JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 2, April 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 72-77 | DOI: 10.30865/mib.v3i2.1161

Principal Component Analysis Sebagai Ekstraksi Fitur Data Microarray Untuk Deteksi Kanker Berbasis Linear Discriminant Analysis

Widi Astuti*, Adiwijaya

Fakultas Informatika, Universitas Telkom, Bandung 40257, Indonesia
Email: 1*astutiwidi@telkomuniversity.ac.id, 2adiwijaya@telkomuniversity.ac.id

Abstrak

Kanker merupakan salah satu penyebab kematian terbesar di dunia. Deteksi dini kanker memungkinkan penanganan yang lebih baik bagi penderitanya. Salah satu cara mendeteksi kanker adalah menggunakan klasifikasi data microarray. Akan tetapi, data microarray memiliki dimensi yang tinggi sehingga mempersulit proses klasifikasi. Linear Discriminant Analysis adalah sebuah teknik klasifikasi yang mudah diimplementasikan dan memiliki akurasi yang bagus. Akan tetapi, Linear Discriminant Analysis memiliki kesulitan dalam menangani data berdimensi tinggi. Oleh karena itu, digunakan Principal Component Analysis, sebuah teknik ekstraksi fitur untuk mengoptimasi kinerja Linear Discriminant Analysis. Berdasarkan hasil penelitian, diperoleh bahwa penggunaan Principal Component Analysis meningkatkan akurasi hingga 29,04% dan f-1 score sebesar 64,28% untuk data tumor usus besar.

Kata Kunci: data microarray, ekstraksi fitur, kanker usus, Linear Discriminant Analysis, Principal Component Analysis

Abstract

Cancer is one of the leading causes of death globally. Early detection of cancer allows better treatment for patients. One method to detect cancer is using microarray data classification. However, microarray data has high dimensions which complicates the classification process. Linear Discriminant Analysis is a classification technique which is easy to implement and has good accuracy. However, Linear Discriminant Analysis has difficulty in handling high dimensional data. Therefore, Principal Component Analysis, a feature extraction technique is used to optimize Linear Discriminant Analysis performance. Based on the results of the study, it was found that usage of Principal Component Analysis increases the accuracy of up to 29.04% and f-1 score by 64.28% for colon cancer data.

Keywords: microarray data, feature extraction, colon cancer, Linear Discriminant Analysis, Principal Component Analysis

1. PENDAHULUAN

Menurut laporan Kementerian Kesehatan RI pada 2015 [1], salah satu penyebab kematian utama di seluruh dunia adalah penyakit kanker. Pada tahun 2012 saja, terdapat 14 juta kasus kanker dan 8,2 juta kematian akibat kanker. Diperkirakan kasus kanker tahunan akan meningkat menjadi 22 juta kasus dalam dua dekade berikutnya. Kanker kolorektal dan kanker payudara adalah dua dari penyebab terbesar kematian akibat kanker setiap tahunnya. Namun demikian, kanker yang terdeteksi lebih awal memiliki peluang untuk mendapatkan penanganan yang lebih baik[1].

Dalam beberapa tahun terakhir, terdapat suatu teknologi bernama microarray yang memungkinkan pemantauan terhadap ribuan ekspresi gen dalam satu waktu [2]. Teknologi ini memungkinkan banyak eksperimen dalam bidang biologi sehingga data microarray dapat diterapkan dalam banyak permasalahan [3], salah satunya adalah deteksi kanker [4]. Deteksi kanker menggunakan data microarray dilakukan dengan menggunakan teknik klasifikasi untuk menentukan apakah sampel termasuk ke dalam kelas kanker atau tidak. Akan tetapi, data microarray memiliki dimensi tinggi karena mengandung ribuan fitur dengan sedikit sampel sehingga klasifikasi data microarray menjadi sulit dilakukan [5]. Untuk menangani klasifikasi data berdimensi tinggi, biasanya dilakukan reduksi dimensi. Reduksi dimensi terbagi ke dalam dua jenis yaitu seleksi fitur dan ekstraksi fitur.

Beberapa penelitian telah dilakukan untuk deteksi kanker menggunakan klasifikasi data microarray. Vanitha [6] menggunakan Support Vector Machine dan Mutual Information pada kanker usus besar dan limfoma. Pada data limfoma, akurasi yang diperoleh berkisar antara 86.36% - 100% sementara pada data kanker usus besar, akurasinya adalah 38.70% hingga 74.19%. Aydadenta pada 2018 [4] melakukan klasifikasi menggunakan random forest yang dikombinasikan dengan seleksi fitur berbasis clustering. Akurasi yang diperoleh adalah 98.9% untuk kanker paru, 89% untuk kanker prostat, dan 85.87% untuk kanker usus besar. Pada [7], dilakukan klasifikasi data kanker usus besar menggunakan k-Nearest Neighbor dan reduksi dimensi menggunakan Independent Component Analysis. Hasilnya menunjukkan bahwa penggunaan reduksi dimensi meningkatkan akurasi dari 77.4% menjadi 88.7% dan sensitivity dari 36.4% menjadi 72.7%. Lebih jauh,Wisesty [8] menggunakan kombinasi Genetic Algorithm dan Momentum Backpropagation untuk melakukan deteksi kanker usus besar dan leukemia. Pada penelitian tersebut, didapatkan bahwa penggunaan reduksi dimensi meningkatkan akurasi deteksi kanker usus besar sebesar 33,81%. Di waktu yang sama, dilakukan juga penelitian menggunakan mutual information dan Naive Bayes serta Bayesian Network [9] untuk lima data kanker. Hasilnya, pada data kanker usus besar, akurasi yang diperoleh berada pada kisaran 66.67%-80%.

Menurut Yip [10], terdapat 6 jenis teknik klasifikasi yang bisa digunakan, salah satunya adalah discriminant analysis yang mudah diimplementasikan dan memiliki akurasi yang baik tapi kurang sesuai bagi data berdimensi tinggi. Di sisi lain, Principal Component Analysis (PCA) adalah salah satu metode reduksi

JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 2, April 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 72-77 | DOI: 10.30865/mib.v3i2.1161

dimensi jenis ekstraksi fitur yang bisa digunakan dalam mereduksi dimensi data microarray [11]. Dalam penelitian ini, dilakukan klasifikasi data microarray menggunakan Linear Discriminant Analysis (LDA) yang dikombinasikan dengan reduksi dimensi PCA untuk deteksi kanker usus besar. Penggunaan PCA diharapkan mampu menaikkan kinerja LDA dan menghasilkan akurasi yang tinggi untuk data kanker usus besar.

2. TINJAUAN PUSTAKA

2.1 Klasifikasi Data Microarray

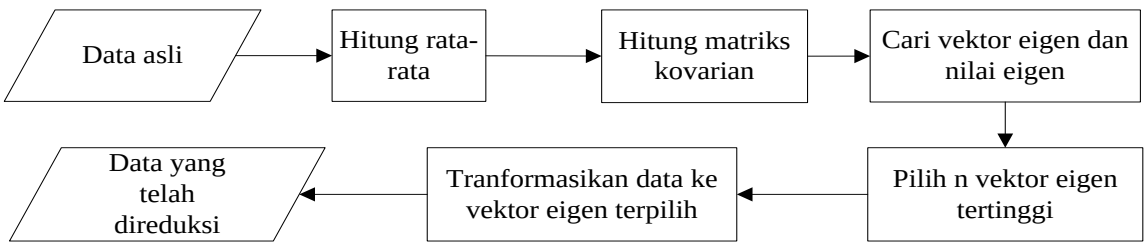
Terdapat beberapa teknik yang bisa digunakan untuk melakukan klasifikasi data microarray. Berdasarkan [10], terdapat 6 teknik klasifikasi yang memiliki kekurangan dan kelebihan sebagaimana ditampilkan dalam Tabel 1 berikut.

Tabel 1. Perbandingan teknik klasifikasi data

Teknik Klasifikasi	Kelebihan	Kekurangan
Decision Tree	• Mudah diimplementasikan • Mudah dipahami	• Rentan terhadap overfitting • Akurasi rendah • Sulit beradaptasi dengan data baru
k-Nearest Neighbor	• Mudah diimplementasikan • Mudah dipahami • Mudah beradaptasi dengan data	• Akurasi rendah • Biaya komputasinya tinggi
Discriminant Analysis	• Mudah diimplementasikan • Akurasi tinggi	• Tidak cocok untuk data dengan dimensi tinggi
Bayesian Network	• Representative power yang tinggi • Bisa menjadi basis bagi teknik klasifikasi lain	• Sulit dilatih • Sulit dipahami • Biaya komputasi bisa mahal
Jaringan Saraf Tiruan	• Akurasi yang tinggi bahkan untuk data berdimensi tinggi atau data yang mengandung noise	• Waktu learning lama • Memerlukan parameter yang sebaiknya ditentukan secara empiris
Support Vector Machine	• Akurasi tinggi • Tidak rentan terhadap over fitting • Mampu memodelkan data yang kompleks	• Biaya komputasi bisa mahal • Memerlukan keahlian untuk memilih kernel dan parameter yang sesuai untuk tiap kasus

Dari Tabel 1 tersebut, terlihat bahwa setiap teknik klasifikasi memiliki karakteristinya masing-masing. Discriminant Analysis dipilih dalam penelitian ini karena memiliki kelebihan mudah diimplementasikan dan akurasinya tinggi. Salah satu algoritma Discriminant Analysis adalah Linear Discriminant Analysis (LDA) yang bisa digunakan untuk melakukan klasifikasi pada data dengan kelas biner. LDA adalah sebuah teknik klasifikasi sederhana tetapi robust yang bekerja dengan mecari kombinasi linear dari fitur-fitur yang ada untuk memisahkan sampel ke dalam kelas yang sesuai [12]. Akan tetapi, LDA memiliki kelemahan yaitu sebaiknya tidak digunakan untuk data dengan sampel yang lebih kecil dari jumlah fiturnya[13]. Oleh karena itu, diperlukan sebuah teknik reduksi dimensi agar LDA bisa memiliki kinerja yang baik saat melakukan klasifikasi data microarray.

2.2. Reduksi Dimensi



Gambar 1. Skema PCA

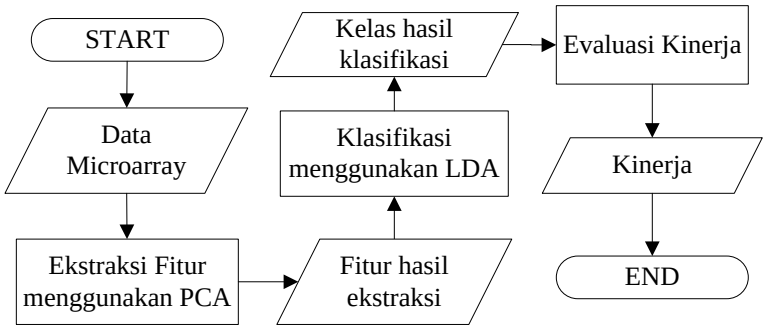
Data microarray yang memiliki ukuran sampel yang kecil dengan jumlah fitur yang tinggi. Menurut [14], hal ini akan menimbulkan permasalahan yang kompleks pada proses klasifikasi yang biasa disebut sebagai curse of dimensionality. Untuk menyelesaikan masalah tersebut, yang sering digunakan adalah reduksi dimensi. Reduksi dimensi terbagi menjadi dua jenis yaitu seleksi fitur dan ekstraksi fitur. Seleksi fitur dilakukan dengan memilih beberapa fitur yang dianggap penting untuk klasifikasi. Sementara itu, ekstraksi fitur dilakukan dengan memproyeksikan data ke dalam fitur baru yang berjumlah lebih sedikit tetapi tetap mencerminkan data aslinya. Salah satu metode ekstraksi fitur yang terkenal adalah Principal Component Analysis (PCA). Langkah-langkah

JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 2, April 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 72-77 | DOI: 10.30865/mib.v3i2.1161

dalam PCA terdiri dari: menghitung rata-rata, menghitung matriks kovarian, menghitung vektor eigen dan nilai eigen, memilih n vektor eigen tertinggi, dan transformasi data. Skema metode PCA digambarkan pada gambar 1 di atas. Langkah lebih lengkap mengenai penggunaan PCA untuk reduksi dimensi dapat ditelusuri di [15].

3. DESAIN SISTEM

Sistem yang dibangun pada penelitian ini digambarkan skemanya pada Gambar 2 berikut.



Gambar 2. Skema penelitian

Masukan untuk penelitian ini adalah data microarray yaitu data tumor usus besar yang diperoleh dari Kent Ridge Biomedical Data Set Repository yang disimpan di repository ELVIRA [16]. Data tersebut terdiri dari 62 sampel dengan 40 sampel berlabel positif dan 22 sampel berlabel negatif. Jumlah fitur pada data tersebut adalah sebanyak 2000. Data tersebut lalu dikurangi fiturnya menggunakan metode esktraksi fitur PCA dan diklasifikasi menggunakan LDA. Evaluasi kinerja dilakukan dengan menggunakan confusion matrix dan mencatat nilai True Positive (TP), False Positive (FP), True Negative (TN), dan False Negative (FN). Selanjutnya, dihitung nilai akurasi, sensitivitas, precision dan f-1 score.

4. IMPLEMENTASI

Pengujian dilakukan dengan menggunakan k-fold cross validation dengan nilai k=10. Hal ini dilakukan untuk memaksimalkan data yang tersedia dan memastikan bahwa kinerja sistem terukur dengan baik. Terdapat tiga pengujian yang dilakukan.

4.1 Pengujian dan Analisis Varians pada PCA

Pengujian dan analisis variansi pada PCA dilakukan untuk melihat pengaruh jumlah fitur yang dipilih setelah transformasi PCA. Pada tabel 2, ditampilkan varians tiap fitur hasil transformasi PCA yang telah diurutkan dari varians tertinggi ke varians terendah. Dari tabel 2 terlihat bahwa jika hanya digunakan 1 fitur hasil PCA, 1 fitur tersebut mewakili 36,10% varians dari data awal. Jika digunakan 2 fitur hasil PCA, maka 2 fitur tersebut mewakili 48,45% varians dari data awal.

Tabel 2. Varians Fitur Hasil PCA

Fitur ke-	Explained Variance	Fitur ke-	Explained Variance	Fitur ke-	Explained Variance	Fitur ke-	Explained Variance
1	36.10%	17	0.72%	33	0.20%	49	0.08%
2	12.35%	18	0.63%	34	0.20%	50	0.08%
3	9.91%	19	0.61%	35	0.19%	51	0.07%
4	7.65%	20	0.56%	36	0.17%	52	0.07%
5	5.31%	21	0.49%	37	0.16%	53	0.06%
6	3.51%	22	0.48%	38	0.16%	54	0.06%
7	2.82%	23	0.45%	39	0.15%	55	0.05%
8	2.49%	24	0.42%	40	0.14%	56	0.05%
9	2.11%	25	0.39%	41	0.12%	57	0.05%
10	1.93%	26	0.34%	42	0.12%	58	0.04%
11	1.42%	27	0.33%	43	0.12%	59	0.03%
12	1.31%	28	0.31%	44	0.10%	60	0.02%
13	0.98%	29	0.30%	45	0.10%	61	0.02%

JURNAL MEDIA INFORMATIKA BUDIDARMA, Vol 3, No 2, April 2019
ISSN 2614-5278 (media cetak)
ISSN 2548-8368 (media online)
Hal 72-77 | DOI: 10.30865/mib.v3i2.1161

Fitur ke-	Explained Variance	Fitur ke-	Explained Variance	Fitur ke-	Explained Variance	Fitur ke-	Explained Variance
14	0.89%	30	0.28%	46	0.09%	62	0.00%
15	0.79%	31	0.24%	47	0.09%	....	....
16	0.77%	32	0.22%	48	0.09%	2,000	0.00%

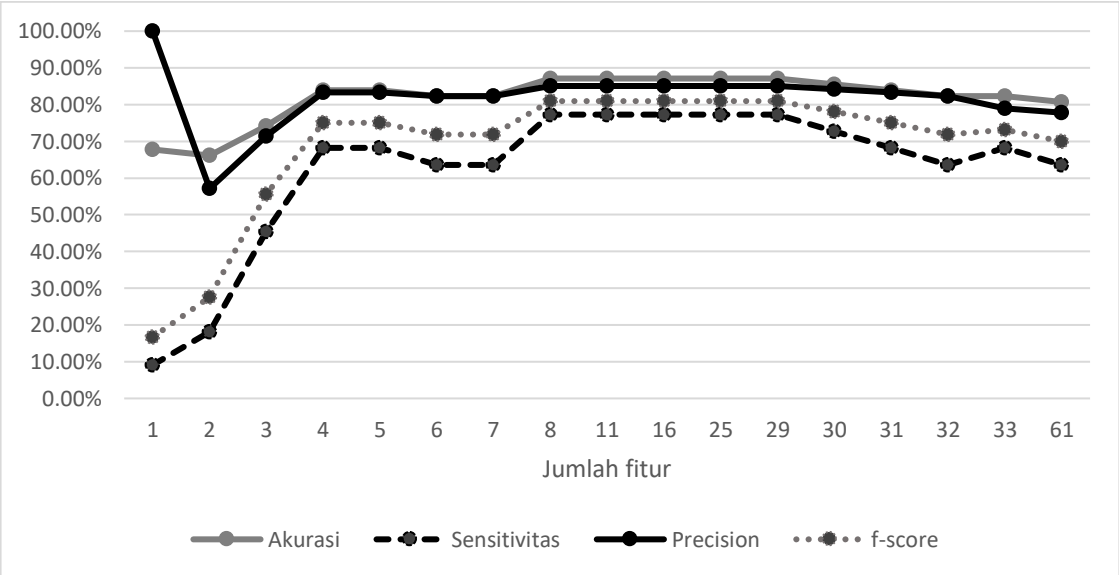
Berdasarkan hasil percobaan, diperoleh bahwa dari 2000 fitur asal,ketika ditransformasikan menggunakan PCA, cukup dengan 61 fitur telah memberikan varians total 100%. Artinya, PCA dalam kasus tumor usus besar mampu mengurangi jumlah fitur yang diperlukan secara signifikan untuk proses klasifikasi. Dengan demikian, pengurangan fitur ini dapat memberikan input data yang sesuai untuk LDA.

4.2 Pengujian dan Analisis Jumlah Fitur yang Diekstraksi menggunakan PCA

Percobaan kedua dilakukan untuk melihat pengaruh jumlah fitur yang dipilih setelah melalui ekstraksi fitur menggunakan PCA. Terdapat 17 skenario jumlah fitur yang berbeda yang diujikan yang dipilih secara empiris. Setiap skenario dihitung jumlah TP, FP, TN, dan FN untuk selanjutnya dihitung nilai akurasi, sensitivitas, precision dan f-1 score-nya. Hasil pengujian ditunjukkan pada Tabel 3 dan ditampilkan trennya pada Gambar 3.

Tabel 3. Kinerja Klasifikasi Menggunakan PCA

Jumlah fitur	TP	FP	TN	FN	Akurasi	Sensitivitas	Precision	f-score
1	2	0	40	20	67,74%	9,09%	100,00%	16,67%
2	4	3	37	18	66,13%	18,18%	57,14%	27,59%
3	10	4	36	12	74,19%	45,45%	71,43%	55,56%
4	15	3	37	7	83,87%	68,18%	83,33%	75,00%
5	15	3	37	7	83,87%	68,18%	83,33%	75,00%
6	14	3	37	8	82,26%	63,64%	82,35%	71,79%
7	14	3	37	8	82,26%	63,64%	82,35%	71,79%
8	17	3	37	5	87,10%	77,27%	85,00%	80,95%
11	17	3	37	5	87,10%	77,27%	85,00%	80,95%
16	17	3	37	5	87,10%	77,27%	85,00%	80,95%
25	17	3	37	5	87,10%	77,27%	85,00%	80,95%
29	17	3	37	5	87,10%	77,27%	85,00%	80,95%
30	16	3	37	6	85,48%	72,73%	84,21%	78,05%
31	15	3	37	7	83,87%	68,18%	83,33%	75,00%
32	14	3	37	8	82,26%	63,64%	82,35%	71,79%
33	15	4	36	7	82,26%	68,18%	78,95%	73,17%
61	14	4	36	8	80,65%	63,64%	77,78%	70,00%



Gambar 3. Kinerja Klasifikasi Menggunakan PCA + LDA

Dari hasil percobaan, terlihat bahwa kinerja klasifikasi meningkat seiring bertambahnya jumlah fitur, selanjutnya kinerja tidak berubah ketika jumlah fitur yang digunakan berkisar antara 8 sampai dengan 29. Ketika jumlah fitur lebih dari 29, kinerja justru mengalami penurunan. Ketika hanya 1 fitur yang digunakan, precision mencapai 100%. Akan tetapi, saat melakukan klasifikasi penyakit, sensitivitas lebih dipertimbangkan. Lebih jauh,ketika jumlah fitur =1, , f-1 score 16,67% yang berarti nilai precision yang dicapai tidak diimbangi dengan sensitivitas yang baik. Dapat disimpulkan bahwa jumlah fitur = 8 memberikan kinerja terbaik dengan akurasi 87,10%, Recall 77,27%, Precision 85% dan F-1 Score 80,95%. Jumlah fitur kurang dari 8 tidak cukup mewakili data yang ada, sementara jumlah fitur yang lebih dari 29 memungkinkan adanya noise yang justru mengurangi kinerja klasifikasi.

4.3 Pengujian dan Analisis Penggunaan PCA

Setelah menemukan jumlah fitur terbaik menggunakan PCA, selanjutnya hasilnya dibandingkan dengan klasifikasi tanpa menggunakan esktraksi fitur. Hasilnya disajikan dalam Tabel 4 di bawah.

Tabel 4. Performansi dengan dan tanpa PCA

Ekstraksi Fitur	Jumlah Fitur	TP	FP	TN	FN	Akurasi	Sensitivitas	Precision	F1-Score
Tidak Ada	2000	16	20	20	6	58,06%	72,73%	44,44%	16,67%
PCA	8	17	3	37	5	87,10%	77,27%	85,00%	80,95%

Dari tabel tersebut terlihat bahwa ketika tidak menggunakan PCA, digunakan 2000 fitur dan menghasilkan akurasi sebesar 58.06%, sensitivitas 72,73% ,Precision 44,44%. dan F-1 Score 16,67% Sementara itu, ketika PCA digunakan, hanya 8 fitur yang diperlukan dan nilai akurasinya naik menjadi 87,10%, Recall 77,27%, Precision 85% dan F-1 Score 80,95%. Kinerja yang diperoleh juga melampaui beberapa penelitian sebelumnya untuk data yang sama. Nilai akurasi untuk data usus besar lebih tinggi 12,91% dari penelitian [6] yang menggunakan SVM yang dikenal memiliki performansi tinggi, 1,23% lebih tinggi dari penelitian [4] dan 7,10% terhadap penelitian [9]. Lebih jauh, penggunaan LDA yang disertai PCA menghasilkan sensitivitas 4,57% lebih tinggi dibandingkan penelitian [7]. Hal ini menunjukkan bahwa PCA mampu menangani kekurangan metode LDA sehingga LDA mampu memberikan hasil klasifikasi yang lebih baik.

5. KESIMPULAN

Berdasarkan penelitian yang dilakukan, diperoleh bahwa kombinasi metode PCA dan LDA mampu mengklasifikasikan data microarray. PCA mampu mengurangi jumlah fitur yang diperlukan dari 2000 fitur menjadi 61 fitur. Kinerja terbaik dihasilkan dengan penggunaan PCA dengan jumlah fitur 8 yang mampu meningkatkan f-1 score klasifikasi data microarray sebesar 64,28% dan akurasi sebesar 29,04% dibandingkan saat tanpa menggunakan PCA. Lebih jauh, hasil penelitian yang diperoleh mampu memperbaiki kinerja beberapa penelitian sebelumnya. Hal ini menunjukkan bahwa PCA juga mampu menghilangkan noise yang terdapat pada data yang digunakan.

REFERENCES

[1] D. Kementerian Kesehatan Republik Indonesia, “Situasi Penyakit Kanker,” Jakarta Selatan, 2015.
[2] Adiwijaya, “Deteksi Kanker Berdasarkan Klasifikasi Microarray Data,” Media Inform. Budidarma, vol. 2, no. 4, pp. 181–186, 2018.
[3] R. Gonzalo and A. Sánche, “Introduction to Microarrays Technology and Data Analysis,” in Comprehensive Analytical Chemistry, 1st ed., vol. 82, J. Jaumot, C. Bedia, and R. Tauler, Eds. Amsterdam: Elsevier BV, 2018, pp. 37–69.
[4] H. Aydadenta and Adiwijaya, “A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest,” J. Inf. Process. Syst., vol. 14, no. 5, pp. 1167–1175, 2018.
[5] M. Ghosh, S. Begum, R. Sarkar, D. Chakraborty, and U. Maulik, “Recursive Memetic Algorithm for gene selection in microarray data,” Expert Syst. Appl., vol. 116, pp. 172–185, 2019.
[6] C. Devi Arockia Vanitha, D. Devaraj, and M. Venkatesulu, “Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection,” in Procedia Computer Science, 2015, vol. 47, pp. 13–21.
[7] Nurfalah, A. Adiwijaya, and Suryani, A.A., (2016). Cancer Detection Based On Microarray Data Classification Using PCA And Modified Back Propagation. Far East Journal of Electronics and Communications, 16(2), p.269.
[8] U. N. Wisesty, R. S. Warastri, and S. Y. Puspitasari, “Leukemia and colon tumor detection based on microarray data classification using momentum backpropagation and genetic algorithm as a feature selection method,” in Journal of Physics: Conference Series (Vol. 971, No. 1), 2018, pp. 12–18.
[9] M. D. Purbolaksono, K. C. Widiastuti, M. S. Mubarok, and F. A. Ma’ruf, “Implementation of mutual information and bayes theorem for classification microarray data,” in Journal of Physics: Conference Series (Vol. 971, No. 1), 2018.
[10] W. K. Yip, S. B. Amin, and C. Li, “A survey of classification techniques for microarray data analysis,” in Handbook of Statistical Bioinformatics, Springer, 2011, pp. 193–223.
[11] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, D. S. Kusumo, "Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification", Journal of Computer Science 14(11), 2018, pp.1521-1530.
[12] M. I. Khalid, T. Alotaiby, S. A. Aldosari, S. A. Alshebeili, F. S. Y. Al-Hameed, M.H. Almohammed, and T. . Alotaibi, “Epileptic MEG

spikes detection using common spatial patterns and linear discriminant analysis," IEEE Access, vol. 4, 2016.
[13] E. Neto, F. Biessmann, H. Aurlien, H. Nordby, and T. Eichele, "Regularized linear discriminant analysis of EEG features in dementia patients," Front. Aging Neurosci., vol. 8, p. 273, 2016.
[14] D. H. Mazumder and R. Veilumuthu, "An Enhanced Gene Selection Methodology for Effective Microarray Cancer Data Classification," Int. J. Simulation--Systems, Sci. Technol., vol. 19, no. 2, 2018.
[15] Z. Jaadi, "A step by step explanation of Principal Component Analysis," Towards Data Science, 2019. [Online]. Available: https://towardsdatascience.com/a-step-by-step-explanation-of-principal-component-analysis-b836fb9c97e2.
[16] "Colon Tumor." [Online]. Available: http://leo.ugr.es/elvira/DBCRepository/ColonTumor/ColonTumor.html.