Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 4 No. 5 (2020) 805 - 812 ISSN Media Elektronik: 2580-0760

Perbandingan CART dan *Random Forest* untuk Deteksi Kanker berbasis Klasifikasi Data *Microarray*

Riska Chairunisa[1], Adiwijaya[2], Widi Astuti[3]
[1,2,3]Informatika, Fakultas Informatika, Universitas Telkom
[1]riskachairunisa@student.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id, [3]astutiwidi@telkomuniversity.ac.id

Abstract

Cancer is one of the deadliest diseases in the world with a mortality rate of 57,3% in 2018 in Asia. Therefore, early diagnosis is needed to avoid an increase in mortality caused by cancer. As machine learning develops, cancer gene data can be processed using microarrays for early detection of cancer outbreaks. But the problem that microarray has is the number of attributes that are so numerous that it is necessary to do dimensional reduction. To overcome these problems, this study used dimensions reduction Discrete Wavelet Transform (DWT) with Classification and Regression Tree (CART) and Random Forest (RF) as classification method. The purpose of using these two classification methods is to find out which classification method produces the best performance when combined with the DWT dimension reduction. This research use five microarray data, namely Colon Tumors, Breast Cancer, Lung Cancer, Prostate Tumors and Ovarian Cancer from Kent-Ridge Biomedical Dataset. The best accuracy obtained in this study for breast cancer data were 76,92% with CART-DWT, Colon Tumors 90,1% with RF-DWT, lung cancer 100% with RF-DWT, prostate tumors 95,49% with RF-DWT, and ovarian cancer 100% with RF-DWT. From these results it can be concluded that RF-DWT is better than CART-DWT.

Keywords: *cancer, microarray, discrete wavelet transform, classification and regression tree, random forest.*

Abstrak

Kanker merupakan salah satu penyakit yang mematikan di dunia dengan tingkat kematian 57,3% pada tahun 2018 di benua Asia. Maka dari itu, diperlukannya diagnosis dini untuk menghindari peningkatan angka kematian yang disebabkan oleh penyakit kanker. Seiring berkembangnya pembelajaran mesin, data gen kanker dapat diolah menggunakan *microarray* untuk deteksi terjangkitnya penyakit kanker sejak dini. Namun permasalahan yang dimiliki *microarray* adalah jumlah atribut yang sangat banyak sehingga perlu dilakukan reduksi dimensi. Untuk mengatasi permasalahan tersebut, dalam makalah ini digunakan reduksi dimensi *Discrete Wavelet Transform* (DWT). Selanjutnya digunakan *Classification and Regression Tree* (CART) dan *Random Forest* (RF) sebagai metode klasifikasinya. Tujuan penggunaan kedua metode klasifikasi tersebut untuk mengetahui metode klasifikasi mana yang menghasilkan performa paling baik. Pada penelitian ini digunakan lima data *microarray* yaitu *Colon Tumor, Breast Cancer, Lung Cancer, Prostate Tumor* dan *Ovarian Cancer* dari *Kent-Ridge Biomedical Dataset*. Akurasi terbaik yang didapat pada penelitian ini untuk data *breast cancer* sebesar 76,92% dengan CART-DWT, *Colon Tumor* sebesar 90,1% dengan RF-DWT, *lung cancer* sebesar 100% dengan RF-DWT, *prostate tumor* sebesar 95,49% dengan RF-DWT, dan *ovarian cancer* sebesar 100% dengan RF-DWT. Dari hasil tersebut maka dapat disimpulkan bahwa RF-DWT lebih baik dibandingkan CART-DWT.

Kata kunci: kanker, *microarray, discrete wavelet transform, classification and regression Tree, random forest.*

1. Pendahuluan

Kanker merupakan salah satu penyakit yang mematikan di dunia dengan peningkatan jumlah kasus yang tinggi di setiap tahunnya. Berdasarkan data dari *International Agency for Research on Cancer* (IARC) pada tahun 2018 terdapat 18,1 juta kasus untuk penyakit kanker di seluruh dunia dan diperkirakan pada tahun 2040, kasus kanker di dunia meningkat sebanyak 11,4 juta kasus [1]. Selain itu, pada tahun 2018 terdapat 9,6 juta kasus kematian yang disebabkan oleh penyakit kanker. Untuk tingkat kematian yang disebabkan peyakit kanker di benua Asia sebesar 57,3% pada tahun 2018 [2]. Maka dari itu, diperlukannya diagnosis dini untuk mengurangi angka kematian yang disebabkan oleh penyakit kanker. Namun pendeteksian secara konvensional menggunakan media gambar seperti CT *scan* dan PET *scan* masih membutuhkan waktu yang cukup lama. Banyak faktor yang membuat pendeteksian menggunakan media

Diterima Redaksi : 24-06-2020 | Selesai Revisi : 25-08-2020 | Diterbitkan Online : 30-10-2020

gambar kurang akurat seperti kesalahan pada pengambilan gambar, kelainan pada organ yang tidak terdeteksi atau diagnosis yang salah [3]. Seiring perkembangan pembelajaran mesin, terdapat suatu teknologi DNA *microarray* yang dapat digunakan untuk mendeteksi terjangkitnya penyakit kanker. DNA *microarray* dapat mendeteksi penyakit kanker dengan waktu yang singkat menggunakan data gen dan juga dapat mengurangi faktor yang memungkinkan pendeteksian kurang akurat [4].

DNA *microarray* adalah suatu teknologi yang digunakan dalam mengumpulkan dan memproses ribuan ekspresi gen dalam waktu yang sama [4][5]. *Microarray* merupakan metode yang efisien dalam menentukan pola ekspresi ribuan gen. Data ekspresi gen yang akan diolah berupa sebuah matriks dimana baris merepresentasikan gen dan kolom merepresentasikan sampel [3]. Pemrosesan ribuan ekspresi gen membuat DNA *microarray* memiliki kendala yaitu jumlah atribut yang dimiliki oleh data *microarray* sangat besar dan setiap atribut memiliki hubungan yang rumit. Kendala tersebut menyebabkan hasil pendeteksian kurang akurat dan waktu komputasi yang tinggi saat melakukan pendeteksian. Reduksi dimensi dapat menjadi solusi untuk kendala tersebut. Penggunaan reduksi dimensi dapat membantu klasifikasi dalam meningkatkan keakuratan hasil klasifikasi. Hasil yang baik bisa didapatkan jika penggabungan reduksi dimensi dan metode klasifikasi cocok.

Hingga saat ini, sudah banyak penelitian *microarray* dengan berbagai metode klasifikasi dan reduksi dimensi yang telah dilakukan. Husna Aydadenta dan Adiwijaya telah melakukan penelitian [6] yang bertujuan untuk menganalisa metode klasifikasi *Random Forest* dalam mengklasifikasikan prediksi kanker dengan data *microarray*. Penelitian ini menggunakan kombinasi metode *K-means* dan *Relief Method* untuk mereduksi dimensi DNA *microarray*. Penelitian ini mendapatkan hasil akurasi 98,9% dan 89% untuk data *Lung* dan *Prostate.*

CHEN, Lei dan Yi-Hui LIU pada penelitian [7] menganalisa penggunaan reduksi dimensi *t-test* dengan metode klasifikasi *Classification and Regression Tree* (CART) untuk klasifikasi DNA *microarray.* Penelitian ini mendapatkan akurasi lebih dari 96% untuk kanker paru-paru.

Pada tahun 2018 telah dilakukan penelitian dengan judul "*Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification*" oleh Adiwijaya, dkk. Penelitian ini membandingkan dua metode klasifikasi yaitu *Support Vector Machine* (SVM) dan *Levenberg Marquardt based Back Propagation* (LMBP) untuk mengetahui metode yang lebih baik dalam mengklasifikasikan DNA *microarray.* Masing-masing metode dipasangkan dengan reduksi dimensi *Principal Component Analysis.* Dalam penelitian ini, metode LMBP menghasilkan akurasi yang lebih baik dibandingkan SVM dengan nilai akurasi LMBP sebesar 92,86% dan SVM sebesar 85,71% untuk salah satu dataset yang digunakan yaitu kanker paru-paru [4].

Penelitian [8] menggunakan reduksi dimensi *Principal Component* (PC) dan metode klasifikasi *Multinominal Logit* dalam mengklasifikasikan data *microarray*. Penelitian ini dilakukan oleh Azka Khoirunnisa, Adiwijaya dan Aniq A. Rohmawati pada tahun 2019 dengan tujuan untuk menganalisa penggunaan reduksi dimensi *Principal Component* (PC) dan metode klasifikasi *Multinominal Logit* untuk mengklasifikasikan data *microarray*. Penelitian ini menghasilkan akurasi sebesar 76,1% untuk dataset *Colon* dan 76% untuk dataset *Leukimia*.

Penelitian yang dilakukan Khadijah dan Sri Hartati pada tahun 2013 [9] menggunakan reduksi dimensi *Discrete Wavelet Transform* dan metode klasifikasi *Extreme Learning Machine* yang diterapkan pada *Radial Basis Function Network*. Namun, hasil akurasi yang didapat masih kurang optimum yaitu sebesar 75%.

Pada penelitian [3], Jason Bennet, dkk relah melakukan percobaan untuk menggabungkan reduksi dimensi DWT dengan beberapa metode klasifikasi yaitu KNN, SVM, *Naïve Bayes*, dan *Hybrid*. Penelitian ini mendapatkan akurasi 100% untuk dataset *breast cancer* menggunakan metode klasifikasi *hybrid* dan jenis dwt db7.

Penelitian ini membangun sebuah sistem klasifiasi *microarray* untuk mendeteksi penyakit kanker. Pada penelitian ini digunakan lima data microarray yaitu *Colon Tumor, Breast Cancer, Lung Cancer, Prostate Tumor,* dan *Ovarian Cancer* yang berasal dari *Kent-Ridge Biomedical Dataset*. Kumpulan data *microarray* dari *Kent-Ridge* berasal dari ekspresi gen yang telah divalidasi dan memenuhi standar, karena banyak digunakan sebagai acuan untuk kinerja ekspresi gen pada bidang bioinformatika.

Metode reduksi dimensi yang digunakan pada penelitian ini yaitu *Discrete Wavelet Transform* (DWT). Metode klasifikasi yang digunakan yaitu *Classification and Regression Tree* (CART) dan *Random Forest*. Penggabungan reduksi dimensi DWT dengan kedua metode klasifikasi tersebut dilakukan untuk mengetahui apakah DWT mampu mengoptimalkan performansi dari CART dan *random forest.* Tujuan dari penelitian ini yaitu mengetahui penerapan CART-DWT dan *random forest*-DWT dalam mengklasifikasikan data microarray serta menganalisis perbandingan performansi kedua metode klasifikasi yang digunakan. Hasil dari kedua metode klasifikasi tersebut dibandingkan untuk mengetahui metode klasifikasi mana yang menghasilkan performansi paling baik jika dipasangkan dengan reduksi dimensi DWT untuk mengklasifikasikan data *microarray* dalam mendeteksi kanker.

2. Metode Penelitian

2.1. Data

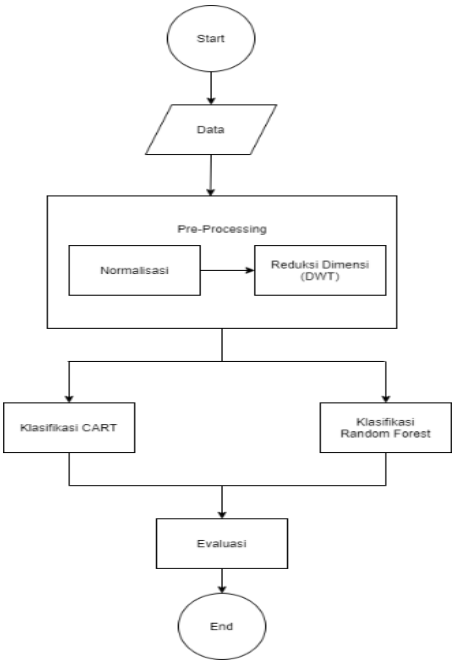
Data yang digunakan dalam penelitian ini adalah data *microarray* yang berasal dari *Kent-Ridge Biomedical Dataset* yang dapat diambil di http://leo.ugr.es/elviraDBCRepository/. Data *microarray* yang digunakan untuk penelitian ini antara lain data *Breast Cancer, Colon Tumor, Lung Cancer, Prostate Tumor,* dan *Ovarian Cancer*. Spesifikasi data yang digunakan dapat dilihat pada tabel 1.

Tabel 1. Data

Data	Jumlah Kelas	Jumlah Fitur	Jumlah Sample
Breast Cancer	2	24482	97
Colon Tumor	2	2000	62
Lung Cancer	2	12533	181
Prostate Tumor	2	12600	136
Ovarian Cancer	2	15155	53

2.2. Skema Umum

Pada penelitian ini, metode klasifikasi yang digunakan untuk mengklasifikasikan *microarray* adalah *Classification and Regression Tree* (CART) dan *Random Forest*. Kedua metode tersebut dipasangkan dengan reduksi dimensi *Discrete Wavelet Transform* (DWT). Pada penelitian ini dianalisa perbandingan hasil klasifikasi untuk kedua metode tersebut. *Flowchart* sistem dapat dilihat pada gambar 1.



Gambar 1. *Flowchart* Sistem

2.3. *Pre-Processing*

Terdapat banyak masalah yang timbul dari pengolahan data antara lain terlalu banyak atribut, nilai data berada di *range* yang sangat jauh, dan *missing value*. Masalah tersebut menyebabkan hasil dari pengolahan data tersebut kurang baik sehingga dibutuhkannya *pre-processing* pada data. Teknik *pre-processing* merupakan bagian dari eksplorasi data dimana data akan dipahami dan diubah sehingga siap untuk di proses [10]. Teknik *pre-processing* akan memperbaiki masalah yang ada di dalam data dan mengubah data menjadi terstruktur [11].

2.3.1. Normalisasi

DNA *microarray* memiliki variasi data yang besar sehingga tidak menutup kemungkinan terdapat atribut *noise* dan atribut yang memiliki hubungan yang lemah terhadap label. Normalisasi adalah salah satu teknik *pre-processing* dalam penskalaan atau pemetaan. Tahap ini merubah skala pada data menjadi skala yang lebih kecil. Skala data yang baru dapat membantu klasifikasi karena dapat menghapus fitur dengan *noise* tinggi dan relevansi yang rendah [8].

Terdapat banyak teknik normalisasi seperti *Min-Max, Z-score* dan *Decimal Scalling*. Penelitian ini menggunakan teknik *Min-Max Normalization* karena teknik ini dapat menghasilkan *range* antar data sebelum dan setelah dilakukan normalisasi pada interval yang sama [12]. Untuk mendapatkan nilai normalisasi setiap data dapat menggunakan rumus 1.

$$A' = \frac{A - nilai\ min.A}{nilai\ max.A - nilai\ min.A} \quad (1)$$

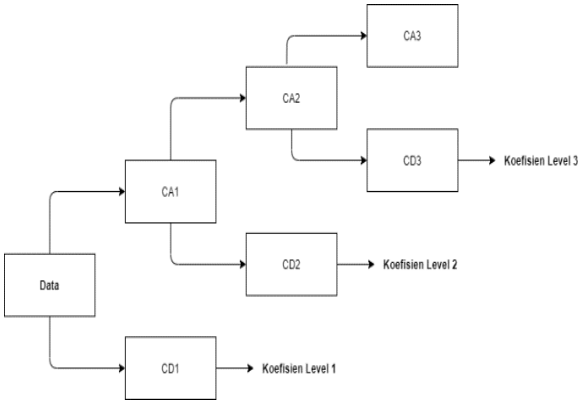
Dengan A' merupakan nilai data baru dari *Min-Max Normalization* dan A merupakan nilai data sebelum di normalisasi.

2.3.2. Reduksi Dimensi *Discrete Wavelet Transform*

Metode ekstraksi fitur DWT merupakan salah satu jenis transformasi teknik *wavelet*. *Discrete Wavelet Transform* merupakan metode yang memproses sinyal dengan cara mendekomposisi sinyal menjadi beberapa frekuensi. DWT melakukan kompresi secara *dydic* atau bilangan integer pangkat dua pada sinyal [13]. Jenis *wavelet* yang digunakan pada penelitian ini adalah *daubechies*. *Daubechies* cenderung stabil dalam melakukan reduksi dimensi untuk data dengan dimensi yang tinggi [14].

Reduksi dimensi DWT melakukan dekomposisi pada sinyal menjadi dua bagian. Dekomposisi tersebut dilakukan menggunakan *highpass decomposition filter* (H) dan *lowpass decomposition filter* (L). Penelitian [10] mengutip penjelasan penelitian [15] bahwa Filter H mengkonvolusi sinyal lalu dilanjutkan dengan *downsampling* dan menghasilkan kelompok sinyal berfrekuensi tinggi atau disebut koefisien detail (cD1). Filter L juga memproses sinyal dengan alur yang sama. Hasil dari konvolusi dan *downsampling* dari filter L

yaitu kelompok sinyal berfrekuensi rendah atau disebut koefisien aproksimasi. Alur dekomposisi pada DWT dapat dilihat pada gambar 2.



Gambar 2. Alur Dekomposisi DWT Level 3

Langkah selanjutnya yaitu dilakukannya *inverse* kepada filter dekomposisi DWT dan akan dikalikan dengan koefisien hasil dekomposisi untuk mendapatkan *detail value* untuk setiap fitur. Sistem ranking diterapkan untuk menyeleksi dan mengambil fitur-fitur yang penting pada data [3]. Penelitian ini menggunakan *mother wavelet daubechies* db2, db4, db7, db8, dan db10. Dekomposisi level pada DWT dilakukan dari level 1 hingga level 6.

2.4. Klasifikasi Menggunakan *Classification* and *Reression Tree*

Algoritma CART merupakan algoritma yang sedehana namun memiliki performansi yang baik. Algoritma CART dapat melakukan proses klasifikasi dengan waktu komputasi yang singkat [16]. Pada tahap *training*, CART akan membuat sebuah model *tree* menggunakan aturan *split*. Semua variabel latih akan menjadi simpul dalam model *tree*. Sistem penentuan simpul pada *tree* dapat menggunakan nilai *Entropy* dan *Information Gain* yaitu dengan mencari nilai *Information Gain* untuk setiap atribut kemudian atribut yang memiliki nilai *Information gain* terbesar makan akan menjadi simpul dalam *tree*. Untuk menghitung nilai Entropy dapat menggunakan rumus 2 dan rumus 3 untuk menghitung *Information Gain*.

$$Entropy\ (S) = -\sum_i^c p_i log_2 p_i \quad (2)$$

Dimana c merupakan label kelas pada data dan p_i merupakan jumlah data yang memiliki kelas i dan c adalah banyaknya kelas.

$$IG(S,a) = Entropy(S) - \sum_v \frac{|S_v|}{|S|} Entropy(S_v) \quad (3)$$

Dimana IG merupakan nilai *information gain*, *v* merupakan semua nilai yang mencakup himpunan data atribut a. S_v merupakan jumlah data yang dimiliki atribut a [17].

Penandaan label kelas deteksi kanker pada suatu simpul dapat menggunakan aturan jumlah terbanyak dimana kelas dari data adalah kelas yang memiliki nilai $P(t_0|i)$ terbesar, nilai $P(t_0|i)$ didapat menggunakan rumus 4.

$$P(t_0|i) = max_t P(t|i) = max_t \frac{N_t(i)}{N(i)} \quad (4)$$

Dengan $P(t|j)$ adalah frekuensi terjadinya kelas t dalam deteksi kanker pada simpul i, $N_t(i)$ adalah jumlah atribut yang memiliki kelas t pada simpul i, $N(i)$ adalah jumlah record pada simpul i [18].

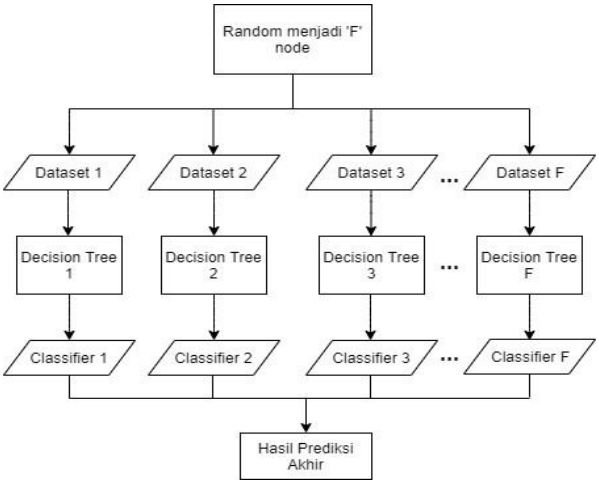
Terdapat dua skenario pengujian untuk mengatahui performa CART-DWT dalam mengklasifikasikan DNA *microarray*. Skenario pertama yaitu melakukan klasifikasi DNA *microarray* tanpa melakukan reduksi dimensi pada data. Nilai dari parameter yang digunakan pada skenario ini diubah-ubah nilainya hingga mendapatkan hasil akurasi yang paling baik. Penginisiasi parameter didasari oleh penelitian sebelumnya.

Skenario kedua yaitu melakukan klasifikasi DNA *microarray* dengan melakukan reduksi dimensi DWT pada data. Dilakukan beberapa *cross validation* dan diambil nilai K yang memiliki akurasi rata-rata tertinggi. Pada skenario ini digunakan parameter yang menghasilkan akurasi yang baik pada skenario pertama.

2.5. Klasifikasai Menggunakan *Random Forest*

Random Forest adalah kumpulan dari pohon klasifikasi hasil dari *sampling bootstrap* data. Langkah awal dalam membangun model *random forest* yaitu menentukan nilai N sebagai jumlah *decision tree* yang dibangun. Pemilihan data yang digunakan untuk pembangunan *tree* menggunakan teknik *bootstrap sample*. Teknik ini akan memilih sampel dari data secara acak dan dilakukan secara berulang hingga jumlah sampel pada *bootstrap sample* sama dengan jumlah data sebenarnya. Tujuan pemilihan ini bukan untuk mencari kemungkinan terbaik dari *split* semua data, namun hanya mencari kemungkinan terbaik dari *split* dari gen-gen yang terpilih [19]. Model yang dibangun menggunakan aturan *entropy* untuk mendapatkan nilai homogenitas pada *decision tree*. Teknik *voting* digunakan dalam penentuan label kelas.

Skenario pengujian untuk pasangan *random forest*-DWT memiliki skenario yang sama dengan skenario CART-DWT. Dilakukan dua kali pengujian dimana pengujian pertama tanpa menggunakan reduksi dimensi DWT dan pengujian kedua menggunakan reduksi dimensi DWT. Parameter yang digunakan pada *random forest* hampir sama dengan parameter yang digunakan CART namun terdapat tambahan parameter N *tree* sebagai banyaknya *tree* yang dibuat pada model *random forest*. Berikut *flowchart* metode *Random Forest* pada gambar 3.



Gambar 3. *Flowchart Random Forest*

2.6. Evaluasi Sistem

Pengukuran performa suatu klasifikasi dapat menggunakan berbagai cara seperti akurasi, *recall*, dan *precision*. Dalam penelitian ini, pengukuran performa dari klasifikasi *microarray* dihitung menggunakan akurasi. Akurasi merupakan pengukuran performa umum yang sering digunakan. Akurasi dapat memberikan gambaran yang baik mengenai seberapa tepat sistem melakukan klasifikasi. Hasil prediksi dari model klasifikasi dibandingkan dengan label sesungguhnya [20]. Jumlah prediksi yang benar dibagi dengan jumlah keseluruhan data lalu dikali dengan 100%. Untuk mendapatkan nilai akurasi dapat menggunakan rumus 5.

$$Akurasi = \frac{jumlah\ prediksi\ benar}{Jumlah\ data} \times 100\% \quad (5)$$

Perhitungan akurasi di terapkan pada kedua metode klasifikasi yang digunakan. Kemudian akurasi dari kedua metode tersebut dibandingkan dan mendapatkan metode klasifikasi mana yang memiliki performa paling baik jika digabungkan dengan reduksi dimensi DWT.

3. Hasil dan Pembahasan

Pada penelitian ini digunakan 5 dataset seperti yang tertera pada tabel 1. Atribut-atribut yang terdapat pada data melalui tahap reduksi dimensi menggunakan *Discrete Wavelet Transform* untuk mengurangi jumlah atribut pada data. Setelah dilakukannya reduksi dimensi, maka dilakukan implementasi model klasifikasi. Teknik *cross-validation* digunakan untuk membagi data *testing* dan *training*. *Cross validation* akan membagi data menjadi K partisi dimana salah satu partisi menjadi data *test* dan partisi lainnya menjadi data *train*. Tahap ini dilakukan sampai semua partisi pernah menjadi data *test*. Penelitian ini menggunakan *cross validation* dengan K = [1,2,3,4,5,6,7,8,9,10]. Pemilihan nilai parameter dan k-fold yang digunakan pada penelitian ini melalui studi empiris.

3.1. Performa CART-DWT dalam Klasifikasi *Microarray*

Parameter dan *k-fold* yang digunakan dapat dilihat pada tabel 2 dan nilai terbaik untuk setiap parameter dapat dilihat pada tabel 3.

Tabel 2. Parameter

Parameter	Nilai
K-fold	1,2,3,4,5,6,7,8,9,10
Maximum Split	1,2,3,4,5,6,7,8,9,10
Maximum Features	'*None', 'sqrt', 'log2'*
Filter DWT	Db2, db4, db7, db8, db10
Level DWT	1,2,3,4,5,6

Tabel 3. Parameter Terbaik

Dataset	*K-fold*	*Max Depth*	Filter DWT	Level DWT
Breast	7	2	Db2	3
Colon	7	9	Db10	1
Lung	3	10	Db7	5
Prostate	2	7	Db10	1
Ovarian	5	4	Db4	5

Parameter CART yang digunakan pada penelitian ini adalah *Max Depth* dan *Max Feature*. *Max Depth* merupakan jumlah level yang dibentuk pada *tree*. Pembentukan level pada *tree* tergantung dengan banyaknya atribut yang dipilih yang dapat berpengaruh terhadap kesalahan prediksi. Pembatasan level ini dibutuhkan untuk mengendalikan kompleksitas data sehingga perhitungan komputasi semakin cepat dan hasil prediksi semakin optimal [21]. Parameter *Max Features* merupakan jumlah fitur yang diambil sebagai pertimbangan ketika mencari simpul pada *tree* sehingga tidak semua data digunakan dalam aturan *split*. Penelitian [22] mengemukakan bahwa semakin kecil nilai *max feature* maka semakin kecil juga hasil pada model. Hasil akurasi yang didapatkan dengan menggunakan parameter terbaik dapat dilihat pada tabel 4.

Tabel 4. Hasil Akurasi CART

Dataset	Akurasi CART-DWT(%)			Akurasi CART No DWT(%)		
	None	*Sqrt*	*Log2*	*None*	*sqrt*	*Log2*
Breast	50,47	**76,92**	66,24	54,86	61,85	54,94
Colon	73,80	**90,07**	83,53	76,98	64,08	76,38
Lung	**98,26**	93,20	89,83	95,55	97,22	89,31
Prostate	**91,79**	61,94	66,41	83,58	62,68	44,77
Ovarian	**98,01**	93,63	92,47	91,66	91,70	82,95

Dapat dilihat pada tabel 4, dari semua pengujian yang dilakukan untuk kelima data penelitian ini mendapatkan hasil terbaik saat model klasifikasi CART digabungkan dengan reduksi dimensi DWT. Proses reduksi dimensi DWT dapat mengurangi atribut yang tidak relevan tanpa menghapus atribut penting pada data dengan baik. Hal ini membantu model klasifikasi untuk mendapatkan hasil akurasi yang lebih baik. *Colon Cancer, Lung Cancer, Prostate Tumor dan Ovarian Cancer* mendapatkan hasil akurasi >90%. Namun, model masih

belum bisa menghasilkan akurasi yang tinggi pada data *Breast Cancer*. Salah satu penyebabnya karena data *Breast Cancer* memiliki kompleksitas yang tinggi dibandingkan data yang lain namun reduksi dimensi DWT belum bisa menangani data tersebut.

Parameter *max feature* merupakan salah satu parameter yang mempengaruhi hasil akurasi. Dilihat pada hasil penelitian ini, teori pada penelitian [22] tidak berlaku untuk semua data. Data *Breast Cancer* dan *Colon Tumor* mendapat akurasi terbaik dengan menggunakan *max feature sqrt* yaitu jumlah fitur sebesar akar dua dari jumlah data sedangkan data lainnya mendapatkan akurasi terbaik dengan menggunakan *max feature none* yaitu tidak ada pembatasan jumlah fitur.

3.2. Performa *Random Forest*-DWT dalam Klasifikasi *Microarray*

Parameter dan *k-fold* yang digunakan dapat dilihat pada tabel 5..

Tabel 5. Parameter

Parameter	Nilai
K-fold	1,2,3,4,5,6,7,8,9,10
N trees	10,20,30,40,50,60,70,80,90,100
Maximum Split	1,2,3,4,5,6,7,8,9,10
Maximum Features	'None','sqrt','log2'
DWT Filter	Db2, db4, db7, db8, db10
DWT Level	1,2,3,4,5,6

N trees, *Max Depth* dan *Max Features* merupakan parameter *random forest* yang digunakan pada penelitian ini. Parameter *Max Depth* dan *Max Feature* pada *random forest* memiliki fungsi yang sama dengan CART. Nilai terbaik untuk setiap parameter dapat dilihat pada tabel 6

Tabel 6. Parameter Terbaik

Data	K-fold	N trees	Max Depth	DWT Filter	DWT Level
Breast	2	80	8	Db4	1
Colon	2	20	9	Db8	3
Lung	3	30	7	Db2	5
Prostate	8	40	6	Db10	1
Ovarian	5	60	10	Db2	5

Parameter *N tree* merupakan jumlah *tree* yang dibangun pada model *random forest*. Semakin besar nilai *N tree* tidak menjamin bahwa model akan menghasilkan akurasi yang optimal [23]. Hasil akurasi yang didapatkan dengan menggunakan parameter terbaik dapat dilihat pada tabel 7.

Tabel 7. Hasil Akurasi *Random Forest*

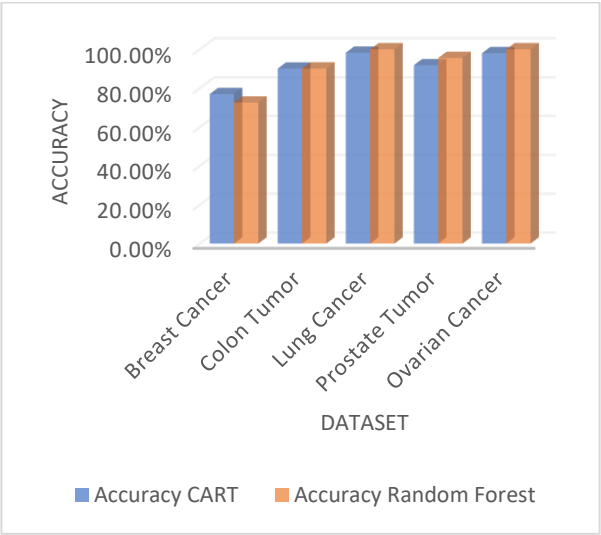
Data	Akurasi RF-DWT(%)			Akurasi RF No DWT(%)		
	None	Sqrt	Log2	None	sqrt	Log2
Breast	69,43	**72,58**	58,86	63,21	65,33	66,28
Colon	81,98	81,88	**90,10**	73,65	68,92	77,04
Lung	**100**	98,33	97,18	98,33	98,88	96,04
Prostate	**95,49**	91,03	88,09	91,72	88,78	85,06
Ovarian	96,42	**100**	99,60	94,04	94,44	92,85

Tabel 7 menunjukan bahwa model klasifikasi *Random Forest* menghasilkan performansi yang lebih baik jika meggunakan reduksi dimensi DWT daripada tidak menggunakan DWT. Data *Colon Tumor* dan *Prostate Tumor* mampu menghasilkan akurasi >90% sedangkan *Lung Cancer* dan *Ovarian Cancer* menghasilkan akurasi sebesar 100%. Hal ini menunjukan bahwa proses reduksi dimensi DWT dapat mengurangi kompleksitasi data dengan baik dengan menghapus atribut-atribut yang tidak relevan. Namun, data *Breast Cancer* mendapatkan akurasi terbaik sebesar 72.58% dimana hasil tersebut dinilai belum menghasilkan performansi yang baik. Penyebab nya sama seperti model CART-DWT dimana reduksi dimensi DWT belum bisa menangani kompleksitas data yang dimiliki *Breast Cancer*.

Dalam penelitian ini, data *Breast Cancer* dan *Ovarian Cancer* mendapat akurasi terbaik dengan menggunakan *max feature sqrt* yaitu sebesar akar dua dari jumlah data, data *Colon Tumor* mendapat akurasi terbaik dengan *max feature log2* yaitu sebesar *log2* dari jumlah data fitur dan data lainnya mendapatkan akurasi terbaik dengan menggunakan *max feature none* yaitu tidak ada pembatasan *max feature*. Pengujian ini mendapatkan kesimpulan yang sama dengan hasil dari pengujian CART dimana semakin kecil nilai *max feature* tidak selalu menghasilkan performansi yang semakin rendah.

3.3. Perbandingan Performa CART dan *Random Forest* dalam Klasifikasi *Microarray*

Penelitian ini melakukan pengujian menggunakan dua model klasifikasi yaitu CART dan *Random Forest*. Tujuannya untuk mengetahui metode klasifikasi yang paling baik dalam mengklasifikasikan data *microarray*. Perbandingan hasil dari kedua metode dapat dilihat pada gambar 4.



Gambar 4. Perbandingan Akurasi CART dan RF

Berdasarkan gambar 4, dari kelima data yang digunakan pada penelitian ini, empat data mendapatkan hasil akurasi terbaik saat menggunakan metode *Random*

Forest dan satu data mendapatkan akurasi terbaik saat menggunakan metode CART. Hasil performansi *Random Forest* dipengaruhi oleh pengambilan atribut dalam pembuatan *tree*. Akurasi akan semakin baik jika *tree* dibangun dapat menggambarkan karakteristik data secara general [23]. Pada penelitian ini, teknik tersebut bekerja baik untuk data *colon tumor*, *lung cancer*, *prostate tumor* dan *ovarian cancer* dimana dapat mengurangi terjadinya *overfitting* namun kurang cocok untuk data *breast cancer*. Hal tersebut disebabkan karena banyak atribut penting yang tidak terpilih untuk menjadi simpul pada *tree* yang dibangun dan menyebabkan tidak tergambarnya karakteristik yang kuat pada data.

4. Kesimpulan

Berdasarkan hasil penelitian yang dilakukan, kesimpulan yang dapat diambil dari penelitian ini adalah metode klasifikasi CART dan *Random Forest* jika digabungkan dengan reduksi dimensi DWT mampu mengklasifikasikan data *microarray* dalam pendeteksian kanker. Reduksi dimensi DWT mampu mengoptimalkan hasil performansi dari metode klasifikasi yang digunakan. Hasil klasifikasi menggunakan reduksi dimensi DWT selalu mendapatkan akurasi terbaik untuk semua dataset. Akurasi tertinggi yang didapat untuk metode klasifikasi *random forest* sebesar 100% untuk data *Lung Cancer* dan *Ovarian Cancer*. Metode klasifikasi CART mendapatkan hasil akurasi tertinggi sebesar 98.26% untuk data *Lung Cancer*. Namun, untuk data *Breast Cancer*, DWT masih belum mampu mengurangi kompleksitas dari data secara maksimal. Akurasi terbaik untuk data *Colon Tumor*, *Lung Cancer*, *Prostate Tumor*, dan *Ovarian Cancer* didapatkan dari kombinasi model klasifikasi *random forest* dan reduksi dimensi DWT, namun model klasifikasi CART dan reduksi dimensi DWT dapat menghasilkan performansi terbaik untuk data *Breast Cancer*. Hal ini disebabkan karena teknik dalam *random forest* belum dapat menjaga variansi data sehingga *tree* yang dibangun tidak menggambarkan karakteristik data secara general.

Parameter *max feature* pada model klasifikasi *random forest* dan CART juga sangat berpengaruh terhadap performansi model. Hasil penelitian ini menunjukan untuk beberapa data, nilai *max feature* yang lebih sedikit dibandingkan jumlah fitur pada data dapat mengoptimalkan performansi yaitu dengan mengurangi terjadinya *overfitting* dan mempercepat proses komputasi. Namun, hal tersebut dapat menurunkan perfomansi untuk data lainnya. Hal ini diakibatkan karena atribut penting terbuang dan tidak dapat dijadikan kandidat untuk simpul tree sehingga untuk beberapa data lebih baik menggunakan jumlah keseluruhan fitur sebagai nilai *max feature*.

Daftar Rujukan

[1] International Agency for Research on Cancer, 2019. Cancer Tomorrow. [Online] (Updated March 2019) Tersedia di: https://gco.iarc.fr/tomorrow/home [Accessed 14 May 2020]
[2] International Agency for Research on Cancer, 2019. All Cancer Fact Sheet. [Online] (Updated March 3019) Tersedia di: https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf [Accessed 14 May 2020]
[3] Bennet, J., Chilambuchelvan A. G. and Kannan A., 2014. A Discrete Wavelet based feature extraction and Hybrid Classification technique for Microarray data analysis. Scientific World Journal, Hindawi publishing corporation, vol. Article ID 195470, no. 9, 2014. doi: https://doi.org/10.1155/2014/195470.
[4] Adiwijaya, Wisesty UN, et al., 2018. Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification. *Journal of Computer Science*, 14(11), pp. 1521-1530. doi: 10.3844/jcssp.2018.1521.1530.
[5] Adiwijaya, 2018. Deteksi Kanker Berdasarkan Klasifikasi Microarray Data. *Jurnal Media Informatika Budidarma*, 2(4), pp. 181-186. doi: http://dx.doi.org/10.30865/mib.v2i4.1043.
[6] Aydadenta, H, 2018. A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest. *Journal of Information Processing Systems*, 14(5), pp. 1167-1175. doi: https://doi.org/10.3745/JIPS.04.0087.
[7] CHEN, Lei, and Yi-hui LIU, 2011. Classification based on CART algorithm for microarray data of lung cancer. *China Journal of Bioinformatics* 3, 9(3), pp. 229-234.
[8] Khoirunnisa A, Rohmawati AA., 2019. Implementing Principal Component Analysis and Multinomial Logit for Cancer Detection based on Microarray Data Classification, In *2019 7th International Conference on Information and Communication Technology (ICoICT)*, pp. 1-6, IEEE. Kuala Lumpur, Malaysia, 2019 Jul 24. doi : 10.1109/ICoICT.2019.8835320.
[9] Khadijah K., Hartati S., 2015. Klasifikasi Data Microarray Menggunakan Discrete Wavelet Transform dan Extreme Learning Machine. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 9(1), pp. 33-42. doi: https://doi.org/10.22146/ijccs.6638.
[10] Sari, P. K., and Purwadinata, A., 2019. Analysis Characteristics of Car Sales In E-Commerce Data Using Clustering Model. *Journal of Data Science and Its Applications*, *2*(1), pp. 19-28. doi: https://doi.org/10.21108/jdsa.2019.2.19.
[11] Daeli, N. O. F., & Adiwijaya, A., 2020. Sentiment Analysis on Movie Reviews using Information Gain and K-Nearest Neighbor. *Journal of Data Science and Its Applications*, *3*(1), pp. 1-7. doi: https://doi.org/10.34818/jdsa.2020.3.22.
[12] Pratiwi, Melati Suci, and Annisa Aditsania, 2018. Cancer Detection Based on Microarray Data Classification using Genetic Bee Colony (GBC) and Conjugate Gradient Backpropagation with Modified Polak Ribiere (MBP-CGP). *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. IEEE, pp. 163-168. doi: 10.1109/IC3INA.2018.8629538
[13] Misiti, M., Misiti Y., Oppenheim G., and Poggi J.M., 2012. *Wavelet Toolbox User's Guide (R2012a)*, Natick, Mass,USA: The MathWorks.
[14] Rohmawati, Aniq A., A Daubechies wavelet transformation to optimize modeling calibration of active compound on drug plants. In *2017 5th International Conference on Information and Communication Technology (ICoIC7).* pp. 1-4, IEEE. Malacca, Malaysia, 2017 May 17-19. doi: 10.1109/ICoICT.2017.8074666.
[15] Fugal, D. L., 2009. *Conceptual wavelets in digital signal processing: an in-depth, practical approach for the non-mathematician*. CA, San Diego: Space & Signals Technical Pub, pp. 1-78.
[16] Mandala, S., Cai Di, T., and Sunar, M. S., 2020. ECG-based prediction algorithm for imminent malignant ventricular arrhythmias using decision tree. *Plos one*, *15*(5), pp. e0231635. doi: https://doi.org/10.1371/journal.pone.0231635.
[17] Mabarti, I., 2020. Implementation of Minimum Redundancy Maximum Relevance (MRMR) and Genetic Algorithm (GA) for Microarray Data Classification with C4. 5 Decision Tree. *Journal*

of Data Science and Its Applications, 3(1), pp. 38-47. doi: https://doi.org/10.34818/jdsa.2020.3.37.

[18] Timofeev, R., 2004. *Classification and regression trees (cart) theory and applications*. M.A. Berlin: Humboldt University of Berlin.

[19] Cutler, A., and Stevens, J. R., 2006. [23] random forests for microarrays. *Methods in enzymology*, 411, pp. 422-432. doi: https://doi.org/10.1016/S0076-6879(06)11023-X.

[20] Purnomoputra, R. B., Adiwijaya, A., and Wisesty, U. N., 2019. Sentiment Analysis of Movie Review using Naïve Bayes Method with Gini Index Feature Selection. *Journal of Data Science and Its Applications*, 2(2), pp. 85-94. doi: https://doi.org/10.34818/jdsa.2019.2.36.

[21] Xia, Y., Liu, C., Li, Y., and Liu, N., 2017. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, pp. 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

[22] Zhang, Y., and Gao, J., 2017. MLFSdel: An accurate approach to discover genome deletions. In *2017 5th International Conference on Machinery, Materials and Computing Technology (ICMMCT 2017)*, Atlantis Press. Beijing, China, 2017 March 25-26.

[23] Agusta, Z.P., Adiwijaya, 2019. Modified balanced random forest for improving imbalanced data prediction. International Journal of Advances in Intelligent Informatics, 5(1), pp.58-65.