan generasi baru. Genetic algorithm untuk seleksi variabel yang berbasis wraper memberikan fungsi output dengan kemampuan generalisasi yang baik (Djellali & Adda, 2017).

Gen adalah istilah untuk menunjukkan variabel, alel adalah nilai gen atau variabel, yang bisa 1 atau 0. Ini adalah 1 yang berarti bahwa variabel dipilih secara acak dan memiliki nilai 0 yang berarti bahwa variabel tidak dipilih. Kromosom adalah pengumpulan gen dan individu yang merupakan salah satu solusi yang mungkin untuk pemilihan variabel (Nurlaily, et al., 2019). Prosedur seleksi variabel menggunakan GA adalah sebagai berikut:

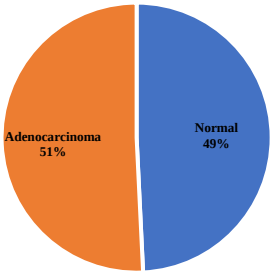
1. Inisialisasi menciptakan dan menginisialisasi individu dalam populasi secara acak.
2. Fitness Assignment adalah mengevaluasi solusi yang dihasilkan untuk dipilih atau tidak. Jika kesalahan tinggi, kebugaran rendah. Individu dengan kebugaran tinggi memiliki peluang lebih tinggi untuk dipilih.
3. Seleksi adalah memilih individu yang akan bergabung kembali dengan generasi berikutnya. Operator pilihan memilih individu berdasarkan nilai kebugaran mereka.
4. Crossover menggabungkan kembali individu yang dipilih untuk menghasilkan populasi baru. Operator ini secara acak memilih dua individu dan menggabungkan variabel mereka untuk mendapatkan keturunan dalam populasi baru.
5. Mutasi adalah solusi untuk masalah crossover yang menghasilkan keturunan yang sama. Dalam proses perubahan mutasi nilai beberapa variabel secara acak.

### 2.3 Data Microarray

Data microarray adalah data yang berisi serangkaian sampel bisa DNA, RNA, jaringan dan protein, sehingga jenis data microarray tergantung dari sampel yang diamati (Tuimala & Laine, 2003). Data microarray bisa memuat ratusan ribu gen yang bisa diamati. Data microarray yang biasa digunakan dalam penelitian adalah DNA microarray. DNA microarray digunakan untuk menentukan tingkat ekspresi gen dalam sampel dan urutan gen dalam sampel. Data microarray yang digunakan pada penelitian ini ada dua yaitu data lung cancer dan prostate cancer. Perlu dilakukan serangkaian penelitian untuk mendapatkan data microarray dan disebut dengan microarray experiment.

## 3. HASIL DAN PEMBAHASAN

Data *lung cancer* pada penelitian ini terdiri dari 65 pengamatan pasien kanker paru-paru dengan jumlah ekpresi gen yang diamati sebanyak 24.257. Pada data lungt cancer, pasien dibedakan menjadi dua yaitu pasien normal dan pasien *adenocarcinoma*. Karakteristik data *lung cancer* berdasarkan kelasnya dapat dilihat di Gambar 1.



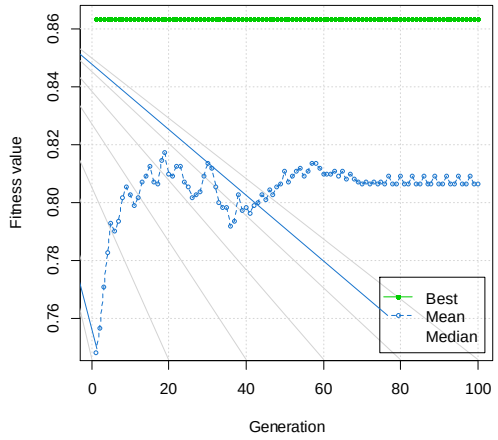

**Gambar 1.** Karakteristik data lung cancer berdasarkan kelas pasien

Berdasarkan Gambar 5.1 terlihat bahwa dari 65 pasien *lung cancer* terdapat 51% pasien masuk kelas normal dan 51% pasien termasuk kelas *adenocarcinoma*. Sehingga dilihat dari persebaran kelas pasien dapat dikatakan bahwa data *lung cancer* ini *balance* atau seimbang untuk setiap kelasnya.

Pada metode *important value* ini, menggunakan metode *decision tree* untuk metode klasifikasinya. Pada proses klasifikasi menggunakan *decision tree* ini terdapat fitur untuk seleksi variabel yaitu *important variable*, dimana saat proses klasifikasi, variabel yang diperkirakan mempengaruhi kelas *lung cancer* akan diseleksi berdasarkan tingkat kepentingannya. Seleksi variabel pada *lung cancer* dengan menggunakan metode important value didapatkan jumlah variabel yang terpilih sebanyak 112 variabel. Sehingga dengan menggunakan metode *important value* dapat mereduksi sebesar 99,53% dari total variabel keseluruhan. Metode *important value* ini memilih variabel berdasarkan nilai tingkat kepentingannya. Berdasarkan perhitungan didapat-kan 112 variabel yang nilai tingkat kepentingannya tidak sama dengan 0

sehingga, variabel yang terpilih yaitu 112 variabel. Hal ini dikarenakan variabel yang nilai kepentingannya 0 dianggap tidak penting dan tidak perlu dimasukkan dalam model klasifikasi.

Seleksi variabel pada data *lung cancer* menggunakan metode *genetic algorithm* dimana iterasi yang dilakukan sebanyal 100 kali, menggunakan peluang *crossover* sebesar 0.8, peluang mutasi 0.05 dan *fitness value* yang digunakan adalah nilai kebaikan model AUC dari metode klasifikasi SVM. Berdasarkan iterasi yang dilakukan sebanyak 100 kali didapatkan nilai *fitness value* sebesar 0.863. berikut adalah nilai *fitness* yang didapatkan selama iterasi berlangsung.




**Gambar 2.** Nilai fitness GA data lung cancer

Menggunakan metode *genetic algorithm* didapatkan jumlah variabel terpilihnya sebanyak 12266 variabel. Jika dibandingkan dengan total variabel aslinya, metode *genetic algorithm* dapat mereduksi variabel sebanyak 49,43% dari jumlah variabel aslinya.

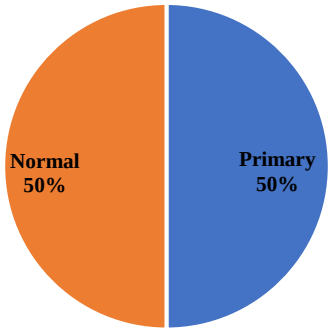
Sehingga jika dibandingkan hasil seleksi variabel untuk data *lung cancer* berdasarkan jumlah variabel terpilhnya dapat dilihat pada Tabel 1.

**Tabel 1.** Perbandingan seleksi variabel data lung cancer

| data | Metode | Total Variabel | Variabel Terpilih | Persentase Variabel Terpilih |
|---|---|---|---|---|
| Lung Cancer | Important Value | 24257 | 112 | 0,46% |
| | Genetic Algorithm | | 12266 | 50,57% |

Pada setiap gambar harus diberikan keterangan di bawah gambar. Keterangan pada tabel diberikan di atas tabel. Keterangan dituliskan dengan huruf kecil kecuali pada karakter pertama pada tiap kalimat. Seluruh gambar harus diberi penomoran secara berurutan. Jika Gambar besar maka diletakkan di tengah halaman (*center alignment*) dengan judul ditengah dan jika gambar kecil maka letakkan di tengah (*center columns*) baik itu pada kolom 1 ataupun pada kolom 2 dengan nama gambar rata *justify*, demikian halnya dengan tabel.

Data *prostate cancer* pada penelitian ini terdiri dari 124 pengamatan pasien kanker prostate dengan jumlah ekpresi gen yang diamati sebanyak 12.626. Pada data *prostate cancer*, pasien dibedakan menjadi dua yaitu pasien normal dan pasien *primary* (utama) . Karakteristik data *prostate cancer* berdasarkan kelasnya dapat dilihat di Gambar 3.

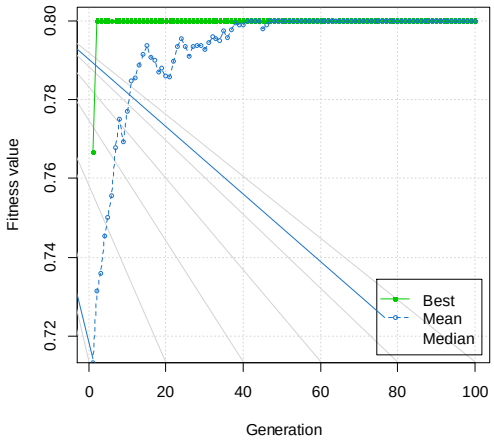



**Gambar 3.** Karakteristik data prostate cancer berdasarkan kelas pasien

Berdasarkan Gambar 3 terlihat bahwa dari 124 pasien *prostate cancer* terdapat 50% pasien masuk kelas normal dan 50% pasien termasuk kelas *primary*. Sehingga dilihat dari persebaran kelas pasien dapat dikatakan bahwa data *prostate cancer* ini *balance* atau seimbang untuk setiap kelasnya.

Pada metode *important value* ini, menggunakan metode decision tree untuk metode klasifikasinya. Pada proses klasifikasi menggunakan decision tree ini terdapat fitur untuk seleksi variabel yaitu important variable, dimana saat proses klasifikasi, variabel yang diperkirakan mempengaruhi kelas prostate cancer akan diseleksi berdasarkan tingkat kepentingannya. Sehingga diakhir proses seleksi variabel pada prostate cancer dengan menggunakan metode important value didapatkan jumlah variabel yang terpilih sebanyak 299 variabel. Metode *important value* ini memilih variabel berdasarkan nilai tingkat kepentingannya. Berdasarkan perhitungan didapatkan 299 variabel yang nilai tingkat kepentingannya tidak sama dengan 0 sehingga, variabel yang terpilih yaitu 299 variabel. Hal ini dikarenakan variabel yang nilai kepentingannya 0 dianggap tidak penting dan tidak perlu dimasukkan dalam model klasifikasi.

Seleksi variabel pada data *prostate cancer* menggunakan metode *genetic algorithm* dimana iterasi yang dilakukan sebanyal 100 kali, menggunakan peluang *crossover* sebesar 0.8, peluang mutasi 0.05 dan *fitness value* yang digunakan adalah nilai kebaikan model AUC dari metode klasifikasi SVM. Berdasarkan iterasi yang dilakukan sebanyak 100 kali didapatkan nilai *fitness value* yang paling optimum sebesar 0.800. berikut adalah nilai *fitness* yang didapatkan selama iterasi berlangsung.




**Gambar 4.** Nilai fitness GA data prostate cancer

Menggunakan metode *genetic algorithm* didapatkan jumlah variabel terpilihnya sebanyak 6359 variabel. Jika dibandingkan dengan total variabel aslinya, metode *genetic algorithm* dapat mereduksi variabel sebanyak 49,63% dari jumlah variabel aslinya.

Sehingga jika dibandingkan hasil seleksi variabel untuk data prostate cancer berdasarkan jumlah variabel terpilhnya dapat dilihat pada Tabel 2.

**Tabel 2.** Perbandingan seleksi variabel data prostate cancer

| data | Metode | Total Variabel | Variabel Terpilih | Persentase Variabel Terpilih |
|---|---|---|---|---|
| Prostate Cancer | Important Value | 12626 | 299 | 23,68% |
| | Genetic Algorithm | | 6359 | 50,37% |

## 4. KESIMPULAN DAN SARAN

Berdasarkan hasil analisis didapatkan kesimpulan bahwa data *lung cancer* dan *prostate canser* merupakan data yang balance jika dilihat dari proporsi kelas pasien. Seleksi variabel pada data lung cancer menggunakan metode *important value* memberikan hasil 122 variabel yang tepilih dan jika menggunakan metode *genetic algorithm* variabel terpilihnya sebesar 12.266 variabel. Seleksi variabel untuk data *prostate cancer* menggunakan metode *important value* jumlah variabel yang terpilihnya sebesar 299 variabel dan jika menggunakan metode *genetic algorithm* jumlah variabel terpilihnya sebanyak 6.359 variabel. Sehingga jika dilihat dari jumlah variabel terpilihnya metode *important value* memberikan reduksi variabel yang lebih tinggi dibandingkan metode *genetic algorithm* untuk data *lung cancer* dan *prostate cancer*.

Saran untuk penelitian ini adalah perlu dilakukan penilitian lanjutan untuk membuktikan metode seleksi variabel mana yang lebih baik jika dilihat dari nilai kebaikan model untuk mengklasifikasikan pasien dari dari data *lung cancer* dan *prostate cancer*. Metode yang bisa digunakan diantaranya adalah regresi logistik, *decision tree*, SCM, *Naïve Nayes*.

## DAFTAR PUSTAKA

Dash, R. (2020). A two stage grading approach for feature selection and classification of microarray data using Pareto based feature ranking techniques: A case study. *Journal of King Saud University –Computer and Information Sciences*, 232-247.

Djellali, C., & Adda, M. (2017). A New Predictive Approach to Variables Selection Through Genetic Algorith and Fuzzy Adaptive Resonance Theory Using Medical Diagnosis as a Case. *Procedia Computer Science*, 448-457.

Hambali, M. A., Oladele, T. O., & Adewole, K. S. (2020). Microarray cancer feature selection: Review, challenges and research directions. *International Journal of Cognitive Computing in Engineering* , 78-97.

Hjerpe, A., 2016. *Computing Random Forest Variable Importance Measures (VIM) on Mixed Continous and Categorical Data.* Stockholm, Sweden: KTH Royal Institute of Technology School of Computer Science and Communication.

Hira, Z. M., & Gillies, D. F. (2015). A Review of Feature Selection and Feature Extraction Methods. *Advaces in Bioinformatics , 1-13.*

Nurlaily, D., Irhamah, Purnami, S. W. & Kuswanto, H., 2019. *Support Vector Machine for Imbalanced Microarray Dataset Classification*

*Using Ant Colony Optimization and Genetic Algorithm.* s.l., AIP Publishing.

Rad, M. R., Koohkan, S., Fanaei, H. R., & Rad, M. R. (2015). Application of Artificial Neural Networks to predict the final fruit weight and random forest to select important variables in native population of melon (Cucumis melo L.). *Scientia Horticulture*, 108-112.

Sayed, S., Nassef, M., Badr, A., & farag, I. (2019). A Nested Genetic Algorithm for Feature Selection in High-dimensional Cancer Microarray Datasets. *Microarray Datasets*.

Wang, L., Huang, Z., & Wang, R. (2021). Discrimination of cracked soybean seeds by near-infrared spectroscopy and random forest variable selection. *Infrared Physics and Technology*.

Wang, X., & Simon, R. (2011). Microarray Based Cancer Prediction using Single Genes. *BMC Bioinformatics , Vol. 12, hal. 391-400.*