

REDUKSI DIMENSI PADA KLASIFIKASI DATA KANKER PROSTAT MENGUNAKAN *PRINCIPAL COMPONENT ANALYSIS* DAN *RANDOM FOREST*

Rino Krisdantoro¹, Rakhmat Kurniawan R², Suhardi³

Universitas Islam Negeri Sumatera Utara, Medan

e-mail: rino.krisdantoro@gmail.com

Abstract: *Cancer is a non-communicable disease characterized by the presence of abnormal tissue that is malignant. The speed of cancer growth cannot be controlled and the cancer can spread to other tissues in the patient's body. There are several types of cancer, one of which is prostate cancer. Prostate cancer usually occurs in parts of the male reproductive system that surrounds the urinary tract and usually occurs in older men, but does not rule out it will occur in younger men. One way to detect prostate cancer is by gene expression using microarray technology. Microarray technology can combine thousands of gene expressions simultaneously in one experiment. However, microarray data has very large dimensions, so a dimension reduction process is needed in microarray data with the aim of eliminating features that are less relevant and to increase the accuracy value in the microarray data classification process later. The dimension reduction process is carried out by selecting features using PCA. After the dimension reduction process is carried out, the next step is to carry out the classification process using RF. In this study, PCA was not able to improve the results of the accuracy of the RF classification as expected because RF without reduction has provided the best accuracy of 95.23%. This is due to the nature of RF, namely the more trees that are built, the higher the resulting accuracy results, while the nature of PCA is to reduce or reduce the number of matrices by reflecting the original data to the new data matrix.*

Keyword: Microarray, Reduction, PCA, RF

Abstrak: Kanker merupakan penyakit tidak menular yang ditandai adanya jaringan abnormal bersifat ganas. Kecepatan dari pertumbuhan kanker tidak dapat dikendalikan dan kanker dapat menyebar ke jaringan lain pada tubuh penderita. Ada beberapa jenis penyakit kanker salah satunya yaitu kanker prostat. Kanker prostat biasanya terjadi pada bagian sistem reproduksi pria yang mengelilingi saluran kemih dan biasanya terjadi pada pria lanjut usia, tetapi tidak menutup kemungkinan akan terjadi pada pria yang berusia lebih muda. Adapun salah satu cara mendeteksi kanker prostat dengan ekspresi gen menggunakan teknologi *microarray*. Teknologi *microarray* dapat memantau ribuan ekspresi gen secara bersamaan dalam satu percobaan. Namun, pada data *microarray* memiliki dimensi yang sangat besar sehingga diperlukannya proses reduksi dimensi pada data *microarray* dengan tujuan untuk menghilangkan fitur yang kurang relevan dan untuk meningkatkan nilai akurasi pada proses klasifikasi data *microarray* nantinya. Proses reduksi dimensi dilakukan dengan menyeleksi fitur menggunakan PCA. Setelah proses reduksi dimensi telah dilakukan, langkah selanjutnya melakukan proses klasifikasi dengan menggunakan RF. Pada penelitian ini, PCA tidak mampu meningkatkan hasil akurasi dari klasifikasi RF seperti yang diharapkan karena RF tanpa reduksi telah memberikan akurasi terbaik sebesar 95,23%. Hal ini dikarenakan sifat dari RF yaitu semakin banyak pohon (*tree*) yang dibangun, maka akan semakin tinggi hasil akurasi yang dihasilkan, sedangkan sifat dari PCA yaitu mereduksi atau mengurangi jumlah matriks dengan mencerminkan data asli ke matriks data baru.

Kata kunci: *Microarray*, Reduksi, PCA, RF

PENDAHULUAN

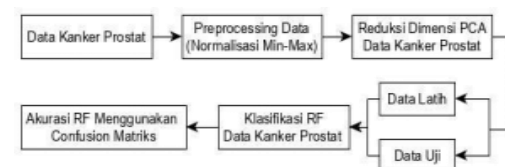
Kanker merupakan penyakit tidak menular yang ditandai dengan adanya jaringan abnormal yang bersifat ganas. Kecepatan dari pertumbuhan kanker tidak dapat dikendalikan dan kanker dapat menyebar ke jaringan lain pada tubuh penderita (Rahayuwati et al., 2020). Pada data Risdeknas tahun 2018 prevalensi kanker di Indonesia terdapat 1,79 per 1.000 penduduk mengidap penyakit kanker. Akibat tingginya prevalensi kanker tersebut maka diperlukan pendeteksian kanker sejak dini (Nita & Indrayani, 2020). Ada beberapa jenis penyakit kanker di Indonesia salah satunya yaitu kanker prostat (Puspita et al., 2025). Kanker prostat biasanya terjadi pada bagian sistem reproduksi pria yang mengelilingi saluran kemih dan biasanya terjadi pada pria lanjut usia, tetapi tidak menutup kemungkinan akan terjadi pada pria yang berusia lebih muda. Adapun salah satu cara mendeteksi penyakit kanker prostat dengan ekspresi gen menggunakan teknologi *microarray*. Teknologi *microarray* dapat memantau ribuan ekspresi gen secara bersamaan dalam satu percobaan.

Proses reduksi dimensi dilakukan dengan menyeleksi fitur menggunakan *Principal Component Analysis* (PCA) (Ardiansyah et al., 2022). PCA merupakan teknik analisis data untuk melakukan pengurangan dimensi atau fitur pada sebuah dataset dengan tetap mempertahankan informasi dari dataset (Arlandy et al., 2025). Setelah proses reduksi dimensi telah dilakukan menggunakan seleksi fitur pada data penyakit kanker prostat, kemudian dilakukan proses klasifikasi dengan menggunakan *Random Forest* (RF). RF merupakan metode klasifikasi yang memiliki banyak pohon (*tree*) sehingga membentuk hutan (*forest*) yang mana setiap pohon keputusan dibangun menggunakan vektor acak. Klasifikasi pada RF dapat diperoleh melalui proses *voting* (jumlah terbanyak) dari pohon-pohon klasifikasi yang terbentuk (Hidayat

et al., 2025). Pohon klasifikasi dibuat dari kumpulan simpul yang saling berkaitan. Setiap simpul mewakili sebuah keputusan, dimana keputusan didasarkan pada salah satu variabel prediktor agar dapat menentukan kelas (Arisandi, 2023).

Sebelumnya, telah dilakukan beberapa penelitian tentang reduksi dimensi dan klasifikasi menggunakan RF pada data penyakit diantaranya yang dilakukan tentang peningkatan kinerja RF melalui seleksi fitur menggunakan PCA untuk mendeteksi penyakit diabetes tahap awal memiliki hasil akurasi 72,55% menggunakan RF tanpa reduksi dengan perbandingan data latih dan uji 80:20%, sedangkan akurasi menggunakan RF dan PCA dengan perbandingan data latih dan uji 80:20% memiliki hasil akurasi terhadap reduksi 1 atribut = 73,20%, terhadap reduksi 2 atribut = 73,86%, terhadap reduksi 3 atribut = 73,20% dan terhadap reduksi 4 atribut = 74,51%. Pada penelitian optimasi RF untuk diagnosis penyakit ginjal kronik menggunakan *Particle Swarm Optimization* (PSO) yang dilakukan oleh (Naufalrifqi, 2022) memiliki hasil akurasi 98% tetapi saat diagnosis penyakit ginjal kronik menggunakan RF tanpa PSO hasil akurasi yang didapat 8,67%. Sehingga dapat dilihat bahwasanya reduksi fitur menggunakan PCA atau optimasi fitur menggunakan PSO dapat mempengaruhi peningkatan hasil akurasi pada klasifikasi RF.

METODE

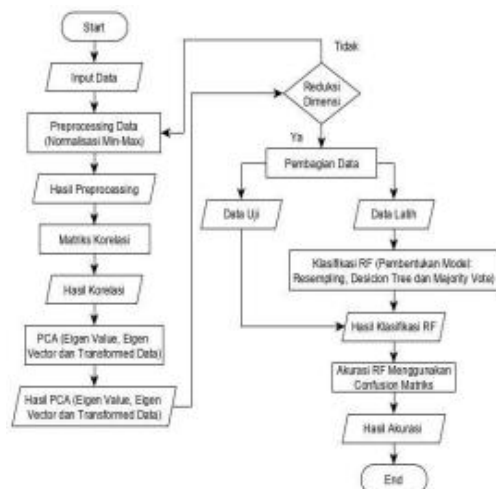


Gambar 1 Diagram Perencanaan Reduksi Dimensi dan Klasifikasi

Adapun proses yang akan dilakukan pertama kali yaitu menyediakan data penyakit kanker prostat, lalu melakukan *preprocessing* dengan

menggunakan metode normalisasi *min-max* agar data menjadi seimbang yaitu antara nilai 0 sampai 1, hal ini dilakukan karena nilai pada data ada yang lebih besar atau lebih kecil. Setelah dilakukan *preprocessing* data, tahap selanjutnya yaitu melakukan reduksi dimensi atau fitur pada data kanker prostat menggunakan *Principal Component Analysis* (PCA), setelah proses reduksi telah dilakukan maka proses selanjutnya yaitu membagi data menjadi data latih dan data uji dengan perbandingan data 80%:20%, setelah dilakukan pembagian data maka tahap selanjutnya yaitu melakukan klasifikasi data kanker prostat apakah terindeks tumor atau normal menggunakan *Random Forest* (RF). Terakhir, setelah dilakukan klasifikasi maka akan dicari kebenaran atau keakuratan dari RF menggunakan metode *confusion matrix*.

Perancangan



Gambar 2 Flowchart Sistem

HASIL DAN PEMBAHASAN

Analisis Data

Beberapa tahapan yang digunakan untuk mengolah data yang diperoleh antara lain sebagai berikut:

1. Menyediakan data penyakit kanker prostat sebanyak 102 data dengan jumlah gen/dimensi data sebanyak 12533 yang terdiri dari kelas positif

tumor (1) dan kelas normal (0).

2. Menerapkan metode PCA pada saat melakukan reduksi dimensi atau fitur data penyakit kanker prostat.
3. Menerapkan metode RF pada saat mengklasifikasi data penyakit kanker prostat pada data yang telah direduksi.

Representasi Data

Adapun data *sample* yang digunakan dapat dilihat pada tabel 1. berikut:

Tabel 1 *Sample* data kanker prostat

Fitur r 1	Fitur r 2	Fitur r 3	Fitur 4	Fitur r 5	Fitur r 6	Class
1	9	0	5	5	-4	0
5	4	0	11	3	-7	0
0	2	-1	2	2	0	1
21	15	1	0	14	-23	1
12	7	4	5	4	-10	1
11	36	8	7	21	-82	1

Hasil Analisis Data

Adapun proses yang dilakukan pertama kali yaitu *preprocessing* data menggunakan normalisasi *min-max*, kemudian reduksi dimensi fitur menggunakan PCA, selanjutnya klasifikasi menggunakan RF dan terakhir menguji keakuratan menggunakan *confusion matrix* yang mana proses tersebut dapat dilihat seperti berikut:

1. Melakukan normalisasi data menggunakan normalisasi min-max

Tabel 2 Hasil *preprocessing* menggunakan normalisasi *min-max*

Fitur 1	Fitur 2	Fitur 3	Fitur 4	Fitur 5	Fitur 6	Class
0,04762	0,20588	0,11111	0,45455	0,1579	0,95122	0
0,2381	0,05882	0,11111	1	0,05263	0,91463	0
0	0	0	0,18182	0	1	1
1	0,38235	0,22222	0	0,63158	0,71951	1
0,57143	0,14706	0,55556	0,45455	0,10526	0,87805	1

0,523 81	1	1	0,636 36	1	0	1
-------------	---	---	-------------	---	---	---

2. Reduksi dimensi atau fitur menggunakan PCA.

Tabel 3 Hasil zero mean pada data sample kanker prostat

Fitur 1	Fitur 2	Fitur 3	Fitur 4	Fitur 5	Fitur 6
$x_1 - \bar{x}_1$	$x_2 - \bar{x}_2$	$x_3 - \bar{x}_3$	$x_4 - \bar{x}_4$	$x_5 - \bar{x}_5$	$x_6 - \bar{x}_6$
-0,349206383	0,093137667	0,222222	3,33E-07	0,166667	0,207318
-0,158730183	0,240195667	0,222222	0,5454547	0,27193	0,170732
-0,396825383	0,299019667	0,333333	0,272727	0,32456	0,256098
0,603174617	0,083333333	0,111111	0,454545	0,307018	0,02439
0,174603217	-0,151960667	0,222223	3,33E-07	0,2193	0,134147
0,126984117	0,700980333	0,666667	0,1818187	0,675439	-0,743903

Mencari nilai matriks kovarian

Tabel 4 Hasil matriks kovarian pada

Tabel 6 Nilai eigen value

	Fitur 1	Fitur 2	Fitur 3	Fitur 4	Fitur 5	Fitur 6
Fitur 1	$\lambda - 0,143$	0,0604	0,0603	-0,0458	0,0925	-0,0573
Fitur 2	0	$\lambda - 0,135$	0,1196	0,008	0,14205	-0,13615
Fitur 3	0	0	$\lambda - 0,143$	0,02828	0,1146	-0,12655
Fitur 4	0	0	0	$\lambda - 0,1223$	-0,0153	-0,02017
Fitur 5	0	0	0	0	$\lambda - 0,16$	-0,14069
Fitur 6	0	0	0	0	0	$\lambda - 0,14$

Menghitung eigen vector

Adapun eigen vector global untuk $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ dan λ_6 adalah sebagai berikut:

$$\begin{pmatrix} - & - & - & 0,98 & 0,98 & -0,99 \\ 0,313 & 0,997 & 0,313 & & & \end{pmatrix}$$

sample data kanker prostat

Fitur 1	Fitur 2	Fitur 3	Fitur 4	Fitur 5	Fitur 6
0,143	0,0604	0,0603	-0,0458	0,0925	-0,057
0,0604	0,1354	0,1196	0,008	0,14205	-0,13615
0,0603	0,1196	0,1432	0,02828	0,1146	-0,12655
-0,0458	0,008	0,02828	0,1223	-0,0153	-0,02017
0,0925	0,14205	0,1146	-0,0153	0,16112	-0,14069
-0,057	-0,13615	-0,12655	-0,02017	-0,14069	0,14193

Mencari nilai eigen value

Tabel 5 Hasil matriks identitas dikali eigen value

λ_1	0	0	0	0	0
0	λ_1	0	0	0	0
0	0	λ_1	0	0	0
0	0	0	λ_1	0	0
0	0	0	0	λ_1	0
0	0	0	0	0	λ_1

Sehingga, didapatkan nilai eigen value dengan membentuk matriks segitiga atas, bawah dan diagonal seperti pada tabel 6. Berikut:

$$\begin{pmatrix} 0,95 & 0,072 & 0,95 & 0,186 & 0,15 & 0,027 \\ M & 0,95 & 0,072 & 0,95 & 0,186 & 0,15 & 0,027 \\ = & 0,95 & 0,072 & 0,95 & 0,186 & 0,15 & 0,027 \\ & 0,95 & 0,072 & 0,95 & 0,186 & 0,15 & 0,027 \end{pmatrix}$$

[0,95 0,072 0,95 0,186 0,15 0,027]

Menghitung *principal component*

Untuk menghitung *Principal Component* (PC) dapat membuat tabel persentase kumulatif dari setiap nilai eigen value dengan cara mengurutkan nilai eigen value dari terbesar ke terkecil, seperti pada perhitungan berikut.

Persentase kumulatif untuk $\lambda_1 = 0,143$

$$= \frac{0,143}{(0,143 + 0,135 + 0,143 + 0,1223 + 0,16 + 0,14)} \times 100\%$$

$$= \frac{0,143}{0,8433} \times 100\% = 0,17 \times 100\% = 17\%$$

Persentase kumulatif untuk $\lambda_2 = 0,135$

$$= \frac{0,135}{(0,143 + 0,135 + 0,143 + 0,1223 + 0,16 + 0,14)} \times 100\%$$

$$= \frac{0,135}{0,8433} \times 100\% = 0,16 \times 100\% = 16\%$$

Persentase kumulatif untuk $\lambda_3 = 0,143$

$$= \frac{0,143}{(0,143 + 0,135 + 0,143 + 0,1223 + 0,16 + 0,14)} \times 100\%$$

$$= \frac{0,143}{0,8433} \times 100\% = 0,17 \times 100\% = 17\%$$

3. Klasifikasi menggunakan *Random Forest*

Menghitung nilai entropy dari seluruh data.

$$Entropy_{Total[4,2]} = \left(-\frac{4}{6} \log_2 \frac{4}{6}\right) + \left(-\frac{2}{6} \log_2 \frac{2}{6}\right)$$

$$Entropy_{Total[4,2]} = 0,39 + 0,582 = 0,91$$

Menghitung entropy dari setiap atribut dengan kriteria fitur ≤ 10 dan fitur > 10 .

$$Entropy_{Fitur \leq 10[2,1][3,0]} = \frac{3}{6} (0,918) + \frac{3}{6} (0) = 0,459 + 0 = 0,459$$

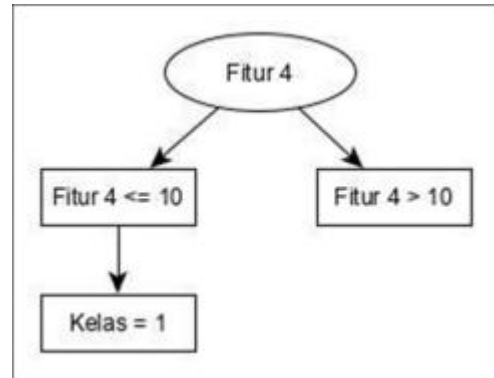
Menghitung nilai *information gain* pada setiap atribut, sehingga diperoleh *information gain* setiap atribut sebagai berikut:

$$Gain_{Fitur1} = 0,918 - 0,459 = 0,459$$

...

$$Gain_{Fitur6} = 0,918 - 0,918 = 0$$

Setelah diketahui nilai *information gain* dari setiap atribut, langkah selanjutnya yaitu memilih nilai *information gain* yang terbesar sebagai *root node*. Adapun analisis yang ditemukan dari Fitur 4 yaitu Fitur 4 ≤ 10 memiliki turunan kelas 1 (*positif tumor*) dan untuk Fitur 4 > 10 tidak memiliki turunan. Sehingga pohon yang terbentuk dapat dilihat pada gambar 3. sebagai berikut:



Gambar 3 Pohon pertama pada RF

Untuk langkah berikutnya, dapat mengulangi proses di atas sebanyak k kali hingga membentuk pohon (*tree*) sebanyak k buah.

Hasil Pengujian

Dalam pengujian reduksi PCA dengan $n_components$ (10, 20, 30, 40, 50, 60, 70, 80, 90 dan 100) dan klasifikasi RF dengan $n_estimators = 100$ hasil akurasi yang didapat dengan menggunakan *confusion matrix* pada persamaan 2.11 seperti pada tabel 7. berikut:

Tabel 7 Hasil Akurasi menggunakan *preprocessing data*

$n_estimators$	RF Tanpa Reduksi	$n_component$	RF + Reduksi PCA
100	95,23%	10	85,71%
		20	85,71%
		30	80,95%
		40	85,71%
		50	85,71%
		60	85,71%
		70	85,71%
		80	85,71%

	90	85,71%
	100	85,71%

Berdasarkan hasil pada tabel 7., akurasi yang didapat lebih tinggi pada klasifikasi RF tanpa reduksi sebesar 95,23%, sedangkan akurasi terendah ada pada klasifikasi RF dengan reduksi PCA $n_estimators = 30$ sebesar 80,95%.

Namun, setelah dilakukan pengujian PCA dengan $n_components$ (10, 20, 30, 40, 50, 60, 70, 80, 90 dan 100) dan klasifikasi RF dengan $n_estimators = 100$ tanpa melakukan preprocessing data atau normalisasi data terlebih dahulu hasil akurasi yang ditemukan seperti pada tabel 8. berikut:

Tabel 8 Hasil Akurasi Tanpa Preprocessing Data

$n_estimators$	RF Tanpa Reduksi	$n_component$	RF + Reduksi PCA
100	95,23%	10	71,42%
		20	85,71%
		30	76,19%
		40	80,95%
		50	85,71%
		60	85,71%
		70	85,71%
		80	80,95%
		90	85,71%
		100	95,25%

SIMPULAN

Berdasarkan pengujian variasi jumlah komponen utama ($n_components$) mulai dari 10 hingga 100, ditemukan bahwa penambahan jumlah komponen tidak memberikan peningkatan akurasi. Nilai akurasi cenderung stagnan di angka 85,71%. Hal ini menyimpulkan bahwa informasi varians terbesar yang relevan untuk klasifikasi kanker prostat pada dataset ini telah terwakili sepenuhnya pada 10 komponen pertama. Pada penelitian ini, PCA tidak mampu meningkatkan hasil akurasi dari

klasifikasi RF karena RF tanpa reduksi mampu memberikan akurasi terbaik sebesar 95,23%. Hal ini dikarenakan sifat dari RF yaitu semakin banyak pohon (*tree*) yang dibangun, maka akan semakin tinggi hasil akurasi yang dihasilkan, sedangkan sifat dari PCA yaitu mereduksi atau mengurangi jumlah matriks dengan mencerminkan data asli ke matriks data baru. Penerapan PCA berhasil mereduksi dimensi data secara ekstrem, dari 12.532 fitur menjadi hanya 10 komponen utama. Meskipun terjadi penurunan akurasi menjadi 85,71% (turun sekitar 9,5%), metode ini menawarkan efisiensi komputasi yang signifikan karena model hanya perlu memproses data yang jauh lebih ringkas.

DAFTAR PUSTAKA

- Ardiansyah, F., Hamdan, F., Sugiyanto, S., & Wahyu Siadi, I. (2022). Klasifikasi Customer Relationship Management Menggunakan Dataset KDD Cup 2009 dengan Teknik Reduksi Dimensi. *Komputika : Jurnal Sistem Komputer*, 11(2), 193–202.
<https://doi.org/10.34010/komputika.v11i2.6498>
- Arisandi, R. (2023). Perbandingan Model Klasifikasi Random Forest Dengan Resampling Dan Tanpa Resampling Pada Pasien Penderita Gagal Jantung. *Jurnal Gaussian*, 12(1), 136–145.
<https://doi.org/10.14710/j.gauss.12.1.136-145>
- Arlandy, K. S., Faqih, A., & Rinaldi, A. R. (2025). Mengoptimalkan Kinerja Naïve Bayes Pada Ancaman Modern Dengan Menggunakan PCA Pada Data Intrusion Detection System (IDS). 8(1), 2621–4962.
- Fitri Adista, N., Apriyanti, I., Kesehatan, P., & Banten, A. (2024). Analisis Dampak Konseling SADARI terhadap Pengetahuan dan Kompetensi WUS dalam Deteksi Dini Kanker Payudara Analysis of

- the Impact of BSE Counseling on WCA Knowledge and Competence in Early Detection of Breast Cancer. *Faletehan Health Journal*, 11(2), 143–149. www.journal.lppm-stikesfa.ac.id/ojs/index.php/FHJ
- Hidayat, R., Tri Saputra, H., Husnah, M., Nabila, N., Hidayatullah, M. B., Naufal Nazhmi, M., Azra, J., & Rana, A. (2025). Implementasi Algoritma Random Forest Regression Untuk Memprediksi Penjualan Produksi di Supermarket. *SIMKOM*, 10(1), 101–109. <https://doi.org/10.51717/simkom.v10i1.703>
- Nita, V., & Indrayani, N. (2020). Pendidikan Kesehatan Dalam Upaya Pencegahan Kanker Serviks Pada Wanita Usia Subur. *Dinamisia : Jurnal Pengabdian Kepada Masyarakat*, 4(2), 306–310. <https://doi.org/10.31849/dinamisia.v4i2.4175>
- Puspita, I. Y. A., Mulyati, S., Mahanani, D., Budianto, E. T., Apriliani, F., Zulfa, H., Wahidatunnisa, I. R., Hafizah, I., Putri, K. F., & Fitri, R. A. (2025). Edukasi Pemeriksaan Payudara Sendiri Pada Wanita Usia Subur Sebagai Upaya Pencegahan Dini Kanker Payudara. *Mitra Mahajana: Jurnal Pengabdian Masyarakat*, 6(1), 72–79. <https://doi.org/10.37478/mahajana.v6i1.5392>
- Rahayuwati, L., Rizal, I. A., Pahria, T., Lukman, M., & Juniarti, N. (2020). Pendidikan Kesehatan tentang Pencegahan Penyakit Kanker dan Menjaga Kualitas Kesehatan. *Media Karya Kesehatan*, 3(1), 59–69.