Indonesian Journal of Computer Science
ISSN 2302-4364 (*print*) dan 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam, No. 1, Padang, Indonesia, Telp. (0751) 7056199, 7058325
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Reduksi Dimensi pada Klasifikasi Data *Microarray* Menggunakan MRMR dan *Random Forest*

Lailan Sofinah Harahap[1], Erna Budhiarti Nababan[2], Syahril Efendi[3]
lailansofinahharahap@gmail.com, ernabrn@usu.ac.id, syahril.efendi@usu.ac.id
Universitas Sumatera Utara

Informasi Artikel

Diterima : 4 Jan 2023
Direview : 25 Jan 2023
Disetujui : 26 Feb 2023

Kata Kunci

Microarray, Random Forets, MRMR, Reduksi Dimensi

Abstrak

Di Indonesia prevalensi kanker pada data Riskesdes tahun 2018 terdapat 1,79 per 1.000 penduduk mengidap penyakit kanker. Akibat tingginya prevalensi kanker maka diperlukan pendeteksian kanker sejak dini. Salah satu cara mendeteksi kanker yaitu dengan teknologi *microarray* dimana teknologi ini dapat memantau ribuan ekpresi gen secara bersamaan dalam satu percobaan. Namun, data *microarray* memiliki dimensi besar sehingga diperlukan proses reduksi dimensi data *microarray* pada penyakit *prostate cancer da gastric cancer* agar dapat menghilangkan atribut yang redundansi dan meningkatkan akurasi pada klasifikasi. Reduksi dilakukan menggunakan MRMR (FCQ dan FCD) dengan *k* 10,20,30,40,50,60,70,80,90 dan 100. Klasifikasi dilakukan menggunakan RF dengan membentuk 100 *tree*. Hasil akurasi terbaik pada klasifikasi data *prostate cancer* yaitu dengan FCQ 100% pada *k*=10, tanpa reduksi 95% dan akurasi terendah dengan FCD 52% pada *k*=90. Sedangkan hasil akurasi terbaik klasifikasi data *gastric cancer* yaitu dengan FCQ dan FCD 100% pada semua *k* dan akurasi terendah yaitu tanpa reduksi 83%.

Keywords

Microarray, Random Forets, MRMR, Dimensional Reduction

Abstrak

In Indonesia, the prevalence of cancer in the 2018 Riskesdes data was 1.79 per 1,000 population with cancer. Due to the high prevalence of cancer, it's necessary to detect cancer early. One way to detect cancer is by using microarray technology, simultaneously monitoring thousands of gene expressions in one experiment. However, microarray data have large dimensions, so it's necessary to reduce the dimensions of microarray data in prostate and gastric cancer to eliminate redundant attributes and improve classification accuracy. The reduction was carried out using MRMR (FCQ and FCD) with k 10,20,30,40,50,60,70,80,90, and 100. The classification was carried out using RF by forming 100 trees. The best accuracy results in classifying prostate cancer data are with FCQ 100% at k=10, without reduction 95%, and the lowest accuracy with FCD 52% at k=90. While the best accuracy for gastric cancer data classification is FCQ and FCD 100% for all k, and the lowest accuracy is without reduction 83%.

A. Pendahuluan

Kanker merupakan penyakit yang tidak menular dengan ditandai adanya sel/jaringan abnormal yang bersifat ganas. Kanker memiliki pertumbuhan yang tidak dapat dikendalikan kecepatannya dan dapat menyebar ke sel/jaringan lainnya di dalam tubuh penderita.[1] Di Indonesia prevalensi kanker pada data Riskesdes tahun 2018 terdapat 1,79 per 1.000 penduduk mengidap penyakit kanker.[2] Akibat tingginya prevalensi kanker tersebut maka diperlukan pendeteksian kanker sejak dini. Adapun salah satu cara dalam mendeteksi penyakit kanker yaitu dengan ekspresi gen menggunakan teknologi *microarray* yang mana teknologi ini dapat memantau ribuan ekpresi gen secara bersamaan dalam satu percobaan.[3] Namun, pada data *microarray* memiliki dimensi yang sangat besar *(curse of dimensionality)* sehingga diperlukannya proses reduksi dimensi pada data *microarray* dengan tujuan untuk menghilangkan atribut/fitur yang kurang relevan dan untuk meningkatkan nilai akurasi pada proses klasifikasi data *microarray* nantinya.[4]

Proses reduksi dimensi dapat dilakukan dengan seleksi fitur menggunakan *Minimum Redundancy Maximum Relevance* (MRMR). MRMR merupakan metode yang dapat memberikan informasi relevansi dan redundansi suatu sub-himpunan. [3] MRMR dirancang untuk menganalisis kualitas dan memberikan kinerja prediktif terbaik pada subset variabel data (atribut kelas).[5] Tujuan utama dari MRMR yaitu untuk memilih fitur terbaik dan mengurangi fitur yang sama. Metode ini menangani setiap fitur secara terpisah dari kumpulan beberapa data dan menggunakan informasi timbal balik di antara fitur-fitur yang ada, dimana dapat juga digunakan untuk mengukur tingkat kesamaan antara dua fitur.[6] MRMR merupakan metode pencarian tambahan yang dapat mengintegrasikan relevansi dan redundansi ke dalam fungsi tujuan tunggal dengan tujuan untuk memaksimalkan relevansi dan meminimalkan redundansi.[7] MRMR juga memiliki tujuan yaitu untuk mempertahankan korelasi maksimal antara variabel-variabel yang dipilih dalam membentuk sub-blok yang masuk akal. MRMR juga memiliki kriteria dalam menemukan variabel dari kumpulan data yang sama-sama memiliki ketergantungan terbesar pada variabel target.[8]

MRMR merupakan metode yang digunakan untuk melakukan penyaringan redundansi di antara redundansi atribut yang dipilih ketika mencoba untuk memilih atribut yang paling mirip dengan tag kelas.[9] Adapun tahapan yang dilakukan oleh metode MRMR seperti berikut:[4]

1. Mencari nilai F-test dan korelasi pearson

$$F_{Test} = \frac{MS_{Between}}{MS_{Within}}$$

$$r = \frac{n\sum(x_1x_2) - \sum(x_1)\sum(x_2)}{\sqrt{[n\sum x_1^2 - (\sum x_1)^2][n\sum x_2^2 - (\sum x_2)^2]}}$$

2. Mencari nilai *F-test Correlation Quotient* (FCQ) dan *F-test Correlation Difference* (FCD) karena data ang digunakan berupa data yang kontiniu.[10]

$$FCQ = \max(F/r)$$

$$FCQ = \max(F - r)$$

Setelah dilakukan proses reduksi dimensi dengan seleksi fitur pada data *microarray*, kemudian dilakukan proses klasifikasi dengan menggunakan *Random Forest* (RF). RF merupakan metode yang digunakan untuk melakukan klasifikasi pada data mining dan *machine learning*. *Random Forest* dapat memetakan atribut dari kelas sehingga dapat digunakan dalam menemukan prediksi terhadap data yang belum muncul.[11] RF merupakan metode klasifikasi yang memiliki banyak pohon *(tree)* sehingga membentuk hutan *(forest)* yang mana setiap pohon keputusan dibangun menggunakan vektor acak.[12] Klasifikasi pada RF didapatkan dengan cara *voting* (jumlah terbanyak) dari pohon-pohon klasifikasi yang telah dibentuk. Pohon klasifikasi dibuat dari kumpulan simpul yang saling berkaitan. Setiap simpul mewakili satu keputusan, yang mana setiap keputusan berdasarkan pada salah satu variabel prediktor agar dapat menentukan kelas.[13] penentuan simpul pada *tree* dapat menggunakan *Entropy* dan *Information Gain*.[14]

$$Entropy\ (S) = -\sum_i^c P_i\ log_2\ P_i$$

$$IG(S,a) = Entropy(S) - \Sigma_v \frac{|s_v|}{|S|} Entropy(s_v)$$

Setelah dilakukan klasifikasi menggunakan RF, langkah selanjutnya yaitu mencari nilai akurasi dengan *confusion matrix*.[3]

$$Akurasi\ (x) = \frac{TP+TN}{TP+FN+FP+TN} \times 100\%$$

Sebelumnya, telah dilakukan beberapa penelitian tentang reduksi dimensi dan klasifikasi pada data *microarray* diantaranya yaitu penelitian yang dilakukan oleh [4] menggunakan metode reduksi dimensi MRMR dengan persamaan FCQ dan FCD. Adapun hasil akurasi yang didapat FCQ = 83,87% dan FCD = 61,29% menggunakan metode klasifikasi RF. Pada penelitian [3] menggunakan metode reduksi dimensi MRMR dan optimasi GA yang dilakukan pada data *microarray* memiliki kenaikan akurasi pada data Colon Tumor dari 5,55% - 16,66%, pada data Lung Cancer dari 2,78% - 5,56% dan pada data Ovarian Cancer dari 2% - 2,67% dengan menggunakan metode klasifikasi FLNN. Pada penelitian yang dilakukan oleh [15] metode reduksi dimensi yang digunakan yaitu Relief dan metode klasifikasi yang digunakan yaitu SVM, ANN, K-NN, C4.5 dan RF dengan hasil akurasi tertinggi 75% pada metode klasifikasi RF. Pada penelitian [16] proses klasifikasi menggunakan RF dan reduksi dimensi menggunakan MRMR serta optimasi data menggunakan IFS dengan hasil akurasi 0,672997 dan MCC 0,347977. Pada penelitian yang dilakukan [13] proses dilakukan untuk mendiagnosa penyakit epilepsi yaitu menggunakan rekaman EEG. Namun, membutuhkan waktu yang cukup lama sehingga diperlukan proses *preprocessing* data menggunakan DWT dan LLF serta diperlukannya proses klasifikasi menggunakan RF dan SVM. Hasil akurasi yang lebih baik pada data *training* yaitu menggunakan RF dan pada data *testing* yaitu SVM.

Dari yang tertulis di atas memiliki tujuan untuk mereduksi dimensi atribut data *microarray* pada penyakit *prostate cancer* dan *gastric cancer* menggunakan *Minimum Redundancy Maximum Relevance* dan melakukan klasifikasi menggunakan *Random Forets* sehingga dapat diketahui seberapa baik akurasi yang dilakukan pada

klasifikasi RF tanpa reduksi dan menggunakan reduksi MRMR dengan persamaan *F-statistic Correlation Quotient* dan *F-statistic Correlation Difference.*

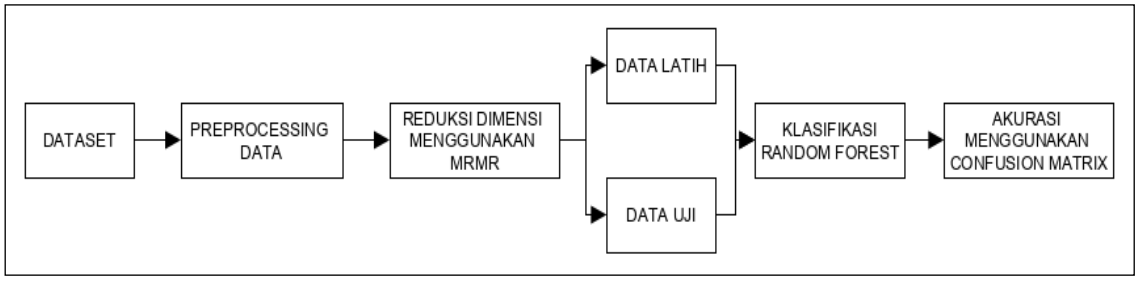
B. Metode Penelitian

Data yang digunakan dalam penelitian ini yaitu data *microarray* dengan menggunakan dua jenis data kanker yaitu *Prostate cancer* dan *Gastric Cancer*. Data diperoleh dari *Bioinformatics Laboratory* pada *website:* www.biolab.si/supp/bi-cancer/projections/info/. Adapun jumlah data digunakan seperti pada Tabel 1.

Tabel 1. Spesifikasi data *microarray*

Nama Data	Jumlah Data	Gen/Dimensi Data	Kelas Positive	Kelas Negative
Prostate Cancer	102	12533	Tumor	Normal
Gastric Cancer	30	4522	Tumor	Normal

Tahapan-tahapan yang akan dilakukan pada metodologi penelitian ini seperti gambar 1 berikut:



Gambar 1. *Framework* reduksi dimensi dataset

Dataset akan dinormalisasikan meggunakan normalisasi *min-max* dengan persamaan berikut:

$$X_{ni} = \frac{X_i - \min(X)}{\max(X) - \min(X)}$$

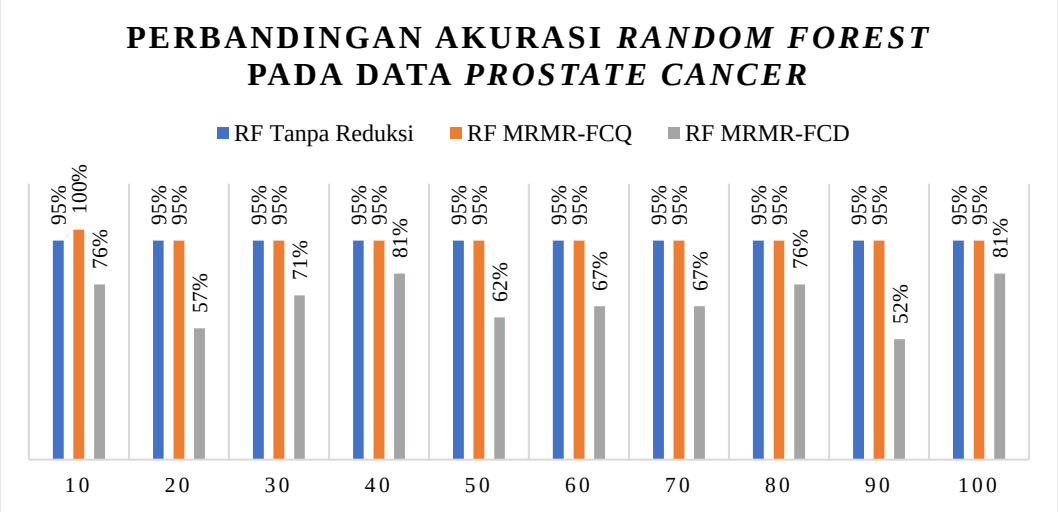
Hasil dari normalisasi pada dataset akan dilakukan reduksi dimensi menggunakan MRMR pada persamaan FCQ dan FCD dengan nilai *k best features* 10, 20, 30, 40, 50, 60, 70, 80, 90 dan 100. Setelah direduksi data akan dibagi menjadi dua bagian yaitu data latih dan data uji dengan perbandingan 80:20 kemudian akan dilakukan klasifikasi data menggunakan RF dengan membangun 100 n_estimator (*tree*), tahapan terakhir akan dilakukan perbandingan akurasi pada klasifikasi RF tanpa reduksi dan klasifikasi RF dengan reduksi MRMR-FCQ dan MRMR-FCD.

C. Hasil dan Pembahasan

Tahapan ini dilakukan pengujian perbandingan hasil akurasi klasifikasi RF tanpa reduksi dan dengan reduksi menggunakan MRMR-FCQ dan MRMR-FCD pada data *prostate cancer* seperti berikut:

Tabel 2. Perbandingan akurasi pada data *prostate cancer*

k best features	RF Tanpa Reduksi	RF MRMR-FCQ	RF MRMR-FCD
10	0.9523809	1,0	0.76190476
20	0.9523809	0.9523809	0.57142857
30	0.9523809	0.9523809	0.71428571
40	0.9523809	0.9523809	0.80952380
50	0.9523809	0.9523809	0.61904761
60	0.9523809	0.9523809	0.66666666
70	0.9523809	0.9523809	0.66666666
80	0.9523809	0.9523809	0.76190476
90	0.9523809	0.9523809	0.52380952
100	0.9523809	0.9523809	0.80952380



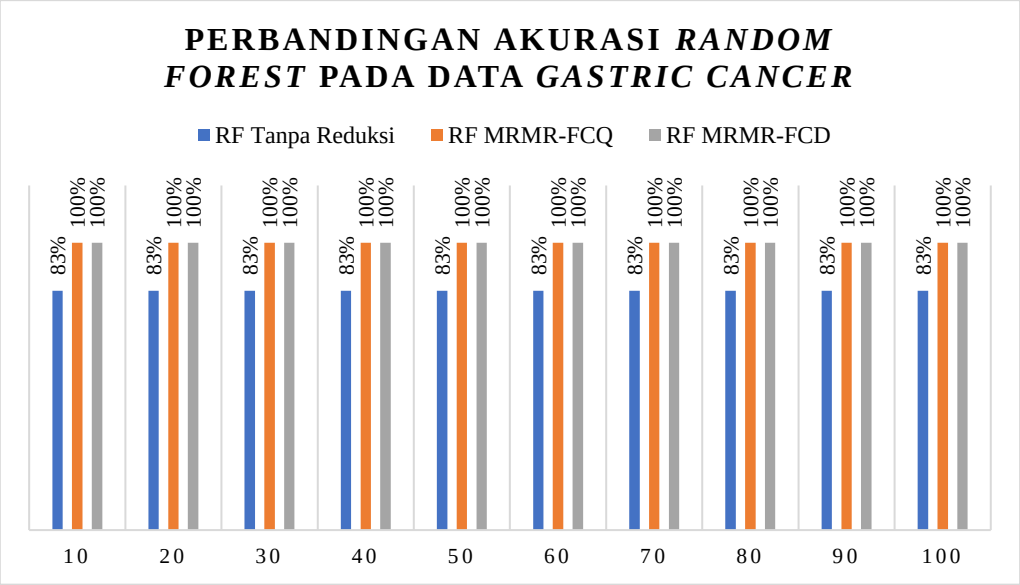
Gambar 2. Grafik perbandingan akurasi pada data *prostate cancer*

Dari gambar 2 hasil akurasi tertinggi terdapat pada RF dengan reduksi MRMR-FCQ (k = 10) sebebsar 100%. Namun, hasil akurasi RF tanpa reduksi dan dengan reduksi MRMR-FCQ pada k = 20 sampai dengan k = 100 memiliki hasil akurasi yang sama sebesar 95,24% dan hasil akurasi terendah terdapat pada RF dengan reduksi MRMR-FCD k = 90 sebesar 52,38%

Pada tahapan ini akan dilakukan perbandingan hasil akurasi klasifikasi RF tanpa reduksi dan dengan reduksi menggunakan MRMR-FCQ dan MRMR-FCD pada data *gastric cancer* seperti berikut:

Tabel 2. Perbandingan akurasi pada data *gastric cancer*

k best features	RF Tanpa Reduksi	RF MRMR-FCQ	RF MRMR-FCD
10	0.8333333	1.0	1.0
20	0.8333333	1.0	1.0
30	0.8333333	1.0	1.0
40	0.8333333	1.0	1.0
50	0.8333333	1.0	1.0
60	0.8333333	1.0	1.0
70	0.8333333	1.0	1.0
80	0.8333333	1.0	1.0
90	0.8333333	1.0	1.0
100	0.8333333	1.0	1.0



Gambar 3. Grafik perbandingan akurasi pada data *gastric cancer*

Dari gambar 3 hasil akurasi tertinggi yaitu RF dengan reduksi MRMR-FCQ dan MRMR-FCD sebesar 100% pada semua nilai *k best features* dan terendah yaitu RF tanpa reduksi sebesar 83%.

D. Simpulan

Pada data *prostate cancer* akurasi terbaik terdapat pada klasifikasi dengan reduksi MRMR-FCQ dengan nilai *k best features* 10 yaitu 100% sedangkan dengan *k best fetaures* 20 sampai dengan 100 hasil akurasinya tidak mengalami peningkatan ataupun penurunan yaitu 95% dan akurasi terendah terdapat pada klasifikasi dengan reduksi MRMR-FCD dengan nilai *k best features* 90 yaitu 52%. Adapun urutan akurasi terbaik yaitu RF MRMR-FCQ kemudian RF MRMR-FCD dan RF tanpa reduksi.

Pada data *gastric cancer* akurasi terbaik terdapat pada klasifikasi dengan reduksi MRMR-FCQ dan MRMR-FCD yaitu 100% pada semua nilai *k best features* yaitu 10 sampai dengan 100. Dan untuk klasifikasi RF tanpa reduksi yaitu 83%.

Hasil akurasi tertinggi pada klasifikasi RF dari ketiga data *microarray* yaitu menggunakan MRMR-FCQ. Namun, untuk akurasi pada klasifikasi RF MRMR-FCD dan RF tanpa reduksi tergantung data yang digunakan dan jumlah *k best features*.

E. Ucapan Terima Kasih

Terima kasih kepada seluruh sivitas akademika Universitas Sumatera Utara, Fakultas Ilmu Komputer dan Teknologi Informasi atas dukungan dan partisipasinya.

F. Referensi

[1] Kemenkes, "Apa itu Kanker?," Jakarta, 2019. [Online]. Available: https://p2ptm.kemkes.go.id/infographic-p2ptm/penyakit-kanker-dan-kelainan-darah/apa-itu-kanker.

[2] Kemenkes, "Hari Kanker Sedunia 2019," Jakarta, 2019. [Online]. Available: https://www.kemkes.go.id/article/view/19020100003/hari-kanker-sedunia-2019.html.

[3] B. Pradana and A. Aditsania, "Implementasi Minimum Redudancy Maksimum Relevance (MRMR) dan Genetic Algorithm (GA) untuk Reduksi Dimensi pada Klasifikasi Data Micorarray Menggunakan Functional Link Neural Network (FLNN)," vol. 6, no. 2, pp. 8966–8977, 2019.

[4] I. G. N. P. V. Geramona and W. Astuti, "Implementasi Minimum Redundancy Maximum Relevance sebagai Teknik Reduksi Dimensi pada Klasifikasi Kanker Usus Besar Menggunakan Random Forest," vol. 7, no. 1, pp. 2490–2497, 2020.

[5] J. Taveira De Souza, A. Carlos De Francisco, and D. C. De Macedo, "Dimensionality Reduction in Gene Expression Data Sets," *IEEE Access*, vol. 7, pp. 61136–61144, 2019, doi: 10.1109/ACCESS.2019.2915519.

[6] M. Toğaçar, B. Ergen, Z. Cömert, and F. Özyurt, "A Deep Feature Learning Model for Pneumonia Detection Applying a Combination of mRMR Feature Selection and Machine Learning Models," *Irbm*, vol. 41, no. 4, pp. 212–222, 2020, doi: 10.1016/j.irbm.2019.10.006.

[7] T. Gangavarapu and N. Patil, "A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets," *Appl. Soft Comput. J.*, vol. 81, p. 105538, 2019, doi: 10.1016/j.asoc.2019.105538.

[8] K. Zhong, D. Ma, and M. Han, "Distributed dynamic process monitoring based on dynamic slow feature analysis with minimal redundancy maximal relevance," *Control Eng. Pract.*, vol. 104, no. September, p. 104627, 2020, doi: 10.1016/j.conengprac.2020.104627.

[9] M. Toğaçar, B. Ergen, and Z. Cömert, "Detection of lung cancer on chest CT images using minimum redundancy maximum relevance feature selection method with convolutional neural networks," *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 23–39, 2020, doi: 10.1016/j.bbe.2019.11.004.

[10] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *J. Bioinform. Comput. Biol.*, vol. 3, no. 2, pp. 1–21, 2005.

[11] A. U. Zailani and N. L. Hanun, "Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera," *Infotech J. Technol. Inf.*, vol. 6, no. 1, pp. 7–14, 2020, doi: 10.37365/jti.v6i1.61.

[12] M. Triyani, A. Adiwijaya, and A. Aditsania, "Discrete Wavelet Transform (DWT) and Random Forest for Cancer Detection Based on Microarray Data Classification," *J. Infotel*, vol. 12, no. 3, pp. 97–104, 2020, doi: 10.20895/infotel.v12i3.484.

[13] M. I. Fachruddin, "Perbandingan Metode Random Forest Classification Dan Support Vector Machine Untuk Deteksi Epilepsi Menggunakan Data Rekaman Electroencephalograph (EEG)," *Fak. Mat. dan Ilmu Pengetah. Alam Inst. Teknol. Sepuluh Nop.*, pp. 1–83, 2015.

[14] Riska Chairunisa, Adiwijaya, and Widi Astuti, "Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 805–812, 2020, doi: 10.29207/resti.v4i5.2083.

[15] H. Aydadenta and Adiwijaya, "On the classification techniques in data mining for microarray data classification," *J. Phys. Conf. Ser.*, vol. 971, no. 1, 2018, doi: 10.1088/1742-6596/971/1/012004.

[16] B. Q. Li, K. Y. Feng, L. Chen, T. Huang, and Y. D. Cai, "Prediction of Protein-Protein Interaction Sites by Random Forest Algorithm with mRMR and IFS," *PLoS One*, vol. 7, no. 8, pp. 1–10, 2012, doi: 10.1371/journal.pone.0043927.