Jurnal Ilmu Komputer & Agri-Informatika
Volume 9 Nomor 1 halaman 101 – 113
e ISSN: 2654-9735, p ISSN: 2089-6026

Tersedia secara *online* di:
http://journal.ipb.ac.id/index.php/jika

Pemodelan Berbasis Jaringan untuk Pengklasifikasian Kanker Payudara Berdasarkan Data Molekuler

Network-Based Modeling for Breast Cancer Classification Using Molecular Data

MUSHTHOFA[1*], CHAMDAN L ABDULBAAQIY[1], SONY HARTONO WIJAYA[1], MUHAMMAD ASYHAR AGMALARO[1], LAILAN SAHRINA HASIBUAN[1]

Abstrak

Sel kanker merupakan sel yang memiliki pertumbuhan tidak terkendali. Keberadaan sel kanker di dalam tubuh ditandai dengan adanya estrogen-reseptor-positif (ER+). Salah satu jenis kanker yang banyak diderita saat ini adalah kanker payudara. Sekitar 67% hasil tes kanker payudara menunjukkan adanya ER+ (estrogen-reseptor positif). Selanjutnya, penanganan kanker payudara ditentukan berdasarkan jenisnya, yaitu: Luminal A, Luminal B, basal-like, dan HER-2 enriched. Saat ini, *biomarker* yang umum digunakan untuk mendeteksi keberadaan sel kanker maupun jenis sel kankernya adalah PAM50. Namun, penelitian-penelitian terkait *biomarker* tetap terus dilakukan untuk meningkatkan hasil identifikasi. Penelitian ini menggunakan pendekatan berbasis jaringan (*network*) untuk menentukan *biomarker* potensial berdasarkan data Copy Number Alteration (CNA) dan ekspresi gen. Hasil pemilihan fitur tersebut dibandingkan dengan akurasi berbasis fitur PAM50 dari studi literatur. Dari hasil penelitian didapatkan bahwa fitur dari metode seleksi berbasis jaringan ini mampu menghasilkan performa yang sebanding dengan fitur PAM50 dan dapat menjadi alternatif untuk melakukan klasifikasi jenis kanker payudara.

Kata kunci**:** estrogen-reseptor, kanker payudara, klasifikasi, model *network-based*, sub-tipe kanker payudara

Abstract

Cancer is a disease characterized by uncontrolled cell growth. One of the characteristics of uncontrolled growth is the presence of estrogen-receptor-positive (ER+). About 67% of breast cancer test results have ER+. Breast cancer profiles are divided into 4 subtypes, namely: Luminal A, Luminal B, basal-like, and HER-2 enriched. Each category has a different effect on adjuvant chemotherapy. In this study, a network-based approach was used to select features/molecular biomarkers that have the potential to assist modeling and classifying sub-types of breast cancer. The molecular features used are Copy Number Alteration (CNA) and gene expression. The feature selection results were compared with the PAM50 feature-based accuracy from the literature study. The results indicate that the features selected from this network-based approach can obtain a comparable performance w.r.t. the original PAM50 features, and can be used as an alternative to perform breast cancer subtyping.

*Keywords**: breast cancer, breast cancer subtypes, classification, estrogen-receptor, network-based model***

PENDAHULUAN

Keberadaan sel kanker ditandai dengan pertumbuhan sel yang tidak terkendali dalam suatu organ. Sel tersebut terus melakukan pembelahan sehingga membentuk benjolan atau tumor pada organ yang terjangkit (Islam *et al*. 2020). Salah satu kanker yang banyak terjadi adalah kanker payudara (Bray *et al*. 2018). Pada tahun 2017, National Institute of Health dan organisasi lainnya mendiagnosa lebih dari 250 000 kasus baru kanker payudara di AS. Sedangkan di Indonesia, Global Cancer Observatory mendiagnosa sebanyak 188.231 kasus kanker baru pada wanita di tahun 2018, 30.9% atau sebanyak 58 256 kasus di antaranya adalah

kanker payudara. Di Asia sendiri, kemunculan kanker payudara telah meningkat sebanyak dua atau tiga kali lipat selama beberapa dekade terakhir (Ghoncheh *et al*. 2016). Tingkat mortalitas kanker payudara juga secara umum cukup tinggi. Pada tahun 2018, tingkat mortalitas karena kanker seluruh dunia tercatat sebesar 83.1 per 100 000 orang, dengan proporsi tertinggi (15%) di antaranya disebabkan karena kanker payudara (Goodarzi *et al.* 2020).

Dalam praktik klinis, kanker payudara diklasifikasikan berdasarkan ekspresi reseptor. Kanker disebut sebagai estrogen-reseptor-positif (ER+) jika sel kanker bersifat seperti sel payudara normal yaitu, memiliki reseptor untuk hormon estrogen yang digunakan untuk meningkatkan pertumbuhan sel. Sebaliknya, kanker disebut sebagai ER- jika tidak memiliki reseptor estrogen. Identifikasi ER bermanfaat untuk menentukan kecenderungan sel kanker, apakah akan merespons perawatan hormonal atau kemoterapi. Statistik menunjukkan bahwa sekitar 67% kanker payudara memiliki hasil tes positif untuk reseptor hormon (DePolo 2020).

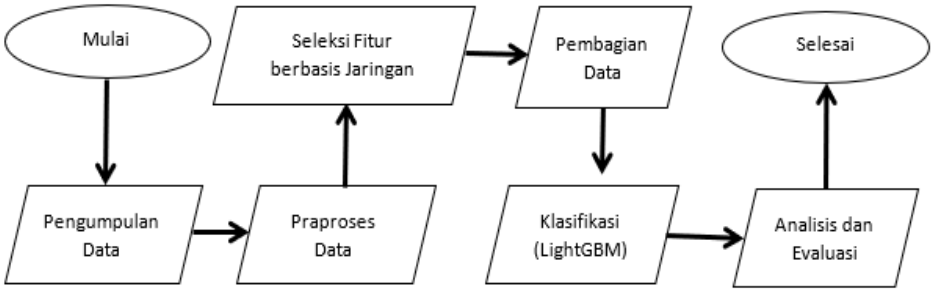
Perou CM *et al*. (2000) menemukan dalam profil ekspresi gen dari kanker payudara bahwa kanker tersebut terbagi menjadi 4 subtipe: luminal A dan luminal B, *basal-like*, dan HER-2 *enriched*. Pengklasifikasian keempat subtipe ini telah terbukti secara prognostik tidak bergantung pada faktor klinikopatologi dan dapat menentukan jenis pasien yang memiliki kemungkinan lebih tinggi mendapat manfaat dari kemoterapi *adjuvant* (Parker *et al*. 2009). Oleh karena itu, berbagai penelitian telah dilakukan untuk memudahkan proses penentuan kelas kanker payudara. Parker *et al*. (2009) mengembangkan *signature* berbasis data ekspresi gen *microarray* 50 buah gen yang terindikasi dapat menjadi prediktor (*biomarker*) bagi klasifikasi kanker payudara. Selanjutnya, berbagai penelitian lain terus dilakukan untuk mengembangkan dan meningkatkan performa dari gen-gen *biomarker* tersebut guna meningkatkan prognosis dari kanker payudara. Pu *et al*. (2020) menggunakan PAM50 untuk membuat model yang memprediksi *survival* dari pasien, sedangkan Raj-Kumar *et al*. (2019) menggunakan pendekatan berbasis Principal Component Analysis (PCA) pada data berbasis PAM50 untuk meningkatkan konsistensi *subtyping* secara klinis pada pasien.

Kanker merupakan penyakit yang kompleks dengan berbagai komponen molekuler yang saling berinteraksi. Oleh karenanya, berbagai pendekatan berbasis jaringan (*network*) untuk memodelkan kanker secara bio-molekuler telah dikembangkan. Vandin *et al*. (2010) mengembangkan pendekatan berbasis *network* pada data mutasi kanker untuk mendeteksi *pathway* biologis yang berperan penting dalam kanker. Verbeke *et al*. (2015) mengembangkan pendekatan berbasis *network* untuk menggabungkan berbagai data molekuler (mutasi, *copy-number alteration*, ekspresi gen dan lain-lain) yang terkait dan kemudian melakukan perangkingan *pathway* yang relevan terhadap setiap subtipe kanker.

Pada penelitian ini dilakukan pemodelan berbasis jaringan untuk mengklasifikasikan estrogen-reseptornya serta untuk mengklasifikasikan profil subtipe dari kanker payudara. Secara lebih spesifik, dalam penelitian ini, metode berbasis *network* digunakan sebagai alternatif pemilihan fitur untuk menentukan subtipe dari kanker payudara, dan untuk mengetahui apakah biomarker berbasis data molekular seperti PAM50 dapat diganti atau ditingkatkan menggunakan penyeleksian fitur berbasis *network*. Selanjutnya, *classifier* yang digunakan adalah *LightGBM* karena *Classifier* ini mampu memproses data dalam ukuran besar dalam waktu lebih singkat dan sumberdaya lebih sederhana (Ke *et al*. 2017).

METODE

Pada penelitian ini, sebuah metode berbasis jaringan (*network-based*), yaitu net*work diffusion* akan digunakan menggabungkan data CNA dan ekspresi gen dan untuk melakukan pemilihan fitur yang dapat digunakan untuk melakukan klasifikasi dan pemodelan pada kanker payudara. Dua permasalahan klasifikasi yang digunakan adalah klasifikasi ER Status dan klasifikasi subtipe pada kanker payudara. Tahapan penelitian yang dilakukan pada penelitian ini tersaji pada Gambar 1.



Gambar 1 Tahapan penelitan.

Pengumpulan Data

Terdapat beberapa jenis data yang digunakan pada penelitian ini yaitu: *Copy Number Alteration* (CNA), *Gene Expression*, Interaksi gen dan PAM50. CNA merupakan data yang menunjukkan perubahan jumlah *copy* gen pada DNA yang normalnya berjumlah dua (Islam *et al*. 2020). Data ini disajikan dalam bentuk matriks berukuran 1565 x 22.546, baris menandakan pasien kanker dan kolomnya adalah perubahan jumlah *copy* dari 22 545 gen, kolom terakhir adalah subtipe kanker payudara. *Gene expression* menunjukkan data keaktifan suatu gen berdasarkan banyaknya jumlah protein yang dihasilkan oleh gen tersebut. Data ini disajikan dalam bentuk matriks berukuran 1565 x 24370, baris menandakan pasien kanker seperti pada data CNA dan kolom menyatakan keaktifan dari 24 370 gen yang disajikan dalam nilai secara kontinyu. Data CNA dan *gene expression* diperoleh dari Molecular Taxonomy of Breast Cancer International Consortium yang tersedia pada alamat https://www.cbioportal.org/study/summary?id=brca_metabric. Gambar 2 dan 3 menunjukkan data CNA dan *gene expression*.

Interaksi gen adalah data yang menunjukkan interaksi antara gen. Data ini diperoleh dari penelitian Mushthofa (2018), terdapat sebanyak 66 000 interaksi gen. Data interaksi disajikan dalam bentuk tabel dua kolom, kolom pertama adalah nama gen dan kolom kedua adalah nama gen yang memiliki interaksi terhadap gen pada kolom pertama. PAM50 adalah data 50 buah gen yang saat ini dijadikan sebagai *biomarker* kanker. Data ini diperoleh dari Forum *BioStar Bioinformatics Explained* dan tersedia pada https://www.biostars.org/p/77590/.

	PatientID	A1BG	A1BG-AS1	A1CF	A2M	A2M-AS1	A2ML1	A2MP1	A3GALT2	A4GALT	...	ZWINT	ZXDA	ZXDB	ZXDC	ZYG11A	ZYG11B	ZYX	ZZEF1	ZZZ3	Subtypes
0	MB-0002	0	0	0	0	0	0	0	0	-1	...	0	0	0	0	0	0	0	-1	0	LumA
1	MB-0005	0	0	0	0	0	0	0	0	0	...	0	1	1	0	0	0	1	0	0	LumB
2	MB-0006	0	0	0	0	0	0	0	0	-1	...	0	1	1	0	0	0	1	1	0	LumB
3	MB-0008	0	0	0	0	0	0	0	0	0	...	0	0	0	1	-1	-1	0	-1	0	LumB
4	MB-0010	1	1	0	-1	-1	-1	-1	0	0	...	0	0	0	0	0	0	0	-1	0	LumB
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1560	MB-7295	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	LumA
1561	MB-7296	0	0	0	1	1	1	1	-1	0	...	0	-1	-1	0	-1	-1	-1	-1	-1	LumB
1562	MB-7297	1	1	0	1	1	1	1	0	-1	...	0	-1	-1	1	0	-1	0	-1	0	LumB
1563	MB-7298	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	-1	0	LumB
1564	MB-7299	-1	-1	-1	-1	-1	-1	-1	-1	0	...	-1	0	0	-1	0	0	0	0	0	LumB

1565 rows × 22546 columns

Gambar 2 Data CNA. Data CNA dikode -1, 0, dan 1 yang mana masing-masing menandakan adanya pengurangan jumlah *copy*, jumlah *copy* tetap dan jumlah *copy* bertambah.

Hugo_Symbol	RERE	RNF165	CD049690	BC033982	PHF7	CIDEA	PAPD4	AI082173	SLC17A3	SDS	...	BX115874	BX107598	UGCGL1	VPS72
MB-0362	8.676978	6.075331	5.453928	4.994525	5.83827	6.397503	7.906217	5.259461	5.702379	6.930741	...	5.271343	5.680321	7.688492	8.084979
MB-0346	9.653589	6.687887	5.454185	5.34601	5.600876	5.246319	8.267256	5.380069	5.521794	6.141689	...	5.942887	5.461069	7.804165	8.349115
MB-0386	9.033589	5.910885	5.501577	5.247467	6.030718	10.111816	7.959291	5.262024	5.689533	6.529312	...	5.174498	5.30403	7.934309	8.406332
MB-0574	8.814855	5.62874	5.471941	5.316523	5.849428	6.116868	9.206376	5.396576	5.43913	6.430102	...	5.116749	5.632249	7.744562	8.310019
MB-0503	9.274265	5.908698	5.531743	5.244094	5.964661	7.828171	8.706646	5.167213	5.417484	6.684893	...	5.402314	5.472185	7.701394	8.137014
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
MB-5465	8.131637	9.101942	5.423027	4.939292	5.644587	5.611189	7.798269	5.219962	5.597732	6.583524	...	5.417529	5.484696	7.643929	8.040024
MB-5453	9.606915	7.427494	5.534115	5.062191	5.927409	5.927031	8.520545	5.129501	5.550549	5.841476	...	5.56632	5.538543	7.048923	7.560101
MB-5471	9.049296	6.85	5.339346	5.166765	6.117095	6.374305	8.499637	4.961279	5.497546	6.351428	...	5.484182	5.386238	7.733413	7.941895
MB-5127	8.858622	6.55045	5.566071	5.140141	5.936371	5.963092	9.320207	5.408996	5.690297	7.280037	...	5.403071	5.436583	7.311774	7.866579
MB-4313	8.415867	6.831722	5.541395	5.266802	7.40896	9.181768	6.804085	5.349958	5.730308	7.642485	...	5.494388	5.381191	9.652446	7.752503

1904 rows × 24368 columns

Gambar 3 Data *gene expression* sebelum dikode menjadi -1, 0, 1 yang mana masing -masing menandakan adanya penurunan ekspresi gen, ekspresi gen tetap, dan ekspresi gen naik.

No	Gene 1	Interaction	Gene 2
1	RUNX3	Interacts with	SMURF2
2	RUNX3	Interacts with	SMURF1
3	RUNX3	Interacts with	SMAD3
…	…	…	….
66978	IL4	Interacts with	IL2RG
66979	MAF	Interacts with	IL4

Gambar 4 Data interaksi gen.

Pra-Proses Data

Pra-proses data perlu dilakukan untuk data *gene expression* dan interaksi gen. Pada data *gene expression*, tingkat keaktifan gen dinyatakan dalam nilai *continue* sehingga perlu diubah ke nilai diskrit -1, 0, 1 yang menyatakan ekspresi gen menurun, ekspresi gen stabil dan ekspresi gen meningkat. Proses konversi diawali dengan menentukan kuartil atas dan bawah pada masing-masing gen (Q3 dan Q1) dan menjadikan nilai tersebut sebagai ambang batas dengan ketentuan:

$$x < Q1 \rightarrow x = -1,$$
$$x > Q3 \rightarrow x = 1,$$
selain itu $x = 0$.

Data *Gene Expression* memiliki jumlah pasien (baris) sebanyak 1 565 untuk kategori kelas subtipe, sedangkan data *Gene Expression* dengan kategori kelas ER memiliki jumlah pasien (baris) sebanyak 1 903 pasien. Kedua kategori kelas tersebut memiliki atribut gen yang sama, yaitu sebanyak ±24 000.

Data Interaksi Gen

Data Interaksi gen ini berupa daftar gen yang berinteraksi dengan gen lainnya. Data tersebut memiliki kurang lebih 66 000 interaksi. Pada penelitian ini, data tersebut diubah bentuknya menjadi sebuah *adjacency matrix*. Contoh data interaksi gen dalam bentuk *adjacency matrix* dapat dilihat pada Gambar 4. Data interaksi gen tersebut didapatkan dari KEGG Human Pathways, Atlas of Cancer Signaling Network (ASCN), dan PPI dari BioGrid. Semua data interaksi gen dan *adjacency matrix* yang digunakan pada penelitian ini dapat diakses dari repositori Github: https://github.com/mushthofa/NBBCFSC.

Seleksi Fitur Berdasarkan *network*

Pada tahapan ini, diterapkan sebuah algoritme *network diffusion* sebagaimana dilakukan pada Mushthofa (2018) yang bertujuan untuk mendapatkan gen-gen yang relevan dijadikan sebagai atribut pengklasifikasian estrogen-reseptor dan subtipe kanker, berdasarkan interaksi pada gen-gen PAM50. Gambar 6 menunjukkan *pseudocode* dari algoritme ini.

Sebagai parameter utama dari algoritme tersebut yaitu nilai α (*alpha*) yang menentukan bobot seberapa besar bobot nilai saat ini dibandingkan nilai awal yang akan didifusikan pada jaringan. Pada penelitian ini, nilai α yang digunakan adalah 0.5 dengan alasan bahwa nilai tersebut dianggap sebagai nilai yang paling netral dan memberikan bobot yang sama antara nilai saat ini dengan nilai awal. Sebagai kriteria konvergensi dari algoritme, digunakan ukuran *Mean Squared Error* (MSE) dengan *threshold* = 10^{-5}. Untuk memenuhi kondisi MSE tersebut diperlukan iterasi sebanyak 3 kali. Gambar 5 menunjukkan ilustrasi difusi nilai fitur pada iterasi 0, 1 dan 2. Setelah mencapai konvergen dan didapatkan vektor akhir, dilakukan *thresholding* kembali pada nilai-nilai gen yang telah didifusikan hingga hanya tersisa kurang dari 100 gen. Dengan kondisi tersebut didapat nilai *threshold* sebesar $2x10^{-3}$ dan gen yang tersisa sebanyak 81 gen.

	14-3-3*	4E-BP*	5T4*	A2M	A4GALT	AACS	AADAT	AANAT	AASDHPPT	AASS	...	p73*	p85*	pre-mRNA~APO*	pre-mRNA~CDC20B*
14-3-3*	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
4E-BP*	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
5T4*	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
A2M	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
A4GALT	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
AACS	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
AADAT	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
AANAT	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
AASDHPPT	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
AASS	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
ABAT	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
ABCB1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0

Gambar 4 Data Interaksi Gen setelah diubah bentuknya menjadi *adjacency matrix*. Matriks tersebut berdimensi 5.354x5.354.

Sebelum memulai proses *network diffusion*, dilakukan normalisasi per kolom pada *adjacency matrix* serta pada nilai awal (F_0). Setelah itu, algoritme tersebut dijalankan hingga tercapai konvergensi. Setelah didapatkan hasil setelah konvergensi, dilakukan pemilihan fitur dengan mengambil gen-gen dengan skor tertinggi dengan melakukan *thresholding* pada nilai skor sehingga terpilih maksimal 100 fitur. Penentuan jumlah 100 fitur ini didasarkan pada pertimbangan bahwa pada literatur, jumlah fitur yang biasa digunakan untuk klasifikasi subtipe kanker biasanya berjumlah < 100, misalnya pada PAM50.

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 & 0.5 & 0 & 0.33 & 0.33 \\ 0.33 & 0 & 0 & 0.33 & 0 \\ 0 & 0 & 0 & 0 & 0.33 \\ 0.33 & 0.5 & 0 & 0 & 0.33 \\ 0.33 & 0 & 1 & 0.33 & 0 \end{bmatrix}$$

Gambar 5 Ilustrasi Normalisasi Matriks berdasarkan kolomnya.

```
SET f0 = NORMALIZE(pam50)
SET a = NORMALIZE(adjacency_matrix)
SET alpha = 0.5
SET ft = f0
WHILE mse >= 10^-5 :
        SET ft1 = (alpha * ft * a) + ([1-alpha] * F0)
        SET mse = COUNT_MSE(ft, ft1)
        SET ft = ft1
```

Gambar 6 *Pseudocode* algoritme *network diffusion.*

Pembagian Data

Birba (2020) menerangkan bahwa membagi dataset menjadi dua kelompok data, yaitu data latih dan data uji penting dilakukan *machine learning*. Data latih digunakan untuk

mendapatkan model, sementara data uji digunakan untuk menguji performa model menggunakan data yang berbeda dalam membangun mode. Pada penelitian ini dilakukan pembagian data 80% dan 20% untuk data latih dan data uji. Pembagian ini dilakukan secara acak menggunakan fungsi cross val_score dari *package* sklearn pada bahasa python.

Klasifikasi

Pada penelitian ini, model klasifikasi yang digunakan adalah *LightGBM* (Ke *et.al.* 2017). *LightGBM* merupakan salah satu algoritme *gradient boosting* yang menggunakan algoritma pembelajaran berbasis *tree*, di mana keunggulannya adalah ia menggunakan *Gradient-Based One Side Sampling* (GOSS) dan *Exclusive Feature Bundling* (EFB) untuk meningkatkan akurasi dengan tetap menjaga kompleksitas algoritme.

Menurut Titov et al. (2021), *LightGBM* ini memiliki beberapa keunggulan, yaitu di antaranya: pelatihan data yang lebih cepat dan efisiensi yang lebih tinggi, pemakaian memori yang lebih rendah, akurasi yang tinggi, mendukung pembelajaran secara paralel maupun terdistribusi dan pembelajaran GPU, terakhir dapat menangani data dengan ukuran yang besar.

HASIL DAN PEMBAHASAN

Praproses Data

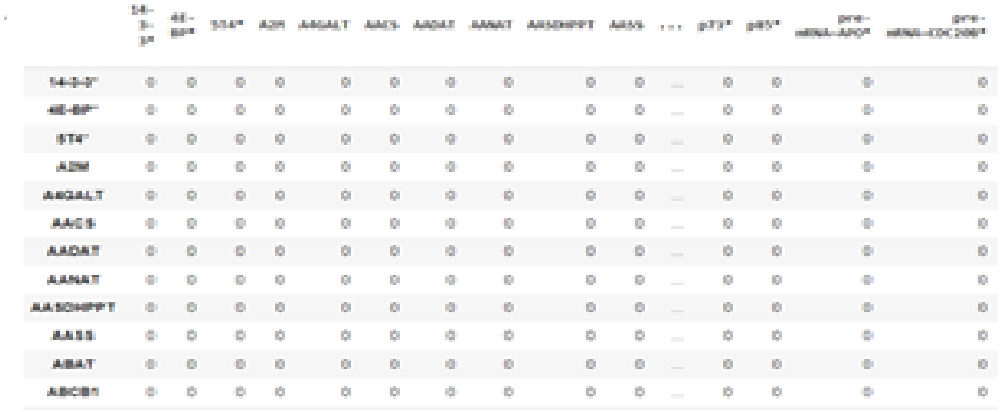
Praproses data untuk mengkodekan perubahan ekspresi gen menghasilkan nilai-nilai -1, 0, atau 1, yang masing-masing menunjukkan apakah terjadi penurunan, tidak terjadi perubahan, atau terjadi kenaikan ekspresi gen pada data pasien yang bersangkutan. Hasil pengkodean data *gene expression* menjadi data diskrit tersaji pada Gambar 7.

- *Row Matching*
  Pada tahap ini dilakukan penyamaan data dari kolom *patient ID*-nya, sehingga kedua data memiliki jumlah baris yang sama. Setelah pasien dari kedua data disamakan kedua data menjadi memiliki 1 565 pasien untuk kategori kelas sub-tipe, dan untuk kategori kelas ER sebanyak 1 903 pasien. Data hasil proses ini dapat diakses di halaman Github: https://github.com/mushthofa/NBBCFSC.

	PatientID	A1BG	A1CF	A2M	A2MAL T2	A4GALT	A4GNT	AA001128	AA013027	AA017152	...	SNpH5	SPPA4	SVPS11	eIA1.Tx	pKBF.2	psITPT622	YPDC	YPDC.II	trans
0	MB-0002	0	-1	-1	1	0	0	1	-1	0	...	0	-1	-1	1	1	-1	0	0	
1	MB-0005	1	0	0	1	0	-1	0	0	0	...	1	1	0	0	1	1	0	0	
2	MB-0006	-1	0	1	-1	-1	1	1	0	-1	...	0	0	1	1	1	-1	1	0	
3	MB-0008	0	-1	0	-1	-1	-1	0	0	0	...	0	-1	0	1	0	0	-1	0	
4	MB-0010	1	-1	0	0	-1	0	-1	0	0	...	1	0	-1	-1	0	-1	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1560	MB-7295	0	1	1	0	0	0	0	0	0	...	1	1	0	0	0	1	1	0	
1561	MB-7296	0	-1	0	1	0	-1	0	0	0	...	1	0	-1	0	0	0	-1	-1	
1562	MB-7297	1	1	0	1	-1	0	0	0	1	...	1	0	0	0	0	0	0	0	
1563	MB-7298	1	-1	0	1	1	0	-1	0	0	...	1	1	0	0	0	0	0	0	
1564	MB-7299	-1	0	0	-1	-1	1	0	-1	-1	...	1	0	0	-1	1	0	-1	0	

1565 rows × 24372 columns

Gambar 7 Hasil pengkodean data *gene expression* menjadi -1, 0 dan 1 yang menyatakan data CNA di kode menjadi -1, 0, 1 yang masing-masing menandakan adanya penurunan ekspresi gen, ekspresi gen tetap, dan ekspresi gen naik.



Gambar 8 Transformasi data interaksi gen menjadi matriks adjacent berukuran 5354 x 5354. Angka 1 pada sel menandakan adanya interaksi antara gen pada baris dan kolom, sementara angka 0 menandakan tidak adanya interaksi.

- Data Interaksi Gen
  Pada praproses ini, data interaksi gen yang semula berupa daftar interaksi antara gen/protein, diubah menjadi bentuk *adjacency matrix*. Setelah diubah, matriks ini berdimensi 5 354x5 354. Dengan kata lain, setidaknya terdapat 5 354 gen yang saling berinteraksi. Gambar 8 menunjukkan hasil matriks hasil transformasi ini.

Pemilihan Fitur Berbasis Jaringan

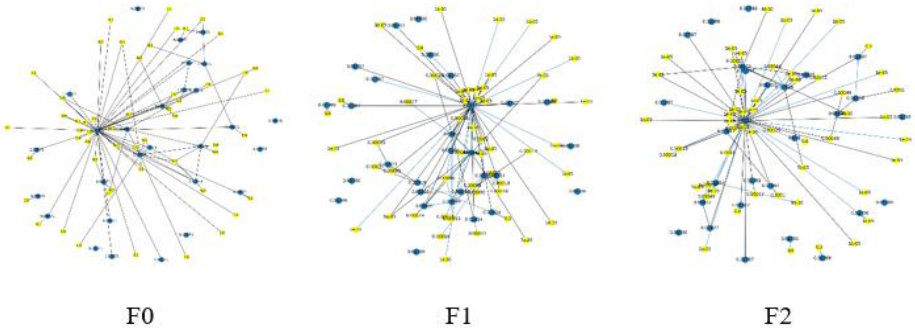
Menggunakan metode *network diffusion* sebagaimana dijelaskan sebelumnya, nilai fitur yang didapatkan dari hasil pra-proses sebelumnya ditransformasi dan dilakukan difusi untuk mendapatkan nilai-nilai yang baru. Pada penelitian ini, nilai α yang digunakan pada proses *network diffusion* adalah 0.5 sedangkan nilai konvergensi ditentukan dari *threshold* $MSE<10^{-5}$. Untuk memenuhi kondisi MSE tersebut ternyata hanya diperlukan iterasi sebanyak 3 kali. Gambar 9 menunjukkan ilustrasi difusi nilai fitur pada iterasi 0, 1 dan 2. Setelah mencapai konvergen dan didapatkan vektor akhir, dilakukan *thresholding* kembali pada nilai-nilai gen yang telah didifusikan hingga hanya tersisa kurang dari 100 gen. Dengan kondisi tersebut didapat nilai *threshold* sebesar $2x10^{-3}$ dan gen yang tersisa sebanyak 81 gen (selanjutnya disebut sebagai himpunan Gen81).

Dari 81 gen tersebut, perlu dilakukan *selection* kembali karena ada beberapa yang berupa protein kompleks (contoh: APC/C*) sehingga perlu diuraikan apa saja gen-nya, ada yang tidak terdapat pada data CNA-nya maupun ekspresi gen,, dan ada juga kondisi lain-lainnya. Setelah melakukan pemilihan kembali, yang semula ada 81 gen berubah menjadi 88 gen yang terpilih (selanjutnya disebut sebagai Gen88). Tabel 1 menunjukkan gen-gen yang terpilih dari proses penguraian protein kompleks berdasarkan Gen81.

Namun, ada juga beberapa gen terpilih dari 81 gen tersebut dibuang. Alasan gen-gen tersebut dibuang adalah diantaranya karena ada gen yang tidak ada pada data CNA maupun *Gene Expression*. Setelah itu, eliminasi gen-gen pada data CNA dan Gene Expression selain gen yang telah terpilih (Gen88), lalu digabungkan antara atribut data CNA dengan atribut data *Gene Expression*.

Pembagian Data

Pembagian data untuk melakukan validasi/*testing* pada penelitian ini adalah 8:2. Data latih sebanyak 80% dan data uji sebanyak 20%. Parameter *random_state* ditetapkan dengan nilai 42.



Gambar 9 Ilustrasi difusi nilai fitur pada iterasi 0, 1 dan 2.

Tabel 1 Daftar gen-gen yang terpilih dari penguraian protein kompleks dari hasil *diffusion network*

No	Gen81	CNA+Expr
1	'APC/C*'	['APC', 'APC2', 'APCDD1', 'APCDD1L', 'APCDD1L-AS1', 'APCS']
2	'cleaved~BCL2*'	['BCL2', 'BCL2A1', 'BCL2L1', 'BCL2L10', 'BCL2L11', 'BCL2L12', 'BCL2L13', 'BCL2L14', 'BCL2L15', 'BCL2L2', 'BCL2L2-PABPN1']
3	'BAD_binding partners*'	'BAD'
4	'HB-EGF*'	'HBEGF'
5	'HER2*'	'ERBB2'
6	PLC_gamma_*	['PDK1', 'AKT1', 'AKT2', 'DAG1']
7	'RRM2*'	['RRM2', 'RRM2B']
8	'ZEB1_ZEB2*'	['ZEB1', 'ZEB2']

Klasifikasi

Untuk melakukan klasifikasi, LightGBM dilatih dengan menggunakan data dengan fitur yang terpilih (Gen88) dan dilakukan 10-*fold cross validation* (CV) untuk melakukan optimasi *hyperparameter*.

- *ER Status Classification*
  Klasifikasi dilakukan dengan menggunakan *LightGBM Classifier*. Data yang digunakan adalah data CNA digabung dengan atribut *Gene Expression*. Ada tiga peng-klasifikasian pada sesi ini, yaitu data dengan hanya terdiri dari gen PAM50, data dengan gen hasil *network diffusion* (Gen88), dan data dengan hanya gen PAM50 yang ada pada Gen88, yaitu sebanyak 28 gen (selanjutnya disebut sebagai Gen28). Hasil klasifikasi ER Status dengan menggunakan masing-masing jenis fitur ditampilkan pada Tabel 2 - 4.

Tabel 2 Hasil klasifikasi ER *Status* menggunakan fitur PAM50

Confusion Matrix, tanpa normalisasi

	Negative	*Positive*
Negative	80	6
Positive	10	285

Confusion Matrix ternormalisasi

	Negative	*Positive*
Negative	0.93	0.07
Positive	0.034	0.97

Tabel 3 Hasil klasifikasi ER *Status* menggunakan fitur Gen88

Confusion Matrix, tanpa normalisasi

	Negative	*Positive*
Negative	82	4
Positive	12	283

Confusion Matrix ternormalisasi

	Negative	*Positive*
Negative	0.95	0.047
Positive	0.041	0.96

Tabel 4 Hasil klasifikasi ER *Status* menggunakan fitur Gen28

Confusion Matrix, tanpa normalisasi |

	Negative	*Positive*
Negative	79	7
Positive	10	285

Confusion Matrix ternormalisasi

	Negative	*Positive*
Negative	0.92	0.081
Positive	0.034	0.97

Dari hasil yang didapatkan tersebut, dapat disimpulkan bahwa secara umum klasifikasi ER *Status* menggunakan fitur terseleksi dari *network diffusion* memiliki akurasi yang cukup tinggi dan tidak berbeda jauh dengan klasifikasi menggunakan fitur dari literatur seperti PAM50. Akurasi untuk Gen88 adalah 97%, sedangkan akurasi untuk PAM50 dan Gen28 adalah 96%. Hasil ini menunjukkan bahwa fitur terpilih dari metode berbasis jaringan ini memberikan hasil yang sebanding dengan fitur dari literatur, yaitu dalam hal ini PAM50, dan bahkan ada sedikit peningkatan dari segi akurasi, meskipun tidak terlalu signifikan. Secara keseluruhan, hasil pengujian dengan 10-fold *cross validation* untuk semua jenis fitur dapat dilihat pada Tabel 5.

Tabel 5 Hasil CV untuk semua jenis fitur

DATA	cv1	cv2	cv3	cv4	cv5	cv6	cv7	cv8	cv9	cv10	**AVG**
PAM50	0.94	0.97	0.95	0.98	0.95	0.99	0.97	0.97	0.96	0.95	**0.96**
Gen88	0.95	0.97	0.97	0.98	0.95	0.99	0.97	0.96	0.97	0.95	**0.97**
Gen28	0.93	0.98	0.95	0.97	0.95	0.98	0.96	0.97	0.97	0.95	**0.96**

- *Subtypes Classification*
  Klasifikasi dilakukan dengan menggunakan *LightGBM Classifier*. Data yang digunakan adalah data CNA digabung dengan atribut *Gene Expression*. Ada 3 peng-klasifikasian pada sesi ini, yaitu data dengan hanya terdiri dari gen PAM50, data dengan gen hasil *network diffusion* (Gen88), dan data dengan hanya gen PAM50 yang ada pada Gen88, yaitu sebanyak 28 gen (Gen28). Hasil klasifikasi subtipe kanker dengan pada masing-masing fitur disajikan pada Tabel 6 - 8.

Tabel 6 Hasil klasifikasi subtipe menggunakan fitur PAM50

Confusion Matrix, tanpa normalisasi | Confusion Matrix ternormalisasi

	Basal	Her2	LumA	LumB	NC
Basal	33	2	2	2	0
Her2	0	36	5	2	0
LumA	0	3	123	14	0
LumB	0	0	7	83	0
NC	0	0	1	0	0

	Basal	Her2	LumA	LumB	NC
Basal	0.85	0.051	0.051	0.051	0
Her2	0	0.84	0.12	0.047	0
LumA	0	0.021	0.88	0.1	0
LumB	0	0	0.078	0.92	0
NC	0	0	1	0	0

Tabel 7 Hasil klasifikasi subtipe kanker dengan fitur Gen88

Confusion Matrix, tanpa normalisasi | Confusion Matrix ternormalisasi

	Basal	Her2	LumA	LumB	NC
Basal	31	4	2	2	0
Her2	0	36	4	3	0
LumA	0	3	124	13	0
LumB	0	3	9	78	0
NC	0	0	0	1	0

	Basal	Her2	LumA	LumB	NC
Basal	0.79	0.1	0.051	0.051	0
Her2	0	0.84	0.093	0.07	0
LumA	0	0.021	0.89	0.093	0
LumB	0	0.033	0.1	0.87	0
NC	0	0	0	1	0

Tabel 8 Hasil klasifikasi subtipe kanker dengan fitur Gen28

Confusion Matrix, tanpa normalisasi | Confusion Matrix ternormalisasi

	Basal	Her2	LumA	LumB	NC
Basal	32	3	2	2	0
Her2	0	36	5	2	0
LumA	0	1	123	16	0
LumB	0	2	11	77	0
NC	0	0	1	0	0

	Basal	Her2	LumA	LumB	NC
Basal	0.82	0.077	0.051	0.051	0
Her2	0	0.84	0.12	0.047	0
LumA	0	0.0071	0.88	0.11	0
LumB	0	0.022	0.12	0.86	0
NC	0	0	1	0	0

Untuk klasifikasi subtipe kanker payudara, didapatkan akurasi untuk PAM50, Gen88 dan Gen28 masing-masing adalah 84%, 83% dan 81%. Hasil dari klasifikasi subtipe kanker ini juga menunjukkan bahwa fitur terpilih dari metode berbasis jaringan ini memberikan hasil yang sebanding dengan fitur dari literatur, yaitu dalam hal ini PAM50, meskipun ada sedikit penurunan, namun tidak signifikan. Secara keseluruhan, hasil *k-fold cross validation* untuk klasifikasi subtipe kanker dengan semua jenis fitur dapat dilihat pada Tabel 9.

Tabel 9 Hasil keseluruhan pengujian CV untuk klasifikasi subtipe kanker payudara

DATA	cv1	cv2	cv3	cv4	cv5	cv6	cv7	cv8	cv9	cv10	**AVG**
PAM50	0.91	0.80	0.86	0.88	0.83	0.84	0.81	0.87	0.81	0.78	**0.84**
Gen88	0.88	0.80	0.82	0.82	0.82	0.86	0.80	0.85	0.82	0.79	**0.83**
Gen28	0.89	0.78	0.84	0.76	0.83	0.82	0.82	0.83	0.80	0.76	**0.81**

Analisis dan Evaluasi

Dalam literatur terkait penelitian kanker payudara, PAM50 telah dikenal sebagai kumpulan 50 gen yang sering digunakan sebagai *biomarker* untuk melakukan klasifikasi/*subtiping* pada kanker payudara. PAM50 biasanya berdasarkan pada data ekspresi gen (*microarray*). Namun, karena adanya keterbatasan data, bisa jadi akses terhadap data *microarray* menjadi sulit untuk didapatkan. Pada penelitian ini, telah diusulkan sebuah metode untuk menggabungkan data *copy number alteration* (CNA) dengan data ekspresi gen, serta

menggunakan teknik berbasis jaringan (*network-based*) untuk melakukan seleksi fitur pada gabungan data CNA dan ekspresi gen untuk membentuk sebuah alternatif seleksi fitur yang berguna untuk klasifikasi kanker payudara. Pendekatan berbasis jaringan dipilih karena sifat kanker yang merupakan penyakit yang dipicu dari kerusakan dan abnormalitas yang muncul secara konsisten pada level *pathway* yang merupakan hasil interaksi dari berbagai komponen molekular, termasuk gen dan protein (Verbeke *et al.* 2015). Hal ini menyebabkan bahwa pendekatan yang lebih tepat untuk memahami dan memodelkan karakteristik kanker adalah pendekatan-pendekatan yang memperhatikan karakteristik jaringan dari entitas-entitas molekuler yang terlibat (gen, protein dan lain-lain).

Untuk menguji pendekatan berbasis jaringan ini, fitur hasil seleksi yang didapatkan kemudian digunakan dalam permasalahan klasifikasi. Sebagai target utama pemodelan klasifikasi, digunakan dua permasalahan klasifikasi pada kanker payudara, yaitu ER *status* dan subtipe dari kanker payudara. Dari penerapan metode ini didapatkan dua buah jenis kelompok fitur yang akan dibandingkan dengan fitur PAM50, yaitu Gen88 (hasil difusi fitur) dan Gen28 (irisan dari Gen88 dan PAM50). Hasil dari masing-masing pengklasifikasian berdasarkan atributnya pada kedua kategori klasifikasi adalah tidak jauh berbeda. Akurasi hasil dari pengklasifikasian kategori subtipe hasilnya adalah 84% untuk PAM50, 83% untuk Gen88, dan 81% untuk Gen28. Namun pada kategori klasifikasi *ER Status*, hasil klasifikasi data yang menggunakan atribut hasil pemodelan *network-based* (Gen88), memiliki nilai akurasi yang lebih tinggi dibandingkan dengan PAM50. Akurasi untuk Gen88 adalah 97%, dan 96% akurasi untuk PAM50 dan Gen28. Dari hasil ini dapat disimpulkan bahwa teknik seleksi fitur berbasis jaringan (*network-based*) yang diusulkan pada penelitian ini dapat digunakan sebagai alternatif dalam melakukan pemilihan fitur untuk berbagai permasalahan pemodelan komputasional dalam penyakit kanker. Namun, di sisi lain, terlihat bahwa meskipun metode ini memilih lebih banyak fitur daripada PAM50 (dan beberapa di antaranya sama dengan metode PAM50), tidak banyak peningkatan akurasi yang didapatkan sehingga beberapa fitur tambahan yang terpilih (pada himpunan Gen88) mungkin tidak begitu relevan untuk kedua permasalahan klasifikasi yang digunakan. Pada penelitian-penelitian selanjutnya, teknik seleksi fitur lebih lanjut dapat diterapkan untuk mengurangi fitur yang terseleksi dengan tanpa mengurangi akurasi dari hasil klasifikasi.

SIMPULAN

Pemodelan *Network-based* pada penelitian ini adalah dengan menggunakan fungsi *Diffusion Network* untuk mencari gen-gen yang juga mungkin relevan untuk dijadikan atribut klasifikasi, dengan gen-gen dari PAM50 sebagai titik awal pencarian fiturnya. Gen-gen yang terpilih dari hasil pemilihan fitur tersebut dijadikan atribut untuk pengklasifikasian, dan hasilnya tidak berbeda jauh dengan hasil klasifikasi data yang menggunakan PAM50. Dengan kata lain dapat dikatakan bahwa pemodelan *network-based* ini dapat digunakan sebagai metode seleksi fitur untuk data molekuler, sekaligus mengintegrasikan data CNA dan ekspresi gen sebagai fitur alternatif untuk *biomarker* klasifikasi subtipe kanker payudara.

Sebagai saran untuk penelitian selanjutnya, perlu dicoba berbagai alternatif data interaksi antar gen yang lain, yang harapannya selain memiliki cakupan data gen yang lebih lengkap dari yang telah digunakan (sehingga tidak perlu menghapus data gen yang tidak ada pada data fitur), juga mungkin memiliki cakupan interaksi antar gen yang lebih lengkap, sehingga mungkin dapat memberikan hasil yang lebih bagus. Selain itu, berbagai algoritma untuk melakukan pemodelan dan difusi pada jaringan, seperti yang dijelaskan pada Fouss (2012) dan Verbeke (2015) dapat menjadi alternatif untuk menghasilkan performa pemilihan fitur yang lebih baik.

DAFTAR PUSTAKA

Birba DE.2020. A Comparative study of data splitting algorithms for machine learning model selection. Stockholm(SE): KTH Royal Institute Of Technology.

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. *ACS Journal*. 68(6):394-424.

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AF, Samarajiwa S, Yuan Y, *et al*. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486(7403): 346–52.

DePolo J. 2020. Breast cancer information and awareness. Estrogen-receptor-negative, progesterone-receptor-positive breast cancers: a distinct subtype?.breastcancer.org [Internet]. [Diakses 2020 Desember 8]; https://www.breastcancer.org/research-news/er-neg-pr-pos-distinct-subtype

Fouss F, Francoisse K, Yen L, Pirotte A, Saerens M. 2012. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural networks* 31:53–72.

Ghoncheh M, Momenimovahed Z, Salehiniya H. 2016. Epidemiology, incidence and mortality of breast cancer in Asia. *Asian Pac J Cancer Prev*; 17: 47-52.

Global cancer observatory (Globocan). 2018. Statistik estimated incidence rates in 2018, indonesia [Internet]. [Diakses pada 2 Desember 2020]; https://gco.iarc.fr/.

Goodarzi E, Beiranvand R, Naemi H, Pordanjani SR, Khazaei Z. 2020. Geographical distribution incidence and mortality of breast cancer and its relationship with the human development index (HDI): An ecology study in 2018. *World Cancer Res J 7*:12.

Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, Ruddy K, Tsang J, Cardoso F. 2019. Breast cancer. *Nature Reviews Disease Primers* 5(1).

Islam MM, Huang S, Ajwad R, Chen C, Wang Y, Hu P. 2020. An integrative deep learning framework for classifying molecular subtypes of breast cancer. *Computational and Structural Biotechnology Journal* 18:2185-2199.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T. 2017. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*. 30.

Pu M, Messer K, Davies SR, Vickery TL, Pittman E, Parker BA, Ellis MJ, Flatt SW, Marinac CR, Nelson SH, et al. 2020. Research-based PAM50 signature and long-term breast cancer survival. *Breast Cancer Res. Treat* 179(1): 197 – 206.

Mushthofa. 2018. Network-based modelling for omics data [Tesis]. Ghent(BE) : Ghent University.

National Institutes of Health; National Cancer Institute. Surveillance, Epidemiology, and End Results Program. Cancer stat facts: female breast cancer [Internet]. [Diakses pada 9 Desember 2020]; https://seer.cancer.gov/statfacts/html/breast.html.

Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 27(8):1160–7.

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. 2000. Molecular portraits of human breast tumours. *Nature* 406(6797):747–52.

Raj-Kumar P, Liu J, Hooke JA, Kovatich AJ, Kvecher L, Shriver CD, Hu H. 2019. PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. *Scientific Reports* 9:7956.

Qi Y, Suhail Y, Lin Y-y, Boeke JD, Bader JS. 2008. Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res* 18:1991–2004.

Suthram S, Beyer A, Karp RM, Eldar Y, Ideker T. eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol Syst Biol* 2008: 4:162.

Titov N, Tsukasa O, Lamb J, Nieva CM.2021.LightGBM Documentation. https://lightgbm.readthedocs.io [Internet] .[Diakses 10 April 2022]; https://lightgbm.readthedocs.io/en/latest/

Vandin F, Upfal E, Raphael BJ. 2010. Algorithms for detecting significantly mutated pathways in cancer. Di dalam: *Proceedings of the 14th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2010)*; Lisbon, 2010 Apr 25-28.

Verbeke LPC, Eynden JVD, Fierro AC, Demester P, Frostier J, Marchal K. 2015. Pathway Relevance Ranking for Tumor Samples through Network-Based Data Integration. *PLoS ONE* 10(7): e0133503.

Wang D, Zhang Y, Zhao Y. 2017. LightGBM: an effective miRNA classification method in breast cancer patients. Di dalam: *ICBB 2017*; Offenburg, 2017 Sep 26-28. hlm 7-11.