Implementasi *Minimum Redundancy Maximum Relevance* sebagai Teknik Reduksi Dimensi pada Klasifikasi Kanker Usus Besar Menggunakan Random Forest

I.G.N.P.Vasu Geramona[1],Adiwijaya[2], Widi Astuti[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]ngurahgera@student.telkomuniversity.ac.id,
[2]adiwijaya@telkomuniversity.ac.id,
[3]widiwdu@telkomuniversity.ac.id.

Abstrak

Kanker merupakan penyakit yang mematikan. Mengutip informasi dari kementrian kesehatan Republik Indonesia pada tahun 2017 sembilan juta orang meninggal akibat kanker. Oleh sebab itu diperlukan sebuah metode untuk mendeteksi kanker salah satunya dengan *gen expression.* Microarray adalah salah satu teknik dari gen expression. Microarray sendiri memiliki feature yang banyak, feature yang banyak ini tidaklah selalu berkaitan dengan masalah yang sedang dihadapi. Sehingga dibutuhkan teknik reduksi dimensi untuk menyeleksi feature yang bersesuaian dengan masalah yang sedang dihadapi.

Pada tugas akhir ini digunakan teknik reduksi dimensi menggunakan *Minimum Redundancy Maximum Relevance* yang selanjutnya akan disingkat dengan MRMR. Adapun Classifier yang digunakan adalah Random Forest, dimana teknik ini membuat beberapa tree untuk mengklasifikasi data lalu dilakukan voting untuk hasil terbanyak. Persamaan MRMR yang digunakan adalah FCD dan FCQ karena data yang digunakan bernilai kontinu. Setelah semua proses telah dilakukan, diperoleh hasil akurasi dari klasifikasi data microarray dengan menggunakan FCQ sebesar 83,87% dan dengan FCD 61,29%.

Kata kunci : microarray, gen expression, random forest, MRMR

Abstract

Cancer is a deadly disease. Quoting information from the Ministry of Health of the Republic of Indonesia in 2017 nine million people died from cancer. Therefore we need a method to detect cancer, one of which is by gene expression. Microarray is a technique of gene expression. Microarray itself has many features, many of these features are not always related to the problem being faced. So we need a dimension reduction technique to select features that correspond to the problem being faced.

In this final project a dimension reduction technique will be used using the Minimum Redundancy Maximum Relevance which will then be abbreviated as MRMR. The Classifier that will be used is Random Forest, where this technique creates several trees to classify data and then will vote for the most results. The MRMR equation used is FCD and FCQ because the data used is continuous. After the process done, the result from classify microarray data using FCQ is 83.87% and with FCD 61.29%

Keywords: microarray, gen expression, random forest, MRMR

1. Pendahuluan

Latar Belakang

Kanker adalah suatu penyakit yang diakibatkan karena perkembangan sel yang tidak normal dari sel biasanya[1]. Kanker sendiri merupakan penyakit yang mematikan. Berdasarkan data dari kementrian kesehatan Republik Indonesia, pada tahun 2017 diprediksi hampir 9 juta orang meninggal dunia akibat kanker dan angka ini diperkirakan akan terus meningkat hingga 13 juta orang pada tahun 2030[2]. Di Indonesia sendiri angka penyakit kanker dapat dikatakan cukup tinggi berdasarkan data Riskesdas pada tahun 2013, prevalensi kanker di Indonesia adalah 1,4 per 100 orang. Hal ini mengakibatkan perlunya deteksi kanker sedini mungkin. Adapun cara untuk mendeteksi kanker itu sendiri diantaranya Rontgen, USG, CT scan, skrining kanker. Namun cara-cara tersebut sangat rentan akan terjadinya kesalahan yang diakibatkan oleh manusia misalnya salah dalam membaca hasil Rontgen maupun CT scan dan lain sebagainya. Dengan demikian diperlukanlah sebuah teknologi yang dapat membantu manusia dalam mendeteksi kanker, salah satunya dengan gen expression. Salah satu dari teknologi gen expression adalah microarray.

Microarray merupakan salah satu teknik untuk menganalisis gen atau DNA. Microarray memungkinkan untuk dilakukannya analisis terhadap ratusan hingga ribuan data gen maupun DNA. Namun permasalahan dari Microarray sendiri adalah banyaknya data dan besarnya dimensi(*Curse of Dimensionality*), hal ini mempersulit untuk didapatkannya sebuah informasi. Sehingga diperlukan sebuah metode untuk mereduksi dimensi dari microarray tersebut. Reduksi dimensi dari microarray sendiri bertujuan untuk menghilangkan variabel atau atribut yang kurang relevan dan juga reduksi dimensi dapat meningkatkan akurasi pada saat klasifikasi nantinya.

Salah satu teknik reduksi dimensi adalah feature selection. Feature selection umumnya digunakan untuk memilih feature yang akan digunakan untuk membangun model nantinya. Reduksi dimensi juga dapat digunakan untuk menghindari *overfitting*. *Overfitting* sendiri adalah sebuah fenomena dimana data train, menghasilkan nilai yang sempurna. Namun hal ini belum tentu bagus untuk data test nantinya.

Teknik reduksi dimensi yang akan digunakan adalah feature selection yaitu *Minimum Redundancy Maximum Relevance (MRMR*). Dipilihnya Minimum Redundancy Maximum Relevance (MRMR) sendiri karena dianggap mampu untuk menghasilkan data yang memiliki redudansi minimum dan relevansi yang maksimum dikarenakan metode ini menghitung data relevansi data ke-n terhadap kelas dan juga redudansi antara data n ke data lainnya. Untuk metode klasifikasi dari data sendiri yang akan digunakan adalah metode random forest. Digunakanya random forest sendiri karena, diharapkan dengan banyaknya decision tree yang dibentuk dari random forest hasil klasifikasi yang didapatkan menjadi lebih baik.

Topik dan Batasannya

Topik dari penelitian ini yaitu bagaimana cara menangani Curse of Dimension dengan menggunakan *feature selection* MRMR, pengaruh *feature selection* terhadap klasifikasi data microarray, dan bagaimana performansi dari sistem yang akan dibangun.

Tujuan

Tujuan dari penelitian ini yaitu mengimplementasikan *feature selection* dan klasifikasi terhadap data microarray dengan menggunakan *Minimum Redundancy Maximum Relevance* (MRMR) sebagai feature selection dan random forest sebagai classifier dan mengetahui performansi dari sistem yang akan dibangun .

Organisasi Tulisan

Bagian pertama adalah pendahuluan. Selanjutnya terdiri dari bagian Studi Terkait, Sistem yang Dibangun, Evaluasi, dan Kesimpulan. Pada bagian Studi Terkait berisikan studi atau riset yang telah dilakukan sebelumnya. Bagian Sistem yang Dibangun menjelaskan bagaimana pembangunan sistem secara umum dengan menggunakan skema dan penjelasan mengenai metode yang akan digunakan dalam membangun sistem. Evaluasi membuat hasil pengujian dan analisis dari sistem. Kesimpulan berisikan kesimpulan dan saran untuk penelitian selanjutnya.

2. Studi Terkait

Banyak metode yang telah dikembangkan untuk mengklasifikasikan data dari microarray diantaranya SVM(Support Vector Machine), Random Forest, KNN, Naïve Bayes, dan masih banyak lagi metode klasifikasi lainnya. Berikut adalah beberapa penelitian yang telah dilakukan sebelumnya pada microarray dapat dilihat pada table 1.

Table 1 Studi Literatur

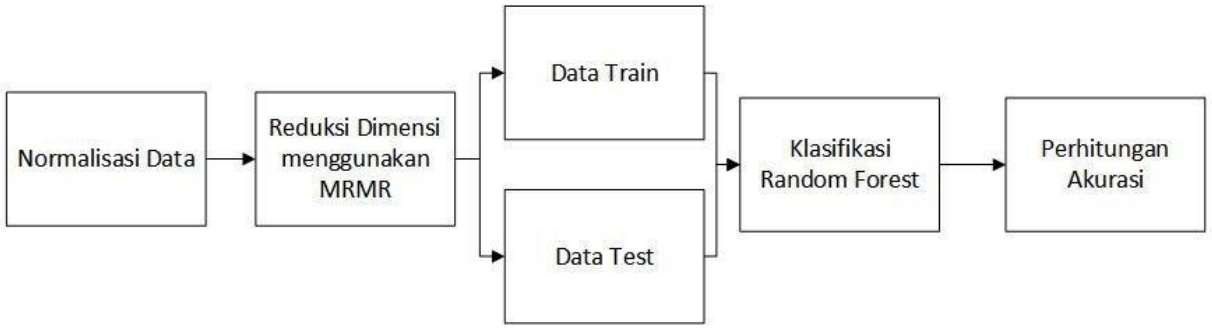
No	Judul Paper	Tahun Terbit	Metode	Hasil
1	Impact of Feature Selection on Support Vector Machine Using Microarray Gene Expression Data	2009	SVM, Feature Selection	Feature selcetion memberikan dampak positife terhadap klasifikasi pada microarray
2	Microarray-based cancer diagnosis: repeated cross-validation-based ensemble feature selection	2018	repeated cross validation, feature selection	repeated cross validation lebih stabil dan memberikan hasil yang sama atau bahkan lebih baik
3	A Clustering Approach for Feature Selection in Microarray Data Classification using Random Forest	2018	Random Forest, K-means	Akurasi yang didapatkan antara 85.87%-98.9%. Lebih tinggi dibandingkan dengan random forest tanpa clustering
4	On the classification techniques in data mining for microarray data classification	2018	SVM, KNN, Naïve Bayes, ANN, C4.5, Random Forest	Hasil akurasi dari random forest lebih tinggi dibandingkan klasifikasi lainnya
5	Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification	2018	PCA, SVM, LMBP	LMBP lebih stabil dibandingkan SVM. LMBP memiliki rata-rata akurasi 96.07% sedangkan SVM 94.98%
6	Analysis of the influence of Minimum Redundancy Maximum Relevance as dimensionality reduction method on cancer classification based on microarray data using Support Vector Machine classifier	2019	MRMR,SVM,PCA	Hasil F1-score kanker usus besar dengan menggunakan MRMR dan SVM lebih baik dibandingkan tanpa reduksi dimensi yaitu 0.84 dan untuk leukimia 0.9657
7	Classification of Electrocardiogram Signals using Principal Component Analysis and Levenberg Marquardt Backpropagation for Detection Ventricular Tachyarrhythmia	2019	PCA, Levenberg Marquardt Backpropagation	Akurasi tertinggi yang didapatkan dengan menggunakan PCA dan Levenberg Marquardt Backpropagation adalah 91,67%

3. Sistem yang Dibangun

Sistem yang dibangun pada tugas akhir ini adalah sebuah sistem yang dapat mengklasifikasikan data kanker berdasarkan data microarray menjadi dua kelas. Data kanker yang akan digunakan adalah data kanker usus besar.

Proses yang akan dilakukan pada sistem ini dibagi menjadi 5 antara lain:

1. Normalisasi data kanker menggunakan metode min-maks yang bertujuan untuk membuat range data menjadi 0-1.
2. Reduksi dimensi menggunakan metode MRMR.
3. Pembagian data hasil MRMR menjadi data train dan test.
4. Klasifikasi data microarray menggunakan Random Forest.
5. Perhitungan akurasi menggunakan teknik confusion matrix.



Gambar 1. Diagram Alur Sistem

A. Normalisasi

Tahap pertama pada sistem yang akan dibangun adalah normalisasi data. Normalisasi data bertujuan untuk menghilangkan redudansi data dan pendependensian dari data. Dengan normalisasi, data akan diubah menjadi sehingga range dari 0-1 menggunakan metode min-maks dengan persamaan 1.

$$Xnormalized = \frac{(X - X\ minimum)}{(Xmaximum - Xminimum)} \quad (1)$$

Dimana Xminimum dan Xmaximum merupakan nilai terendah dan tertinggi dari range nilai yang diinginkan. Sedangkan X merupakan nilai dari dataset sebelum dinormalisasi.

B. Reduksi Dimensi

Tahap selanjutnya adalah reduksi dimensi. Pada tugas akhir ini akan digunakan metode MRMR yaitu FCD yang menggunakan persamaan 4 dan FCQ dengan persamaan 5 karena dataset yang digunakan bertipe continue. Secara garis besar hal yang akan dilakukan pada MRMR adalah mencari nilai F-test dengan menggunakan persamaan (2) lalu nilai korelasi menggunakan pearson correlation. Setelah melakukan reduksi dimensi data akan dibagi 2, menjadi data train dan data test untuk proses klasifikasi nantinya.

Redundancy adalah duplikasi data, pada kasus ini redundancy yang dilihat adalah redundancy antar feature. Redundancy antar feature akan dihitung menggunakan persamaan korelasi pearson (3). Nilai korelasi berkisar antara -1 hingga 1, dimana apabila nilai mendekati -1 atau 1 korelasi variabel sangat kuat namun jika nilai mendekati 0 maka korelasinya lemah. Untuk relevansi akan dihitung menggunakan F-test

$$F_{een} = \frac{MS_{betw}}{MS_{within}} \quad (2)$$

Dimana

F = F-test

$$MS_{between} = \frac{SS_{betw\ een}}{df_{between}}$$

$$MS_{within} = \frac{SS_{w\ ithin}}{df_{within}}$$

$$SS_{within} = \sum (Xi - \overline{Xi})^2$$

Xi = *data kelas i*

$\overline{Xi}$ = *nilai rata − rata kelas i*

$SS_{total} = \sum (Xi - \overline{X})^2$

$\overline{X}$ = *nilai rata − rata seluruh data*

$SS_{between} = SS_{total} - SS_{within}$

$df_{between} = k - 1$

k = jumlah kelas

$df_{within} = N - k$

N = jumlah data

$$r = \frac{n(\sum X_i X_j) - (\sum X_i)(\sum X_j)}{\sqrt{(n\sum Xi^2 - (\sum X_i)^2)(n\sum Xj^2 - (\sum X_j)^2)}} \quad (3)$$

Dimana
n = jumlah data

$X_{i,j}$ = *nilai feature i, j*

$$FCD = max(F - r) \quad (4)$$

$$FCQ = max(F/r) \quad (5)$$

C. Klasifikasi

Metode klasifikasi yang akan digunakan pada tugas akhir ini adalah Random Forest dengan menggunakan bantuan tool sklearn.

D. Perhitungan Akurasi

Tahapan terakhir adalah perhitungan akurasi. Akurasi akan dihitung menggunakan metode Confusion Matrix. Tujuan dari perhitungan performansi sendiri adalah untuk mengukur ketepatan sistem dalam melakukan klasifikasi dengan benar sehingga menghasilkan keluaran yang benar.

Pada sistem yang akan dibangun nantinya apabila sistem memberi label benar dan label asli adalah benar maka akan diberi label *True Positive* (TP). Apabila sistem memberi label salah dan label asli salah maka akan diberi label *True Negative* (TN). Apabila label yang diberikan sistem berbeda dengan label asli maka akan diberikan label False Positive (FP) apabila sistem memberi label benar dan label asli salah atau False Negative (FN) sistem memberi label salah dan label asli benar.

Table 2 Confusion Matrix

		Nilai Sebenarnya	
		Positive	Negative
Nilai Prediksi	Positive	TP	FP
	Negative	FN	TN

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (6)$$

4. Evaluasi

Data set yang digunakan pada tugas akhir ini adalah data kanker usus besar dengan rincian sebagai berikut:

Table 3 Dataset

Dataset	
Gene	2000
Sample	62
Class Name	Sample
Tumor(0)	40
Normal(1)	22

Dari data set tersebut (tabel 3) akan dilakukan normalisasi lalu dilanjutkan dengan seleksi fitur menggunakan MRMR. Pemilihan jumlah fitur dalam MRMR dilakukan dengan cara intuisi dengan trial dan error[5]. Pada tugas ini jumlah feature yang diujikan adalah 10 hingga 100 namun yang akan ditampilkan adalah jumlah feature 10,30,50,70 , dan 90 dikarenakan keterbatasan ruang . Seleksi fitur akan menggunakan metode MRMR adapun metode MRMR yang akan digunakan adalah FCQ dan FCD. Digunakannya kedua metode ini dikarenakan data set adalah data continue. Tahap selanjutnya adalah klasifikasi, klasifikasi yang akan digunakan adalah random forest. Untuk klasifikasi sendiri akan dijalankan sebanyak tiga kali untuk setiap jumlah fiturnya.

4.1 Hasil Pengujian

Dari hasil pengujian MRMR dengan melakukan pengambilan sejumlah feature antara lain 10,30,50,70, dan 90 terhadap dataset kanker usus besar lalu dilakukan klasifikasi menggunakan Random Forest didapatkanlah hasil sebagai berikut:

Tabel 4 Evaluasi Kinerja Klasifikasi

Seleksi Fitur	Jumlah Fitur				
	10	30	50	70	90
FCQ	68.81%	77.41%	77.41%	83.87%	83.87%
FCD	61.29%	61.29%	61.29%	61.29%	61.29%
Tanpa MRMR	82.79%	82.79%	82.79%	82.79%	82.79%

Dari hasil pengujian dapat dilihat bahwa reduksi dimensi menggunakan metode MRMR menghasilkan akurasi yang lebih baik dibandingkan tanpa MRMR, khususnya FCQ dimana klasifikasi dengan FCQ menghasilkan akurasi sebesar 83.87% yang akurasinya sedikit lebih tinggi dari klasifikasi tanpa MRMR yaitu 82,79%.

Table 5 Evaluasi Waktu

Jumlah Feature	MRMR	Tanpa MRMR
10	2037.63s	2316.05s
30	2088.73s	
50	2046.35s	
70	2086.74s	
90	2068.01s	

Sedangkan untuk waktu komputasi klasifikasi dengan MRMR khususnya FCQ membutuhkan waktu yang lebih sedikit yang berkisar antara 2037.63s - 2068.01s untuk jumlah fitur 10-90 dibandingkan dengan tanpa MRMR yang membutuhkan waktu 2316.05s.

4.2 Analisis Hasil Pengujian

Dari hasil pengujian dapat dilihat bahwa MRMR yang menggunakan metode FCQ menghasilkan akurasi yang lebih baik dibandingkan dengan menggunakan metode FCD. Akurasi tertinggi yang didapatkan dari klasifikasi Random Forest menggunakan feature selection MRMR dengan metode FCQ adalah 83,87% sedangkan dengan menggunakan metode FCQ akurasi tertinggi yang didapatkan adalah 61,29%. FCQ sendiri melakukan pembagian dalam persamaannya (5) sedangkan FCD menggunakan pengurangan (4). Untuk masalah relevansi dan redundancy persamaan yang menggunakan pembagian lebih disarankan.

Penggunaan teknik reduksi dimensi MRMR dapat menghasilkan nilai akurasi yang sama dengan klasifikasi tanpa MRMR namun waktu komputasi yang dibutuhkan sedikit lebih lama dibandingkan dengan MRMR. Perbedaan hasil yang tidak terlalu jauh ada kemungkinan diakibatkan oleh penggunaan teknik klasifikasinya karena klasifikasi yang digunakan adalah Random Forest. Pada random forest terdapat proses pencarian information gain dimana ini merupakan proses pencarian informasi yang relevan dengan data sehingga dapat meningkatkan hasil akurasi dari klasifikasi.

5. Kesimpulan

Berdasarkan pengujian yang telah dilakukan pada tugas akhir ini, dapat disimpulkan bahwa klasifikasi dengan menggunakan feature selection khususnya MRMR mempersingkat waktu komputasi. Klasifikasi dengan MRMR membutuhkan waktu komputasi yang berkisar antara 2037.63s - 2068.01s sedangkan klasifikasi tanpa MRMR membutuhkan waktu komputasi selama 2316.05s. Untuk pengujian yang dilakukan terhadap dua metode MRMR dimana klasifikasi dengan metode MRMR FCQ memperoleh hasil akurasi sebesar 83,87% sedangkan klasifikasi dengan metode FCD sebesar 61,29%. Pada kasus ini FCQ memiliki hasil akurasi yang lebih baik dibandingkan dengan FCD.

Adapun saran untuk studi selanjutnya perlu dilakukan perbandingan antara Random Forest dengan teknik klasifikasi lainnya, menguji metode MRMR untuk data yang bernilai diskrit, dan melakukan uji coba terhadap data kanker lainnya.

Daftar Pustaka

[1] Husna Aydadenta and Adiwijaya. 2018. A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest, *Journal of Information Processing Systems*, 14, 5, (2018), 1167~1175.

[2] Adiwijaya, Kang & Wisesty, Untari Novia & Lisnawati, E. & Aditsania, Annisa & Kusumo, Dana. (2018). Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification. Journal of Computer Science. 14. 1521-1530. 10.3844/jcssp.2018.1521.1530.

[3] Aydadenta, H., & Adiwijaya. 2018, March. On the classification techniques in data mining for microarray data classification. In Journal of Physics: Conference Series (Vol. 971, No. 1, p. 012004). IOP Publishing.

[4] Z. Cao, Y. Wang, Y. Sun, W. Du and Y. Liang, "Effective and stable feature selection method based on filter for gene signature identification in paired microarray data," *2013 IEEE International Conference on Bioinformatics and Biomedicine*, Shanghai, 2013, pp. 189-192.

[5] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, Stanford, CA, USA, 2003, pp. 523-528

[6] *Kamus Besar Bahasa Indonesia*. Retrieved from https://kbbi.kemdikbud.go.id/entri/kanker.

[7]*Kementerian Kesehatan Republik Indonesia*. Retrieved from http://www.depkes.go.id/article/print/17020200002/kementerian-kesehatan-ajak-masyarakat-cegah-dan-kendalikan-kanker.html.

[8]*Kent Ridge Biomedical Data Set Repository*,. (n.d.). Retrieved from http://leo.ugr.es/elvira/DBCRepository/index.html.

[9] D. Pavithra and B. Lakshmanan, "Feature selection and classification in gene expression cancer data," *2017 International Conference on Computational Intelligence in Data Science(ICCIDS)*, Chennai, 2017, pp. 1-6.

[10] S. Dasgupta, G. Saha, R. Mondal, R. K. Pal and A. Chanda, "A comparison between methods for generating differentially expressed genes from microarray data for prediction of disease," *Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT)*, Hooghly, 2015, pp. 1-5.

[11] C. M. M. Wahid, A. B. M. S. Ali and K. Tickle, "Impact of Feature Selection on Support Vector Machine Using Microarray Gene Expression Data," *2009 Second International Conference on Machine Vision*, Dubai, 2009, pp. 189-193.

[12]Ma'ruf, Firda Aminy, Adiwijaya, and Untari Novia Wisesty. "Analysis of the influence of Minimum Redundancy Maximum Relevance as dimensionality reduction method on cancer classification based on microarray data using Support Vector Machine classifier." Journal of Physics: Conference Series. Vol. 1192. No. 1. IOP Publishing, 2019.

[13]Manik, A., Adiwijaya, A., & Utama, D. Q. (2019). Classification of Electrocardiogram Signals using Principal Component Analysis and Levenberg Marquardt Backpropagation for Detection Ventricular Tachyarrhythmia. Journal of Data Science and Its Applications, 2(1), 78-87.