Reduksi Fitur Untuk Optimalisasi Klasifikasi Tumor Payudara Berdasarkan Data Citra *FNA*

Made Satria Wibawa[1)], Kadek Dwi Pradnyani Novianti[2)]
STMIK STIKOM BALI
Jalan Raya Puputan No. 86 Renon, Denpasar, Bali. (0361) 244445
[1)]satria.wibawa@stikom-bali.ac.id, [2)]novianti@stikom-bali.ac.id

Abstrak

Klasifikasi tumor jinak dan ganas dapat digunakan sebagai penentu adanya kanker payudara pada pasien. Salah satu modalitas untuk klasifikasi tumor adalah Fine Needle Aspiration (FNA). Beberapa fitur seperti tekstur dan kontur dapat diekstrak dari citra FNA untuk klasifikasi tumor menggunakan Computer Aided Diagnosis (CAD). Namun, tidak semua fitur hasil ekstraksi dapat berkontribusi positif terhadap klasifikasi. Penelitian ini bertujuan untuk meningkatkan performa klasifikasi tumor payudara menggunakan fitur citra FNA. Adapun metode yang diajukan adalah tiga jenis metode reduksi fitur, yaitu PCA, RFE dan RFECV. Data sebelum dan setelah reduksi fitur digunakan untuk klasifikasi tumor dengan klasifier KNN. Hasil klasifikasi menunjukkan akurasi tertinggi dicapai dari penggunaan PCA dengan KNN, yaitu sebesar 0.9736. Tingkat akurasi tersebut mengalami peningkatan jika dibandingkan dengan penggunaan fitur awal. Waktu komputasi menggunakan PCA juga mengalami penurunan, yaitu menjadi 1.231 detik.

Kata kunci: *kanker payudara, CAD, reduksi fitur, KNN, PCA*

1. Pendahuluan

Kanker merupakan salah satu penyebab utama morbiditas dan kematian di seluruh dunia dengan angka mencapai 14 juta lebih kasus. Kanker merupakan penyebab kematian nomor dua di seluruh dunia dan pada tahun 2015 angka kematian karena kanker mencapai 8.8 juta lebih. Secara global, penyebab 1 dari 6 kematian adalah penyakit kanker (1). Beberapa jenis kanker yang paling umum ditemui adalah kanker paru-paru, hati (liver), perut dan payudara. Kanker payudara merupakan penyakit kanker dengan persentase kasus baru tertinggi, yaitu 43.3% dan persentase kematian akibat kanker payudara sebesar 12.9%.

Menurut WHO (World Health Organization), deteksi dini dapat meningkatkan peluang penanganan yang sukses dalam kasus penyakit kanker. Diagnosa kanker payudara dapat dilakukan dengan cara biopsi atau metode screening. Metode screening meliputi pemeriksaan fisik, MRI, dan mamografi. Jika hasil screening inkonklusif, maka biopsi melalui analisa mikroskopis *seperti fine needle aspiration* (FNA) dapat digunakan untuk menentukan hasil diagnosa. Biopsi merupakan cara yang paling definitif dalam diagnosa kanker payudara diantara metode diagnosis lainnya.

Istilah kanker merujuk pada pertumbuhan sel abnormal yang berpotensi menyerang dan menyebar ke bagian tubuh lainnya. Pertumbuhan sel abnormal ini akan menimbulkan jaringan yang disebut dengan tumor. Namun, tidak semua tumor berpotensi untuk menyebabkan kanker. Tumor dapat dibagi menjadi tumor jinak (*benign*) dan tumor ganas (*malignant*). Tumor ganas-lah yang disebut dengan penyakit kanker sedangkan tumor jinak tidak termasuk kanker. Oleh karena itu, penentuan tumor termasuk ganas atau jinak merupakan tahapan utama yang penting dalam diagnosa kanker. Penyebab kanker merupakan gabungan faktor genetik dan 3 faktor eksternal. Tiga faktor eksternal termasuk faktor karsinogen fisik (sinar ultraviolet dan radiasi), faktor karsinogen kimiawi (arsenik, komponen asap rokok) dan faktor karsinogen biologis (infeksi dari virus, bakteri atau parasit)(2).

Untuk membantu dalam diagnosa penyakit kanker khususnya kanker payudara, *Computer-Aided Diagnosis* (CAD) dapat digunakan. CAD telah secara luas digunakan untuk mendeteksi dan diagnosa dalam kanker payudara. Oleh karena itu, peningkatan kinerja dari CAD dalam bentuk akurasi, sensitivitas dan spesifitas telah menjadi salah satu bidang riset yang paling utama dan penting. Kelebihan utama dalam penerapan penggunaan komputer untuk membantu diagnosis medis terletak pada pengolahan data mining yang mampu mengekstrak pola tertentu dari data. Diagnosis dengan bantuan komputer jika

dilakukan dengan penanganan yang tepat memiliki potensi untuk menangkap keahlian interpretasi dari seorang pakar. Sehingga dapat meningkatkan akurasi diagnosis dan tingkat keyakinan seorang pakar (3).

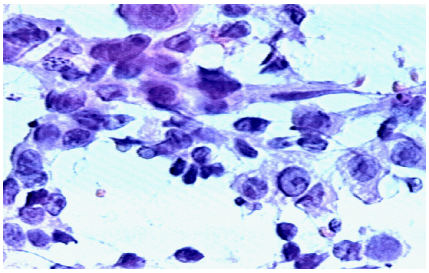
Penelitian untuk mengidentifikasi penyakit kanker payudara menggunakan CAD telah banyak dilakukan. Salah satunya adalah diagnosa kanker payudara menggunakan data biopsi FNAC dari repositori UCI (4). Dataset tersebut memiliki 32 atribut yang merupakan pengukuran dari setiap sel nukleus yang diambil dalam biopsi. Namun, tidak semua atribut yang ada dapat memberikan kontribusi positif terhadap proses pengenalan tumor jinak dan tumor ganas pada dataset. Selain hal tersebut, jumlah atribut yang banyak dapat membebani waktu komputasi proses pembelajaran model klasifikasi. Penelitian ini bertujuan untuk menganalisa performa tiga metode reduksi fitur terhadap optimalisasi hasil klasifikasi

2. Metode Penelitian

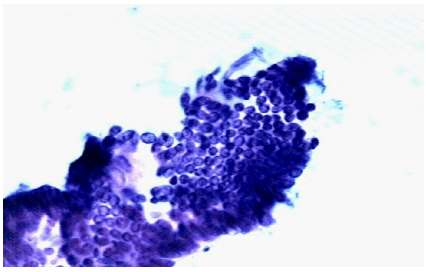
Metode Penelitian memberikan penjelasan tentang langkah-langkah, data, lokasi penelitian, metode evaluasi yang digunakan serta penjelasan terstruktur tentang algoritma atau metode dari penelitian yang dibahas.

2.1. Data

Klasifikasi yang dilakukan adalah klasifikasi biner dengan kelas tumor jinak (*benign*) dan tumor ganas (*malignant*). Jumlah data yang digunakan adalah sebanyak 569 data dengan 357 data kelas *benign* dan 212 kelas *malignant.* Fitur merupakan properti pengukuran terhadap suatu proses/objek yang sedang diamati. Menggunakan kumpulan fitur tersebut metode *machine learning* dapat melakukan klasifikasi. Fitur biasanya disusun oleh ahli di domain dimana *machine learning* diterapkan. Pada penelitian ini, fitur yang digunakan adalah beberapa parameter pengukuran citra jaringan sel tumor yang diambil dari proses *Fine Needle Aspiration* (FNA). Contoh citra tumor payudara ganas dapat dilihat pada Gambar 1a, sedangkan untuk tumor payudara jinak dapat dilihat pada Gambar 1b.



(a) Tumor Ganas



(b) Tumor Jinak

Gambar 1. Contoh Citra Tumor Payudara FNA

Dari citra tumor payudara tersebut diambil 10 fitur kontur dan tekstur, dari 10 fitur utama tersebut diambil lagi 3 ciri statistiknya, yaitu rerata, standar deviasi dan nilai terbesar. Adapun 10 fitur kontur dan tekstur yang diambil adalah sebagai berikut :

a. *Radius*
   Radius merupakan rerata jarak dari tepi sel tumor ke *centroid* sel tumor tersebut.
b. *Perimeter*
   Perimeter merupakan panjang dari keliling sel tumor.
c. *Area*
   Area merupakan luas (jumlah piksel) dari sel tumor.
d. *Compactness*
   Compactness merupakan kombinasi dari fitur *area* dan *perimeter*, yaitu dengan persamaan :

$$\frac{Perimeter^2}{Area}$$

e. *Smoothness*
   Smoothness diukur dari perbedaan antara panjang garis radial dengan rerata panjang garis radial yang mengelilingi garis radial tersebut.
f. *Concavity*
   Concavity merupakan ukuran ‘lekukan’ dari sel tumor.
g. *Concave Points*
   Fitur ini hampir sama dengan *Concavity*, namun hanya mengukur jumlah lekukan yang ada pada sel tumor bukan besarnya lekukan yang ada.
h. *Symmetry*

Untuk mengukur kesimetrisan sel tumor, poros terpanjang yang melalui pusat sel dibentuk terlebih dahulu. Kemudian perbedaan antara panjang garis tegak lurus terhadap poros tadi dihitung.

i. *Fractal Dimension*
Merupakan rasio yang memberikan indeks statistik tentang kompleksitas detail dibandingkan dengan pola yang dibentuk saat penskalaan.

j. *Texture*
Tekstur merupakan perbedaan intensitas keabuan pada piksel di sel tumor.

2.2. Seleksi dan Reduksi Dimensi Fitur

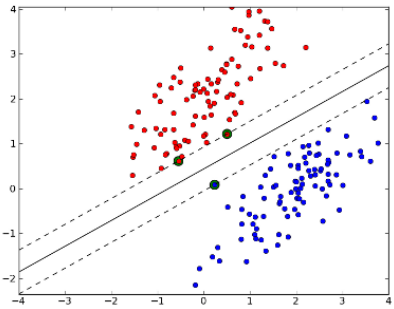
Beberapa tahun belakangan ini jumlah fitur yang diterapkan pada aplikasi *machine learning* dan *pattern recognition* cenderung bertambah. Karena pada dasarnya kita tidak dapat mengetahui bagaimana pengaruh suatu fitur terhadap hasil klasifikasi saat fitur tersebut dibuat. Oleh karena itu, terjadi kecenderungan penambahan jumlah fitur pada proses *machine learning*. Namun, dari sejumlah fitur yang digunakan terdapat kemungkinan fitur yang tidak relevan dan redundan. Beberapa metode telah dikembangkan untuk mengatasi permasalahan tersebut. Metode tersebut dapat dibagi menjadi dua bagian, yaitu metode reduksi fitur dan metode seleksi fitur. Kedua metode tersebut dapat membantu dalam memahami data, mengurangi kebutuhan komputasi dan meningkatkan performa prediksi (5).

a. PCA
Principal Component Analysis (PCA) merupakan salah satu teknik untuk mereduksi dimensi dari suatu data. Tidak seperti metode seleksi fitur yang mengurangi jumlah fitur dengan cara menghilangkan fitur yang dianggap tidak penting tanpa membentuk fitur baru, PCA mengurangi dimensi data dengan cara 'mengkombinasikan' intisari dari atribut dengan membentuk alternatif subset fitur yang lebih kecil. Jadi, pada metode PCA terbentuk fitur yang baru. Jumlah *principal component* dapat berkisar 1-*n*, dimana *n* adalah jumlah fitur. Namun, tentu saja jumlah *principal component* sebaiknya paling banyak sebesar *n-1*, karena tujuan awal adalah untuk mengurangi waktu komputasi (6).
Pemilihan jumlah *principal component* pada penelitian ini dilakukan dengan cara membentuk *principal component* sejumlah 1 hingga *n-1* terlebih dahulu. Kumpulan fitur hasil pca yang terbentuk adalah sejumlah 29 set. Semua 29 set diujikan pada klasifier, set yang menghasilkan klasifikasi terbaik digunakan sebagai pembanding metode reduksi fitur lainnya.

b. *Recursive Feature Elimination* (RFE)
Metode seleksi RFE pada dasarnya adalah proses rekuksif yang meranking fitur berdasarkan tingkat pentingnya terhadap proses prediksi. Pada setiap iterasi, ranking pentingnya fitur diukur dan fitur yang kurang relevan dihilangkan. Ranking tersebut dapat dihitung menggunakan metode *support vector machine* (SVM) kernel linear (7). Untuk klasifikasi biner, SVM membentuk fungsi linear, seperti yang dirumuskan pada Persamaan 1.

$$D(x) = sign(x \cdot w) \quad (1)$$

Dimana x menandakan vektor input dan w adalah vektor yang tegak lurus terhadap *hyperplane* yang terbentuk dari fungsi linear. Algoritma SVM mengalokasikan *hyperplane* dengan jarak terjauh terhadap vektor per kelas. Komponen w merupakan ukuran tingkat pentingnya fitur terhadap fungsi yang terbentuk.

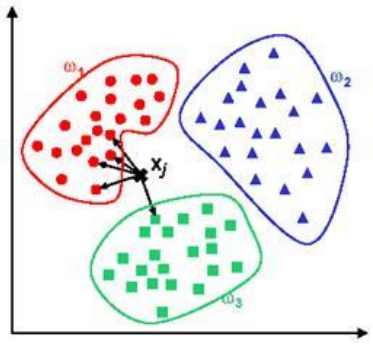


Gambar 1. *Hyperplane* pada *Support Vector Machine* (SVM)

c. *Recursive Feature Elimination Cross Validation* (RFECV)
Pada metode RFE, pemilihan *training set* dalam pembuatan model klasifier sangat berpengaruh terhadap ranking yang dibentuk. Metode RFE menghasilkan ranking dengan variasi yang tinggi atau dengan kata lain metode RFE sangat sensitif terhadap *training set*. RFECV mengatasi permasalahan tersebut dengan penggunaan *cross validation*. *Cross validation* memberi kesempatan pada seluruh dataset untuk menjadi *testing set* sebanyak *k-1*, dimana *k* adalah jumlah partisi pada *cross validation*. Dengan *cross validation*, metode RFE akan lebih stabil dan lebih handal dalam perankingan fitur (7).

2.3. KNN

KNN menggunakan prinsip bahwa data yang memiliki karakteristik yang sama akan memiliki 'kedekatan' dalam ruang vektor. Data baru yang belum diketahui kelasnya dapat diprediksi cara mengobservasi kelas data terdekat. Pada Gambar 2, kelas x_j dapat dicari dari perhitungan jarak ke kelas w_1, w_2 dan w_3.



Gambar 2. Skema Klasifikasi KNN

Beberapa rumus jarak dapat digunakan untuk menghitung kedekatan data. Penelitian ini menggunakan perhitungan jarak Euclidean yang dapat dilihat pada Persamaan 2. Untuk mengetahui jumlah tetangga yang dapat memberikan hasil prediksi terbaik dilakukan uji coba dengan menggunakan jumlah k = 1 hingga 25.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} \quad (2)$$

3. Hasil dan Pembahasan

3.1. Hasil Seleksi dan Reduksi Dimensi Fitur

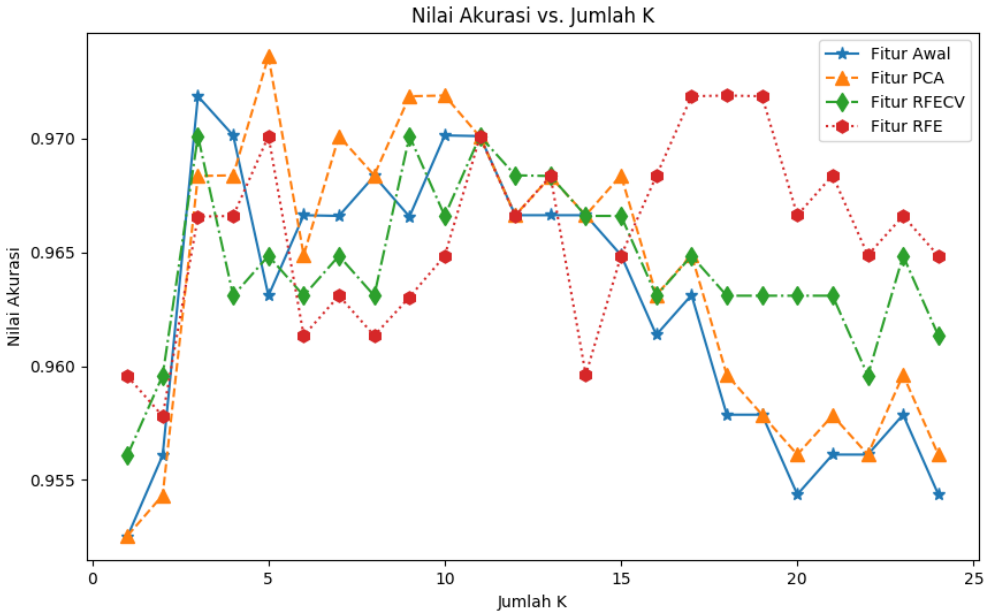
a. PCA
Berdasarkan hasil eksperimen yang dilakukan, jumlah *principal component* yang paling optimal dalam menghasilkan nilai akurasi adalah 10 dimensi. Dengan kata lain, fitur optimal yang dihasilkan oleh PCA adalah 10 buah fitur.

b. RFE
Fitur hasil seleksi fitur metode RFE berjumlah 15 buah. Fitur tersebut yaitu rerata radius, standar deviasi radius, radius terbesar, rerata tekstur, rerata perimeter, standar deviasi perimeter, standar deviasi area, rerata *compactness*, *concavity* terbesar, rerata *concave points*, standar deviasi *concave points*, *concave points* terbesar, rerata simetri, rerata dimensi fraktal dan standar deviasi dimensi fractal.

c. RFECV
Fitur yang dihasilkan dari metode metode RFECV lebih banyak dibandingkan fitur yang dihasilkan metode RFE, yaitu sejumlah 19 buah. Fitur tersebut yaitu rerata radius, standar deviasi radius, radius terbesar, rerata tekstur, rerata perimeter, standar deviasi perimeter, rerata area, standar deviasi area, rerata *smoothness*, standar deviasi *smoothness*, rerata *compactness*, *concavity* terbesar, rerata *concavity points*, standar deviasi *concavity points*, *concavity points* terbesar, rerata *symmetry*, *symmetry* terbesar, rerata dimensi fraktal, standar deviasi dimensi fraktal.

3.2. Performa Klasifikasi

Hasil dari penggunan metode PCA, RFE dan RFECV terhadap akurasi klasifikasi tumor jinak dan ganas dapat dilihat pada Gambar 4, detail hasil klasifikasi berserta parameter yang digunakan ditampilkan pada Tabel 1. Sumbu X menyatakan jumlah *k* atau parameter tetangga yang digunakan pada klasifier KNN. Nilai parameter *k* yang diujikan adalah 1 hingga 25. Sumbu Y merupakan tingkat akurasi yang diperoleh dari metode yang digunakan.



Gambar 4. Perbandingan Akurasi Berdasarkan Fitur dan Nilai *k*

Hasil klasifikasi menggunakan fitur awal, yaitu sebanyak 30 fitur tanpa fitur kelas ditunjukkan dengan garis berwarna biru. Nilai akurasi tertinggi, yaitu sebesar 0.9718 didapatkan saat parameter *k* sebesar 3 tetangga. Pembuatan model klasifikasi menggunakan fitur awal membutuhkan waktu sebesar 2.003 detik. Waktu ini merupakan waktu terlama diantara pembuatan model dengan fitur lainnya.

Tabel 1. Hasil Klasifikasi

No	Fitur	Jumlah Fitur (Non Fitur Kelas)	Nilai *k* Optimal Terkecil	Akurasi Tertinggi	Waktu Komputasi (detik)
1.	Awal	30	**3**	0.9718	2.003
2.	PCA	**10**	5	**0.9736**	**1.231**
3.	RFECV	19	**3**	0.9701	1.717
4.	RFE	15	17	0.9718	1.453

Klasifikasi dengan sepuluh fitur hasil reduksi dimensi PCA ditunjukkan dengan garis berwarna oranye pada Gambar 3. Akurasi tertinggi pada fitur ini sebesar 0.9736, nilai akurasi ini dihasilkan saat parameter *k* sebesar 5. Penggunaan parameter *k* lebih dari 5 cenderung menurunkan tingkat akurasi. Klasifikasi dengan fitur hasil reduksi PCA memakan waktu sebesar 1.231 detik, waktu yang paling sedikit diantara penggunaan metode reduksi fitur lainnya.

Metode RFECV menghasilkan fitur sejumlah 19 atribut. Dengan fitur tersebut, klasifikasi terbaik dihasilkan saat *k* sejumlah 3 buah. Tingkat akurasi tertinggi yang dicapai adalah sebesar 0.9701 dan waktu yang dibutuhkan untuk membuat model adalah 1.717 detik. Tingkat akurasi yang diperoleh melalui fitur metode RFECV adalah akurasi terendah dibandingkan metode lainnya.

Dengan jumlah fitur sebanyak 15 fitur yang dihasilkan oleh metode RFE, tingkat akurasi tertinggi diperoleh saat jumlah *k* sebesar 1 tetangga. Tingkat akurasi tertinggi yang dicapai adalah 0.9718, nilai ini sama dengan tingkat akurasi yang dihasilkan dari penggunaan fitur awal. Namun, waktu

komputasi menggunakan fitur RFE lebih cepat dibandingkan menggunakan fitur awal, yaitu sebesar 1.453 detik.

Berdasarkan hasil diatas, penggunaan metode reduksi fitur mampu meningkatkan performa klasifikasi tumor payudara. Peningkatan performa dapat dilihat dari waktu komputasi dan tingkat akurasi. Waktu komputasi menurun di semua skema reduksi fitur, waktu komputasi terkecil dicapai oleh metode reduksi fitur PCA dengan waktu sebesar 1.231 detik. Penurunan waktu komputasi disebabkan penurunan jumlah fitur dibandingkan jumlah fitur awal, sehingga klasifikasi oleh KNN dapat dilakukan dengan lebih cepat. Tingkat akurasi hanya meningkat pada penggunaan metode PCA, kedua metode reduksi fitur lainnya tidak dapat meningkatkan tingkat akurasi bahkan pada metode RFECV malah menurunkan tingkat akurasi.

4. Simpulan

Penelitian ini dirancang untuk mendiagnosa kanker payudara menggunakan citra digital *fine needle aspirate* (FNA) dari jaringan payudara. 30 fitur berupa tekstur dan kontur citra diekstraksi dari citra tersebut. Dataset yang digunakan adalah dataset sekunder yang diperoleh repositori UCI. Fokus penelitian adalah meningkatkan performa klasifikasi menggunakan klasifier KNN. Metode yang diusulkan untuk meningkatkan performa adalah penggunaan reduksi fitur, yaitu *principal component analysis* (PCA), *recursive feature elimination* (RFE) dan *recursive feature elimination cross validation* (RFECV). Untuk menguji penggunaan reduksi fitur, hasil klasifikasi dengan reduksi fitur dibandingkan dengan hasil klasifikasi dengan fitur awal. Parameter uji yang digunakan adalah tingkat akurasi dan waktu komputasi.

Hasil yang diperoleh menunjukkan penggunaan metode reduksi fitur dapat meningkatkan hasil klasifikasi, yaitu tingkat akurasi dan menurunkan waktu komputasi. Namun, tidak semua metode reduksi fitur dapat meningkatkan performa hasil klasifikasi. Metode RFECV malah menurunkan tingkat akurasi dan tingkat akurasi pada metode RFE tidak mengalami perubahan, namun waktu komputasi berhasil diturunkan di semua metode reduksi fitur yang diusulkan. Tingkat akurasi tertinggi diperoleh dari metode PCA dengan jumlah fitur sebanyak 10 fitur. Waktu komputasi terendah juga diperoleh penggunana metode PCA, yaitu 1.231 detik. Penelitian berikutnya akan berfokus pada penggunaan metode klasifikasi dan analisa klasifier untuk meningkatkan hasil klasifikasi lebih jauh lagi.

Daftar Pustaka

1. World Health Organization. Breast Cancer: Prevention and Control. World Health Organization; 2017.
2. Ganpisetti R, Chandluri P, Lakshmi BVS, Swami PA. World Journal Of Pharmaceutical And Medical Research. Am Cancer Soc. 2016;
3. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph. 2007;31(4):198–211.
4. Lichman M. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences; 2013.
5. Wibawa MS, Nugroho HA, Setiawan NA. Performance evaluation of combined feature selection and classification methods in diagnosing parkinson disease based on voice feature. In: 2015 International Conference on Science in Information Technology (ICSITech). 2015. hal. 126–31.
6. Bro R, Smilde AK. Principal component analysis. Anal Methods. 2014;6(9):2812.
7. Zhang F, Kaufman HL, Deng Y, Drabier R. Recursive SVM biomarker selection for early detection of breast cancer in peripheral blood. BMC Med Genomics. 23 Januari 2013;6(Suppl 1):S4.