JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 3, Juli 2020, Page 486-494
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i3.2096

Deteksi Kanker Berdasarkan Data *Microarray* Menggunakan Metode *Naïve Bayes* dan *Hybrid Feature Selection*

Bintang Peryoga*, Adiwijaya, Widi Astuti

Fakultas Informatika, Universitas Telkom, Bandung, Indonesia
Email: [1,*]bintangperyoga@student.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id
[3]astutiwidi@telkomuniversity.ac.id
Email Penulis Korespondensi: bintangperyoga@student.telkomuniversity.ac.id

Abstrak−Kanker merupakan penyakit mematikan yang bertanggung jawab atas kematian 9.6 juta jiwa pada 2018 berdasarkan data WHO sehingga diperlukan pendeteksian kanker sejak dini agar dapat diobati segera dan kematian akibat kanker dapat dikurangi. *Microarray* merupakan teknologi yang dapat memonitor dan menganalisis ekspresi gen kanker pada data *microarray* akan tetapi memiliki dimensi data yang tinggi dan sampel yang sedikit sehingga dibutuhkan reduksi dimensi agar proses klasifikasi optimal. Reduksi dimensi dapat mengurangi penggunaan fitur untuk proses klasifikasi dengan cara memilih beberapa fitur yang paling berpengaruh. Metode *Hybrid* merupakan salah satu reduksi dimensi dengan cara menggabungkan metode *Filter* dengan *Wrapper* sehingga mendapatkan sisi positif dari keduanya. Dalam penelitian ini, peneliti menggabungkan *Naïve Bayes* dengan *Hybrid Feature Selection*(*Information Gain – Genetic Algorithm*) pada data kanker *microarray Lung Cancer, Ovarian Cancer, Breast Cancer, Colon Tumor, dan Prostate Tumor*. Data kanker *microarray* didapat dari *Kent-Ridge Biomedical Dataset*. Hasil dari penelitian menunjukkan dari 5 data yang digunakan, 4 data mendapatkan tingkat akurasi 87-100% sedangkan data tumor prostat mendapatkan akurasi terkecil yaitu 61.14%. Implementasi dari metode seleksi fitur serta klasifikasi terhadap 5 data kanker diatas hanya menggunakan kurang dari 63 fitur saja untuk mendapatkan akurasi tersebut.

Kata Kunci: Kanker, *Microarray*, *Naïve Bayes*, *Information Gain*, *Genetic Algorithm*, *Hybrid Feature Selection*.

Abstract−Cancer is a deadly disease that is responsible for 9.6 million death in 2018 based on WHO data so early cancer detection is needed so can be treated immediately and cancer deaths can be reduced. Microarray is technology that can monitor and analyze the expression of cancer genes in microarray data but has high data dimension and small sample so dimensional reductions are needed for the optimal classification process. Dimension reduction can reduce the use of features for the classification process by selecting some influential features. Hybrid method is one dimension reduction by combining Filter method with Wrapper so it gets the both advantage. In this case, researchers combined Naïve Bayes with Hybrid Feature Selection (Information Gain - Genetic Algorithm) on cancer data for microarray Lung Cancer, Ovarian Cancer, Breast Cancer, Colon Tumors, and Prostate Tumors. These data were obtained from Kent-Ridge Biomedical Dataset. The results showed that from 5 data used, 4 data obtained an accuracy between 87-100% while the prostate tumor data obtained the smallest accuracy of 61.14%. The implementation of the feature selection method and the classification of the 5 cancer data above only uses less than 63 features to obtain this accuracy.

Keywords: Cancer, Microarray, Naïve Bayes, Information Gain, Genetic Algorithm, Hybrid Feature Selection.

1. PENDAHULUAN

Kanker merupakan penyakit mematikan yang dapat menyerang bagian tubuh mana pun. Menurut World Health Organization[1], kanker merupakan penyakit mematikan kedua dan bertanggung jawab atas 9.6 juta kematian pada tahun 2018 di dunia dengan kasus kanker yang banyak terjadi yaitu kanker paru-paru (2.09 juta kasus) dan kanker payudara (2.09 juta kasus) sehingga diperlukan pendeteksian kanker sejak dini agar dapat penanganan segera dan tingkat kematian akibat kanker dapat dikurangi. Salah satu teknologi yang dapat dimanfaatkan untuk mendeteksi kanker yaitu *microarray*. Microarray mampu membantu peneliti untuk memantau dan menganalisis perubahan ekspresi gen dalam suatu organisme[2]. Teknologi *Microarray* pada data kanker mempelajari identifikasi ekspresi dan karakteristik yang berbeda pada gen pasien kanker yang hasilnya dapat diaplikasikan untuk memprediksi keadaan pasien tersebut[3]. Akan tetapi, data *microarray* memiliki dua masalah penting yaitu *high-dimensional* dan *high-complexity*[4]. Data *microarray* bersifat *high-dimensional* karena memiliki fitur yang mencapai ribuan lebih. Dimensi data yang tinggi akan berdampak pada *learning algorithm* karena akan menurunkan kinerja program ketika fitur yang tidak terlalu penting menambah ruang pencarian dan membuat generalisasi menjadi lebih sulit[5]. Oleh karena itu, dibutuhkan proses reduksi dimensi untuk mengurangi kompleksitas data tersebut[6].

Reduksi dimensi dapat mengurangi penggunaan fitur yang dianggap tidak penting untuk proses klasifikasi. Pemilihan reduksi dimensi yang tepat dapat mengoptimalkan waktu pengklasifikasian dan akurasi[7]. Seleksi fitur merupakan salah satu cara untuk mereduksi dimensi. Menurut Pengyi Yang pada penelitiannya[8], seleksi fitur dibagi menjadi 3 yaitu *Filter*, *Wrapper*, dan *Embedded(Hybrid)*. Metode *Filter* bekerja tanpa pengaruh dari teknik klasifikasi yang dipakai sehingga secara komputasi akan lebih efisien[9]. Cara kerja metode *Filter* yaitu dengan menghitung nilai peringkat dari tiap fitur. *Information Gain* merupakan salah satu metode *Filter*. Metode *Wrapper* memiliki kelemahan yaitu komputasi yang tidak efisien karena ia mengambil hipotesis model ke dalam *training* dan *testing* pada ruang fitur yang dipakai, juga mengambil lebih banyak *CPU time* dan memori untuk *running program*[9]. Kelebihan dari *Wrapper* adalah ia dapat mendeteksi sifat ketergantungan antar fitur. *Genetic Algorithm* merupakan salah satu metode *Wrapper* dengan jenis *Randomize* yang paling sering

Submitted: **30/03/2020**; Accepted: **03/05/2020**; Published: **20/08/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 3, Juli 2020, Page 486-494
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i3.2096

dipakai[9]. Di antara semua metode *Wrapper*, *Genetic Algorithm* mendapatkan akurasi tertinggi dengan jumlah gen yang dipilih paling sedikit[4].

Hybrid Feature Selection merupakan salah satu metode seleksi fitur. Metode *Hybrid* dapat menggabungkan metode *Filter* dan *Wrapper* menjadi suatu kesatuan sehingga secara *computational time* lebih cepat dan secara performansi lebih baik[4]. Pada penelitian Nada Almugren[4] tahun 2019 yang berisi tabel komparasi penelitian sebelumnya tentang penggunaan metode *Hybrid* yang beragam, hasil dari banyaknya penelitian tersebut mendapatkan tingkat akurasi diatas 83% untuk data *microarray Colon*, *Leukemia*, *Prostate*, *Lung*, dan *Breast* sehingga terbukti bahwa metode *Hybrid* dapat mengurangi penggunaan fitur gen pada saat klasifikasi tanpa mengurangi tingkat akurasi. Beberapa peneliti yang menggunakan metode *Hybrid* diantaranya yaitu Hanaa Salem[10], Chen-Sang Yang[11], dan Cheng-Huei Yang[12], dan Abdul Hasnat[13].

Hanaa Salem[10] membuat program klasifikasi terhadap data tumor usus besar dan kanker prostat menggunakan seleksi fitur (*IG/SGA*) dengan *Genetic Programming*. Ia menyetel parameter pada *Genetic Algorithm*, diantaranya *crossover rate* = 0.8 dan *mutation rate* = 0.1. Populasi yang dimunculkan sebanyak 100 individu dengan regenerasi sebanyak 10 kali. Dari penelitian tersebut, didapatkan akurasi pada data tumor usus besar 85.48% dan kanker prostat sebesar 100%.

Chen-Sang Yang[11] melakukan penelitian pada data kanker prostat dengan menggunakan *Information Gain-Chaotic Genetic Algorithm* dan *K-Nearest Neighbor*. Parameter yang digunakan yaitu *mutation rate* = 0.1 dan *crossover rate* = 1.0. Ia membuat populasi sebanyak 30 individu dengan generasi sebanyak 100 generasi. Ia mendapatkan akurasi untuk kanker prostat sebesar 99.96%.

Cheng-Huei Yang[12] bereksperimen dengan berbagai macam data kanker mendapatkan akurasi diatas 85% untuk semua data kanker yang ia gunakan. Parameter yang digunakan dalam programnya yaitu *mutation rate* = 0.1 dan *crossover rate* = 0.8. Ia memunculkan 30 populasi secara acak. Generasi pada program ini dibatasi hingga 100 generasi saja.

Abul Hasnat[13] melakukan klasifikasi pada data tumor usus besar menggunakan *CC-MOGA* dengan *K-NN*. Parameter *crossover rate* yang dipakai yaitu 0.8 dengan *mutation rate* sebesar 0.05. Populasi yang dibuat sebanyak 100 individu dengan generasi sebanyak 15000 generasi. Dari penelitian tersebut, ia mendapatkan skor akurasi untuk data kanker usus besar sebesar 82.3%. Pada penelitian ini, peneliti menggunakan seleksi fitur *Hybrid* dengan menggabungkan *Information Gain* dan *Genetic Algorithm* serta menggunakan metode klasifikasi *Gaussian Naïve Bayes* yang bertujuan data kanker yang dipakai mendapatkan akurasi diatas 95% dengan fitur yang dipakai kurang dari 50 fitur.

2. METODE PENELITIAN

2.1 Dataset

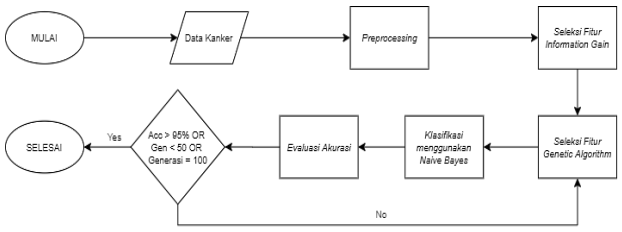
Dataset yang digunakan pada penelitian ini yaitu data kanker *microarray* dari Kent-Ridge Biomedical *Dataset*. Ada lima macam data yang dipakai yaitu data *Lung Cancer*, *Colon Tumor*, *Ovarian Cancer*, *Breast Cancer*, dan *Prostate Tumor*. Rincian data dapat dilihat pada Tabel 1.

Tabel 1. Rincian *Dataset*

Data	Jumlah Sampel	Jumlah Fitur	Jumlah Kelas
Lung Cancer	181	12533	2 (31 *Mesothelioma*, 150 *ADCA*)
Colon Tumor	62	2001	2 (40 *Negative*, 26 *Positive*)
Ovarian Cancer	253	15155	2 (91 *Normal*, 162 *Cancer*)
Breast Cancer	97	24482	2 (51 *non-relapse*, 46 *relapse*)
Prostate Tumor	136	12600	2 (59 *Normal*, 77 *Tumor*)

2.2 Skema Umum

Sistem yang dibangun pada penelitian ini bertujuan agar sistem dapat mengelompokkan data uji kanker *microarray* menggunakan metode klasifikasi *Naïve Bayes* dan seleksi fitur metode *Hybrid*, yaitu dengan menggabungkan *Information Gain* dengan *Genetic Algorithm* sehingga fitur yang dipakai lebih sedikit daripada reduksi dimensi non-*Hybrid*. *Flow Chart* sistem dapat dilihat pada Gambar 1.



Gambar 1. Skema umum metode yang diusulkan

Submitted: **30/03/2020**; Accepted: **03/05/2020**; Published: **20/08/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 3, Juli 2020, Page 486-494
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i3.2096

2.3 Preprocessing

Proses yang dilakukan pada saat *preprocessing* ada dua, yaitu membersihkan *missing value* dan standarisasi. Membersihkan *missing value* dilakukan agar hasil klasifikasi tidak kacau. Data kanker merupakan jenis data numerik sehingga cara untuk mengisi *missing value* yaitu dengan mencari nilai median yang satu tipe kelas dengan data tersebut. Standarisasi dilakukan agar data memiliki distribusi normal. Standarisasi dilakukan dengan rumus Z-Score berikut.

$$Z = \frac{X_i - \bar{X}_j}{\sigma_j} \qquad (1)$$

Keterangan:
Z = Z-Score
X_i = Nilai data ke-i
$\bar{X}_j$ = Nilai rata-rata dari fitur ke-j
σ_j = Nilai standar deviasi dari fitur ke-j

2.4 Seleksi Fitur *Information Gain*

Seleksi fitur dilakungan untuk memilih beberapa fitur yang dianggap penting untuk proses klasifikasi[14]. *Information Gain* merupakan salah satu metode *Filter*. Metode *Filter* bekerja tanpa pengaruh dari teknik klasifikasi yang dipakai sehingga secara komputasi akan lebih efisien[9]. Cara kerja metode *Filter* yaitu dengan menghitung nilai peringkat dari tiap fitur. Berikut rumus *Information Gain.*

$$Info(S) = \sum_{i=1}^{k} - P(C_i, S) * log_2(P(C_i, S)) \qquad (2)$$

$Info(S)$ merupakan rumus untuk mencari *Entrophy* dengan $P(C_i, S)$ adalah peluang kelas C_i pada himpunan S.

$$Info_A(S) = \sum_{i=1}^{v} - \frac{|S_i|}{|S|} * Info(S_i) \qquad (3)$$

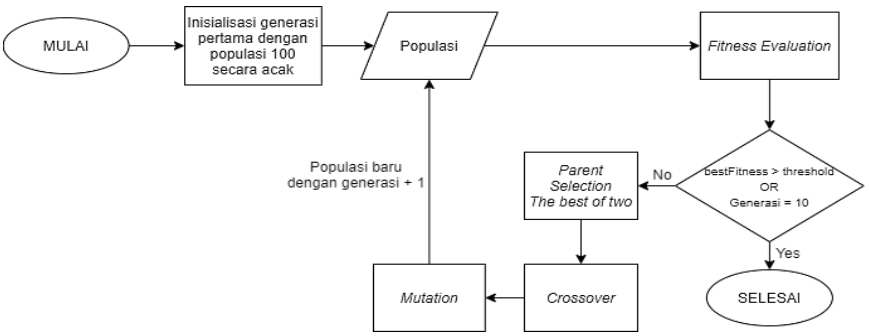
S_i adalah jumlah kasus pada partisi ke-i dengan A_i adalah nilai pada atribut atau fitur A.

$$Gain(A) = Info(S) - Info_A(S) \qquad (4)$$

$Gain(A)$ merupakan rumus *Information Gain* terhadap fitur atau atribut A.

2.5 Seleksi Fitur *Genetic Algorithm*

Genetic Algorithm bekerja dengan cara menentukan solusi terbaik dengan nilai fitness terbaik. Berikut *Flowchart* untuk algoritma *Genetic Algorithm*.



Gambar 2. Alur *Genetic Algorithm*

Mula-mula dibangkitkan populasi sebanyak 100 individu secara acak. Satu individu memiliki gen sebanyak jumlah fitur data yang dipakai. Gen bernilai 0 menandakan fitur tersebut tidak dipakai, sebaliknya gen bernilai 1 menandakan fitur tersebut dipakai untuk perhitungan *fitness value*. *Fitness Evaluation* dilakukan untuk menghitung skor akurasi tiap individu sebagai *fitness value* menggunakan metode klasifikasi *Gaussian Naïve Bayes*. *Fitness value* terbaik dijadikan sebagai *bestFitness*. Jika *bestFitness* > *threshold* atau populasi telah mencapai generasi ke-10, maka program berakhir. Jika tidak, maka dipilih dua *parent* dengan *fitness value* terbaik lalu dilakukan proses *Crossover* dan *Mutation* untuk membuat keturunan generasi selanjutnya dengan populasi baru. *Crossover* merupakan proses pertukaran silang gen antara dua individu untuk menghasilkan individu yang baru. *Mutation* merupakan proses termutasinya gen atau berubahnya nilai pada gen. Proses *Crossover* dan *Mutation* ditentukan oleh peluang yang dipanggil secara acak sehingga tidak semua individu dan gen melewati proses *Crossover* dan *Mutation*. Rincian parameter yang digunakan dapat dilihat pada Tabel 2.

Tabel 2. Parameter yang dipakai peneliti

Parameter	Nilai
Mutation Rate	0.1
Crossover Rate	0.8
Populasi	100

Submitted: **30/03/2020**; Accepted: **03/05/2020**; Published: **20/08/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 3, Juli 2020, Page 486-494
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i3.2096

Parameter	Nilai
Generasi	10
CV	5

2.6 Klasifikasi menggunakan *Naïve Bayes*

Metode klasifikasi yang dipakai pada penelitian ini yaitu metode klasifikasi *Naïve Bayes.* Menurut Mubarok[15], metode *Naïve Bayes* telah terbukti memiliki kinerja yang baik untuk banyak klasifikasi masalah. Pengklasifikasian *Naïve Bayes* adalah klasifikasi *Bayesian Network* sederhana yang dibangun atas asumsi kuat bahwa atribut yang berbeda independen satu sama lain[16]. Rumus probabilitas metode *Naïve Bayes* yaitu sebagai berikut.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (5)$$

P(*H*|*X*) merupakan peluang hipotesis *H* berdasarkan kondisi atribut *X*. Variabel *X* adalah data sampel dengan kelas(label) yang tidak diketahui. Variabel *H* merupakan data hipotesis. *P(H)* adalah peluang dari hipotesis *H*. *P(X)* adalah peluang dari *X* yang diamati. *P(X)* dapat diabaikan karena sama dengan satu[17]. *P(X|H)* adalah peluang *X* berdasarkan kondisi pada hipotesis *H*.
Dikarenakan data kanker *microarray* bertipe numerik, maka peneliti memakai rumus *Gaussian Naïve Bayes* dengan menggunakan distribusi normal berikut.

$$\hat{P}(X_j|C = c_i) = \frac{1}{\sigma_{ji}\sqrt{2\pi}} exp\left(-\frac{(X_j-\mu_{ji})^2}{2\sigma_{ji}^2}\right) \quad (6)$$

Variabel μ_{ji} merupakan rata-rata(*mean*) dari variabel X_j dengan $C = c_i$. Variabel σ_{ji} merupakan standar deviasi dari variabel X_j dengan $C = c_i$.

2.7 Evaluasi Akurasi

Perhitungan akurasi, presisi, *recall*, dan *f1-score* menggunakan *Confusion Matrix*. Hasil tersebut didapat berdasarkan parameter *True Positive(TP)*: hasil prediksi kelas positif(*cancer*) bernilai benar sesuai data aktualnya, *True Negative(TN)*: hasil prediksi kelas negatif(non-*cancer*) bernilai sama dengan data aktualnya, *False Positive(FP)*: hasil prediksi kelas positif(*cancer*) bernilai beda dengan data aktualnya, *False Negative(FN)*: hasil prediksi kelas negative(non-*cancer*) bernilai beda dengan data aktualnya. Ilustrasi tabel *Confusion Matrix* dapat dilihat pada Tabel 3 berikut.

Tabel 3. *Confusion Matrix*

		Data Aktual	
		Positif	**Negatif**
Data Prediksi	**Positif**	*True Positive (TP)*	*False Positive(FP)*
	Negatif	*False Negative(FN)*	*True Negative(TN)*

Perhitungan akurasi, presisi, *recall*, dan *f1-score* dapat dilihat pada formula berikut.

$$Akurasi = \frac{TP+TN}{TP + TN + FP + FN} \quad (7)$$

Akurasi digunakan untuk menentukan nilai prediksi bernilai benar terhadap semua sampel data.

$$Presisi = \frac{TP}{TP + FP} \quad (8)$$

Presisi digunakan untuk menentukan nilai prediksi benar bernilai positif terhadap semua prediksi positif.

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

Recall digunakan untuk menentukan nilai prediksi benar bernilai positif terhadap semua data bernilai positif.

$$F1 = 2 * \frac{Recall*Precision}{Recall+Precision} \quad (10)$$

F1-Score digunakan untuk menentukan nilai perbandingan rata-rata presisi dan *recall* yang dibobotkan.
Perhitungan skor akurasi, presisi, *recall*, dan *f1-score* dilakukan dengan cara *k-fold cross-validation* dengan menetapkan nilai k=5 yang bertujuan agar hasil skor optimal. Program berakhir setelah mendapatkan skor akurasi diatas 95% dan jumlah fitur dibawah 50 atau program berakhir setelah melakukan iterasi sebanyak generasi yang ditetapkan sebelumnya, yaitu 10 generasi. *Output* dari program adalah gen yang terbaik beserta akurasi, presisi, *recall*, *f1-score*, dan jumlah fitur yang dipakai.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengujian

Berikut hasil pengujian menggunakan metode seleksi fitur *Hybrid(IG+GA)* dan metode klasifikasi *Gaussian Naïve Bayes* dengan parameter yang telah ditentukan sebelumnya.

Submitted: **30/03/2020**; Accepted: **03/05/2020**; Published: **20/08/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 3, Juli 2020, Page 486-494
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i3.2096

Tabel 4. Setelah dilakukan *Filtering* dengan *Information Gain*

Data	Jumlah Fitur	Setelah IG	**Avg. Accuracy**	**Avg. Precision**	**Avg. Recall**	**Avg. f1-score**	**Comp. Time**
Colon	2000 fitur	67 fitur	72.56%	58.44%	81%	67.5%	3 s
Prostate	12600 fitur	115 fitur	51.53%	25%	40%%	30.77%	2.2 s
Lung	12533 fitur	66 fitur	98.89%	100%	98.67%	99.32%	1.8 s
Breast	24481 fitur	90 fitur	61.68%	61.67%	42.89%	49.08%	2.4 s
Ovarian	15154 fitur	99 fitur	98.81%	99.38%	98.75%	99.06%	1.9 s

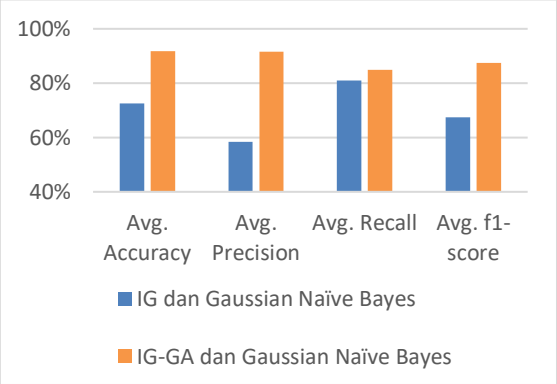
Tabel 5. Hasil akhir setelah dilakukan *Wrapping* dengan *Genetic Algorithm*

Data	Setelah IG	Setelah GA	**Avg. Accuracy**	**Avg. Precision**	**Avg. Recall**	**Avg. f1-score**	**Comp. Time**
Colon	67 fitur	28 fitur	91.8%	91.67%	85%	87.47%	109.23 s
Prostate	115 fitur	62 fitur	58.94%	32.5%	40%	35.39%	137.4 s
Lung	66 fitur	34 fitur	100%	100%	100%	100%	11.04 s
Breast	90 fitur	33 fitur	83.47%	81.23%	84.89%	82.59%	133.18 s
Ovarian	99 fitur	44 fitur	100%	100%	100%	100%	12.43 s

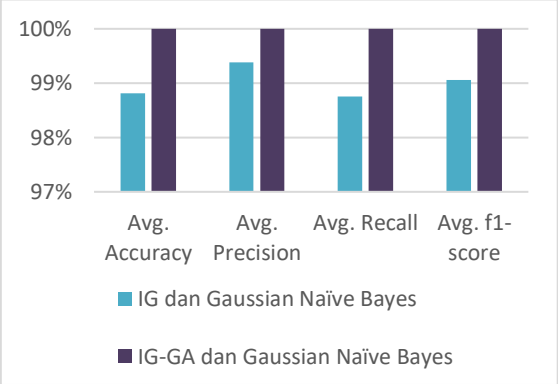
Hasil pengujian menunjukkan bahwa 4 data kanker menggunakan fitur kurang dari 50 fitur dan data *Prostate Tumor* menggunakan total 62 fitur serta dua data kanker(*Lung Cancer* dan *Ovarian Cancer*) mendapatkan akurasi lebih dari 95%. Akurasi tertinggi didapatkan dari data *Lung Cancer* dan *Ovarian Cancer* sebesar 100%. Akurasi terendah terdapat pada data *Prostate Cancer* sebesar 58.94%. Fitur paling sedikit digunakan berada pada klasifikasi data *Colon Tumor* sebanyak 28 fitur sedangkan fitur paling banyak digunakan pada data *Prostate Tumor*.

3.2 Pengaruh metode *Hybrid* terhadap akurasi

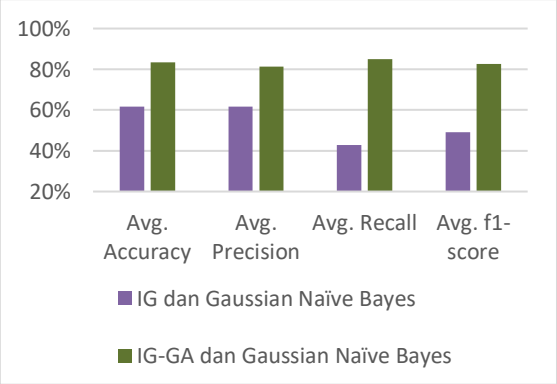
Peneliti melakukan pengujian lima *dataset* kanker yang dipakai menggunakan parameter yang telah ditentukan sebelumnya. Nilai parameter yang dipakai didapat dari rujukan penelitian sebelumnya. Hasil penelitian dapat dilihat pada gambar berikut.



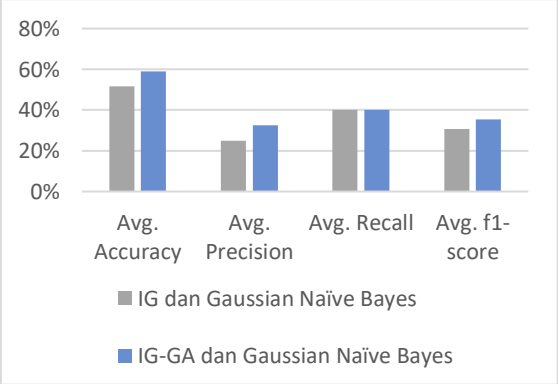
Gambar 3. Perbandingan nilai rataan akurasi, presisi, *recall*, dan f1 pada data *Colon Tumor*



Gambar 4. Perbandingan nilai rataan akurasi, presisi, *recall*, dan f1 pada data *Ovarian Cancer*



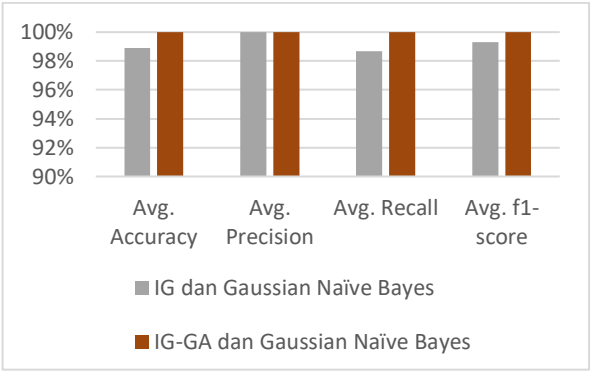
Gambar 5. Perbandingan nilai rataan akurasi, presisi, *recall*, dan f1 pada data *Breast Cancer*



Gambar 6. Perbandingan nilai rataan akurasi, presisi, *recall*, dan f1 pada data *Prostate Tumor*

Submitted: **30/03/2020**; Accepted: **03/05/2020**; Published: **20/08/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 3, Juli 2020, Page 486-494
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i3.2096

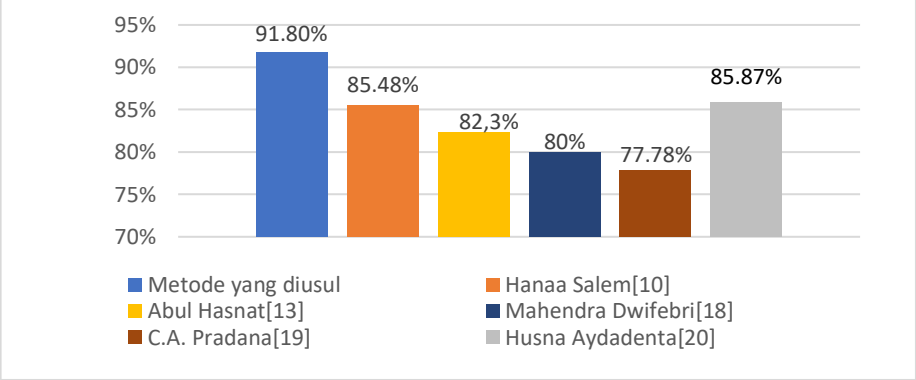


Gambar 7. Perbandingan nilai rataan akurasi, presisi, *recall*, dan f1 pada data *Lung Cancer*

Pada gambar diatas, diketahui bahwa pengujian pada lima data kanker dengan penggunaan metode *Hybrid* dapat meningkatkan akurasi. Data yang dapat terklasifikasi dengan baik adalah data *Lung Cancer* dan data *Ovarian Cancer*. Data dengan hasil terendah yaitu data *Prostate Tumor*.

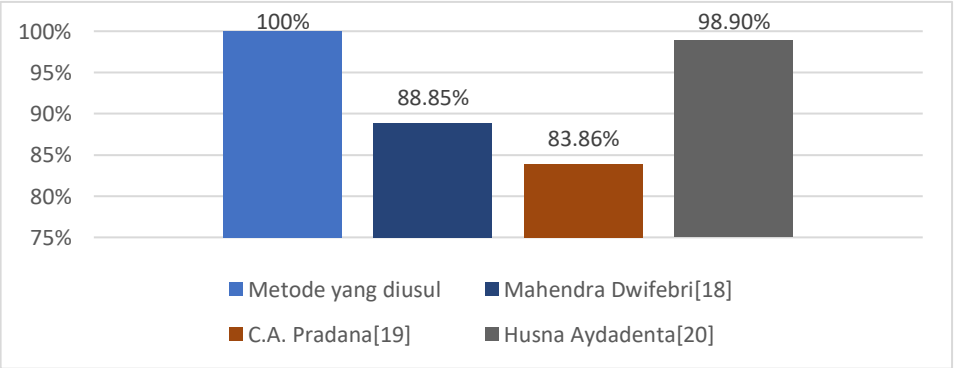
3.3 Komparasi Dengan Metode Lain

Peneliti juga melakukan perbandingan hasil akurasi dengan metode lain yang dilakukan oleh peneliti lain. Tujuan dari komparasi ini untuk melihat metode mana yang paling sesuai dengan data kanker yang digunakan. Komparasi metode dapat dilihat pada gambar berikut.



Gambar 8. Komparasi akurasi pada data *Colon Tumor*

Komparasi pada data *Colon Cancer* dilakukan oleh enam peneliti, yaitu metode yang diusul menggunakan *IG-GA* dengan *classifier Gaussian Naïve Bayes*, Hanaa Salem[10] menggunakan *IG-GA* dengan *classifier Genetic Programming*, Abul Hasnat[13] menggunakan *CC-MOGA* dengan *classifier k-Nearest Neighbor*, Mahendra Dwifebri[18] menggunakan *Mutual Information* dengan *classifier Bayes Theorem*, Pradana[19] menggunakan *iBPSO* dengan *classifier C4.5 Decision Tree*, dan Aydadenta[20] menggunakan *Relief Method* dengan *classifier Random Forest*. Dari hasil yang didapat pada Gambar 8, metode yang diusulkan mendapatkan akurasi tertinggi dibandingkan dengan metode lainnya sehingga metode yang paling cocok untuk penggunakan klasifikasi pasien kanker pada data *Colon Cancer* adalah metode yang diusulkan yaitu *IG-GA* dengan *classifier Gaussian Naïve Bayes*.

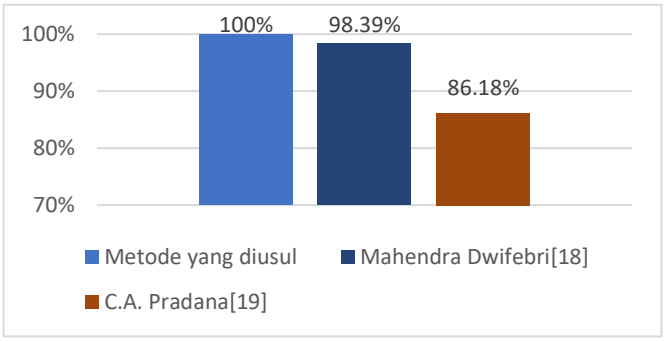


Gambar 9. Komparasi akurasi pada data *Lung Cancer*

Submitted: **30/03/2020**; Accepted: **03/05/2020**; Published: **20/08/2020**

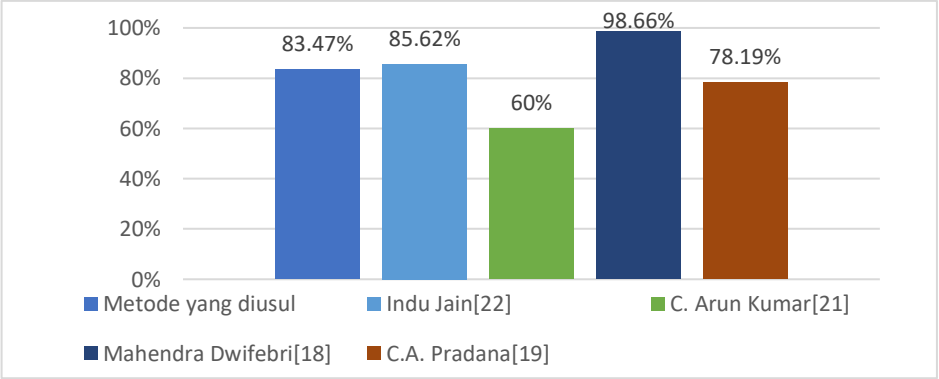
JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 3, Juli 2020, Page 486-494
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i3.2096

Perbandingan akurasi pada data *Lung Cancer* dengan tingkat akurasi tertinggi dimiliki oleh metode yang diajukan yaitu sebesar 100% sedangkan Mahendra Dwifebri[18] melakukan percobaan pada data *Lung Cancer* mendapat tingkat akurasi 88.85%. Pradana[19] mendapatkan akurasi sebesar 83.86%. Aydadenta[20] melakukan klasifikasi data *Lung Cancer* dan mendapatkan akurasi 98.9%. Dari Gambar 9 metode yang memiliki akurasi tertinggi yaitu metode yang diusulkan dengan *IG-GA* dan *classifier Gaussian Naïve Bayes* untuk klasifikasi data *Lung Cancer.*



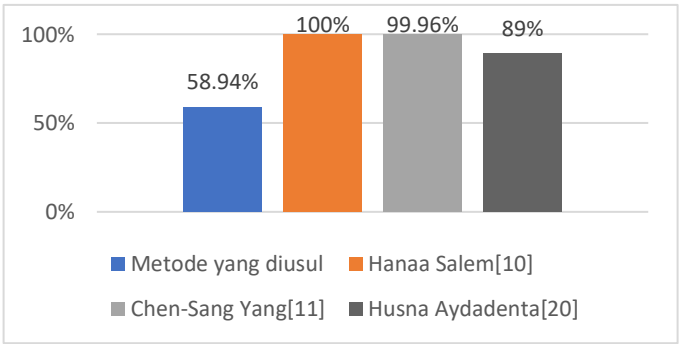
Gambar 10. Komparasi akurasi pada data *Ovarian Cancer*

Pada perbandingan akurasi di Gambar 10, ketiga peneliti yaitu metode yang diajukan, Mahendra Dwifebri[18], dan Pradana[19] mendapatkan tingkat akurasi yang tinggi untuk data *Ovarian Cancer*. Metode yang diusulkan mendapat akurasi tertinggi sebesar 100% disusul dengan Mahendra Dwifebri[18] sebesar 8.39% dan Pradana[19] sebesar 86.18%. Dari ketiga metode diatas, metode yang paling cocok adalah metode yang diajukan dengan menggunakan metode *IG-GA* dan *classifier Gaussian Naïve Bayes*.



Gambar 11. Komparasi akurasi pada data *Breast Cancer*

Perbandingan data *Breast Cancer* di Gamber 11 menampilkan lima peneliti dengan metode berbeda. Metode yang diusulkan tidak mendapatkan akurasi tertinggi. Metode yang diusulkan mendapat akurasi sebesar 83.47% diatas Arun Kumar[21] sebesar 60% dan Pradana[19] sebesar 78.19%. Peneliti yang mendapat akurasi lebih tinggi daripada metode yang diusul adalah Indu Jain[22] dengan tingkat akurasi sebesar 85.62% dan akurasi tertinggi oleh Mahendra Dwifebri[18] dengan akurasi 98.66% sehingga klasifikasi data *Breast Cancer* lebih cocok menggunakan metode yang dipakai oleh Mahendra Dwifebri[18] yaitu *Mutual Information* dengan *classifier Bayes Theorem*.



Gambar 12. Komparasi akurasi pada data *Prostate Tumor*

Submitted: **30/03/2020**; Accepted: **03/05/2020**; Published: **20/08/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 3, Juli 2020, Page 486-494
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i3.2096

Perbandingan akurasi pada Gambar 12 menunjukkan metode yang diusul mendapatkan akurasi kurang dari 80% dan akurasi terendah yaitu 58.94%. Adapula perbandingan dengan peneliti lain yang memiliki tingkat akurasi tinggi yaitu Hanaa Salem[10] sebesar 100%, Chen-Sang Yang[11] sebesar 99.96%, dan Aydadenta[20] sebesar 89%. Rendahnya hasil akurasi yang didapat dikarenakan pada saat implementasi metode terdapat kelas yang tidak terdeteksi oleh program sehingga berdampak pada hasil akurasi. Sudah dilakukan pengecekan beberapa kali pada program yang dibangun akan tetapi tetap tidak mendapatkan jawaban atas kendala tersebut. Oleh karena kendala tersebut, metode yang diusul tidak cocok untuk metode klasifikasi data *Prostate Tumor*.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, metode *Hybrid(IG-GA)* dapat mengoptimalkan akurasi dan konsumsi waktu yang dibutuhkan. Ada beberapa faktor yang mempengaruhi hasil pengujian tersebut. Penggunaan *Information Gain* berfungsi untuk mengoptimalkan konsumsi waktu komputasi sedangkan penggunaan *Genetic Algorithm* berfungsi untuk mengoptimalkan akurasi data yang telah diseleksi sebelumnya oleh *Information Gain*. Hasil seleksi fitur oleh *IG* berpengaruh kepada banyaknya fitur yang dipilih sehingga berpengaruh juga pada waktu komputasi yang dijalankan pada saat proses klasifikasi. Akurasi pada data *Prostate Tumor* rendah dikarenakan pada saat klasifikasi program menemukan suatu kejanggalan yaitu mendapatkan label kelas yang dianggap kosong pada saat proses klasifikasi. Semakin banyak fitur yang digunakan belum tentu mendapatkan hasil akurasi yang optimal. Hasil dari penelitian menunjukkan dari 5 data yang digunakan, 4 data mendapatkan tingkat akurasi 87-100% sedangkan data tumor prostat mendapatkan akurasi terkecil yaitu 61.14%. Implementasi dari metode yang diusulkan terhadap 5 data kanker diatas hanya menggunakan kurang dari 63 fitur untuk mendapatkan akurasi tersebut. Metode yang diusulkan ini cocok untuk data yang sudah memiliki akurasi diatas 90% seperti data *Lung Cancer*(akurasi 100%), *Ovarian Cancer*(akurasi 100%), dan *Colon Cancer*(akurasi 91.8%). Metode klasifikasi ini belum cocok untuk data *Breast Cancer* dan *Prostate* cancer karena memiliki akurasi yang masih rendah yaitu dibawah 90% sehingga diperlukan pengujian lebih lanjut untuk kedua data tersebut. Ada beberapa teknik *Crossover* dan *Mutation* yang dapat diuji untuk meningkatkan akurasi pada penelitian selanjutnya. Penelitian selanjutnya juga dapat mengubah nilai-nilai parameter yang sudah ditentukan pada penelitian ini sebagai bentuk pengujian dan perbandingan.

REFERENCES

[1] World Health Organization, “Cancer,” 12-Sep-2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed: 18-Mar-2020].
[2] M. M. Babu, “An Introduction to Microarray Data Analysis,” *Comput. genomics Theory Appl.*, vol. 225, p. 249, 2004.
[3] S. Michiels, S. Koscielny, and C. Hill, “Interpretation of microarray data in cancer,” *British Journal of Cancer*. 2007.
[4] N. Almugren and H. Alshamlan, “A survey on hybrid feature selection methods in microarray gene expression data for cancer classification,” *IEEE Access*. 2019.
[5] N. Sánchez-Maroño, O. Fontenla-Romero, and B. Pérez-Sánchez, “Classification of Microarray Data,” in *Microarray Bioinformatics*, V. Bolón-Canedo and A. Alonso-Betanzos, Eds. New York, NY: Springer New York, 2019, pp. 185–205.
[6] A. Adiwijaya, “Deteksi Kanker Berdasarkan Klasifikasi Microarray Data,” *J. MEDIA Inform. BUDIDARMA*, 2018.
[7] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, and D. S. Kusumo, “Dimensionality reduction using Principal Component Analysis for cancer detection based on microarray data classification,” *J. Comput. Sci.*, 2018.
[8] P. Yang, B. B. Zhou, Z. Zhang, and A. Y. Zomaya, “A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data,” *BMC Bioinformatics*, 2010.
[9] Z. M. Hira and D. F. Gillies, “A review of feature selection and feature extraction methods applied on microarray data,” *Adv. Bioinformatics*, 2015.
[10] H. Salem, G. Attiya, and N. El-Fishawy, “Classification of human cancer diseases by gene expression profiles,” *Appl. Soft Comput. J.*, 2017.
[11] C. S. Yang, L. Y. Chuang, J. C. Li, and C. H. Yang, “Information gain with chaotic genetic algorithm for gene selection and classification problem,” in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2008.
[12] C. H. Yang, L. Y. Chuang, and C. H. Yang, “IG-GA: A hybrid filter/wrapper method for feature selection of microarray data,” *J. Med. Biol. Eng.*, 2010.
[13] A. Hasnat and A. U. Molla, “Feature selection in cancer microarray data using multi-objective genetic algorithm combined with correlation coefficient,” in *Proceedings of IEEE International Conference on Emerging Technological Trends in Computing, Communications and Electrical Engineering, ICETT 2016*, 2017.
[14] W. Astuti and A. Adiwijaya, “Principal Component Analysis Sebagai Ekstraksi Fitur Data Microarray Untuk Deteksi Kanker Berbasis Linear Discriminant Analysis,” *J. MEDIA Inform. BUDIDARMA*, 2019.
[15] M. S. Mubarok, A. Adiwijaya, and M. D. Aldhi, “Aspect-based sentiment analysis to review products using Naïve Bayes,” in *AIP Conference Proceedings*, 2017.
[16] R. Aziz, C. K. Verma, and N. Srivastava, “A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data,” *Genomics Data*, 2016.
[17] E. Alpaydin, “Introduction to Machine Learning Ethem Alpaydin.,” *Introd. to Mach. Learn. Third Ed.*, 2014.
[18] M. D. Purbolaksono, K. C. Widiastuti, M. S. Mubarok, Adiwijaya, and F. A. Ma’ruf, “Implementation of mutual

Submitted: **30/03/2020**; Accepted: **03/05/2020**; Published: **20/08/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 3, Juli 2020, Page 486-494
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i3.2096

information and bayes theorem for classification microarray data," in *Journal of Physics: Conference Series*, 2018.

[19] A. C. Pradana, Adiwijaya, and A. Aditsania, "Implementing binary particle swarm optimization and C4.5 decision tree for cancer detection based on microarray data classification," in *Journal of Physics: Conference Series*, 2019.

[20] H. Aydadenta and Adiwijaya, "A clustering approach for feature selection in microarray data classification using random forest," *J. Inf. Process. Syst.*, 2018.

[21] C. Arun Kumar, M. P. Sooraj, and S. Ramakrishnan, "A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Microarray Datasets," in *Procedia Computer Science*, 2017.

[22] I. Jain, V. K. Jain, and R. Jain, "An improved Binary Particle Swarm Optimization (iBPSO) for Gene Selection and Cancer Classification using DNA Microarrays," in *2018 Conference on Information and Communication Technology, CICT 2018*, 2018.

[23] Mabarti, I., Aditsania, A., "Implementation of Minimum Redundancy Maximum Relevance (MRMR) and Genetic Algorithm (GA) for Microarray Data Classification with C4.5 Decision Tree". Journal of Data Science and Its Applications, 3(1), 2020.

[24] Purnomoputra, R.B., Adiwijaya, A. and Wisesty, U.N., 2019. Sentiment Analysis of Movie Review using Naïve Bayes Method with Gini Index Feature Selection. Journal of Data Science and Its Applications, 2(2), pp.85-94.

[25] Ma'ruf, F. A., Adiwijaya & Wisesty, U. N. "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, 2019.

[26] Manik, A., Adiwijaya, A., & Utama, D. Q. "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, 2019

[26 Daeli, N.O.F, Adiwijaya. Sentiment analysis on movie reviews using Information gain and K-nearest neighbor. Journal of Data Science and Its Applications, 3(1), 2020.

.

Submitted: **30/03/2020**; Accepted: **03/05/2020**; Published: **20/08/2020**