ANALISIS SELEKSI FITUR *BINARY PARTICLE SWARM OPTIMIZATION* PADA KLASIFIKASI KANKER BERDASARKAN DATA *MICROARRAY* MENGGUNAKAN *DISTANCE WEIGHTED K-NEAREST NEIGHBORS*

Yanche Kurniawan Mangalik[1], Triando Hamonangan Saragih[*,2], Dodon Turianto Nugrahadi[3], Muliadi[4], Muhammad Itqan Mazdadi[5]

[1,2,3,4,5] Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat
[1]yanche.k.m@gmail.com, [2]triando.saragih@ulm.ac.id, [3]dodonturianto@ulm.ac.id, [4]muliadi@ulm.ac.id, [5]mazdadi@ulm.ac.id

Abstrak

Salah satu penyakit mematikan yang menjadi penyebab kematian terbesar secara global adalah kanker. Kematian akibat kanker dapat diredam dengan deteksi dini terhadap kanker dengan memanfaatkan teknologi *microarray*. *Microarray* merupakan teknologi yang berisi kumpulan gen manusia yang telah direaksikan dengan cDNA yang dilabeli dengan pewarna *fluorescent* sehingga menghasilkan warna-warna tertentu yang kemudian diterjemahkan menjadi data *microarray*. Namun, teknologi ini memiliki kekurangan, yaitu jumlah gen (fitur) yang terlalu banyak. Kekurangan tersebut dapat diatasi dengan melakukan seleksi fitur terhadap data *microarray*. Salah satu algoritma seleksi fitur yang dapat digunakan adalah *Binary Particle Swarm Optimization* (BPSO). Pada penelitian ini, dilakukan seleksi fitur dengan BPSO pada data *microarray* dan klasifikasi menggunakan *Distance Weighted K-Nearest Neighbors* (DWKNN). Kemudian akan dilihat perbandingan hasil akurasi, presisi, *recall*, dan *f1-score* antara DWKNN dan DWKNN dengan penambahan pada seleksi fitur. Seleksi fitur dan klasifikasi pada *dataset Leukemia* menghasilkan akurasi, presisi, *recall*, dan *f1-score* tertinggi beturut-turut sebesar 93,12%, 94,39%, 95,92%, dan 94,8%. Pada *dataset Lung Cancer* diperoleh akurasi, presisi, *recall*, dan *f1-score* tertinggi beturut-turut sebesar 98,36%, 98,77%, 99,35%, dan 99,03%. Pada *dataset Prostate Cancer* diperoleh akurasi, presisi, *recall*, dan *f1-score* tertinggi beturut-turut sebesar 86,81%, 89,13%, 88,04%, dan 88,07%. Pada *dataset Diffuse Large B-Cell Lymhpome* (DLBCL) diperoleh akurasi, presisi, *recall*, dan *f1-score* tertinggi beturut-turut sebesar 85,8%, 93,21%, 88,1%, dan 89,76%. Hasil perbandingan menunjukkan peningkatan akurasi, presisi, *recall*, dan *f1-score* pada algoritma DWKNN dengan seleksi fitur BPSO dibandingkan dengan algoritma DWKNN tanpa seleksi fitur BPSO.

Kata kunci : *microarray*, seleksi fitur, klasifikasi, *binary particle swarm optimization, distance weighted k-nearest neighbors.*

1. Pendahuluan

Salah satu penyakit paling mematikan dan menjadi salah satu penyebab kematian terbesar secara global adalah kanker. Resiko kematian akibat kanker dapat diredam dengan menggunakan strategi deteksi dini terhadap kanker. Deteksi dini terhadap kanker dapat dilakukan dengan memanfaatkan teknologi *microarray*. Adiwijaya (2018) berpendapat ribuan ekspresi genetik dari berbagai sampel DNA dapat diamati secara bersamaan menggunakan teknologi DNA *microarray* sehingga teknologi DNA *microarray* banyak dimanfaatkan untuk melakukan deteksi kanker. Rani & Ramyachitra (2018) berpendapat bahwa data *microarray* memiliki kekurangan dimana data *microarray* memiliki dimensi data yang tinggi yang diakibatkan oleh jumlah gen yang sangat banyak. Kekurangan ini dapat menyebabkan proses klasifikasi yang tidak baik.

Pada penelitian terdahulu yang dilakukan oleh Manikandan et al. (2017), permasalahan dimensi tinggi diatasi dengan melakukan seleksi fitur menggunakan pendekatan *wrapper subset selection*. Dalam penelitiannya, seleksi fitur dilakukan pada 2 jenis data berdimensi tinggi, yaitu data *microarray* dan data *text mining*. Penerapan seleksi fitur tersebut berhasil meningkatkan akurasi dari 2 metode klasifikasi yang berbeda. Salah satu algoritma yang dapat digunakan untuk melakukan seleksi fitur adalah *Binary Particle Swarm Optimization* (BPSO). Amrullah et al. (2020) berpendapat bahwa BPSO mampu menyeleksi fitur data *microarray* sebesar 48% dari total keseluruhan fitur. Dalam penelitiannya BPSO diuji terhadap 5 data *microarray* kemudian diklasifikasi

[*] Penulis Korespondensi (T. H. Saragih)
Email: triando.saragih@ulm.ac.id

menggunakan algoritma *Backpropagation* dan *Combining Conjugate Gradient Backpropagation* (CGBP) dengan hasil bahwa BPSO berhasil meningkatkan rata-rata akurasi dari algoritma CGBP hingga mencapai 86,11%. Untuk melakukan deteksi kanker, diperlukan algoritma komputasi berupa algoritma klasifikasi. Salah satu algoritma yang banyak digunakan untuk masalah klasifikasi adalah *K-Nearest Neighbor* (KNN). Ma'wa et al. (2019) melakukan penelitian untuk mengklasifikasikan data *microarray* menggunakan KNN. Dalam penelitiannya diperolah bahwa KNN memiliki permasalahan sensitivitas pemilihan nilai k yang dapat menimbulkan *noise* (jika terlalu kecil) dan batasan klasifikasi yang kabur (jika terlalu besar). Pada penelitian yang dilakukan oleh Hardianti et al. (2017), kekurangan algoritma KNN tersebut diatasi menggunakan algoritma *Distance Weighted K-Nearest Neighbors* (DWKNN). Dalam penelitiannya algoritma diuji pada nilai k yang berubah-ubah untuk prediksi masa studi mahasiswa. Hasilnya algoritma DWKNN memiliki akurasi yang lebih stabil dibandingkan dengan algoritma KNN. DWKNN tidak akan terpengaruh oleh kelas mayoritas sehingga dapat mengurangi sensitivitas pemilihan nilai k pada KNN karena menggunakan pembobotan ganda dengan memberikan bobot yang lebih besar kepada tetangga yang memiliki jarak terdekat. Penelitian lain yang dilakukan Chrismanto et al. (2020) menguji DWKNN untuk klasifikasi komentar spam dan non spam pada Instagram menggunakan k = 1 sampai dengan k = 5 juga memberikan hasil yang sama dimana DWKNN selalu mengungguli KNN pada setiap pemilihan nilai k.

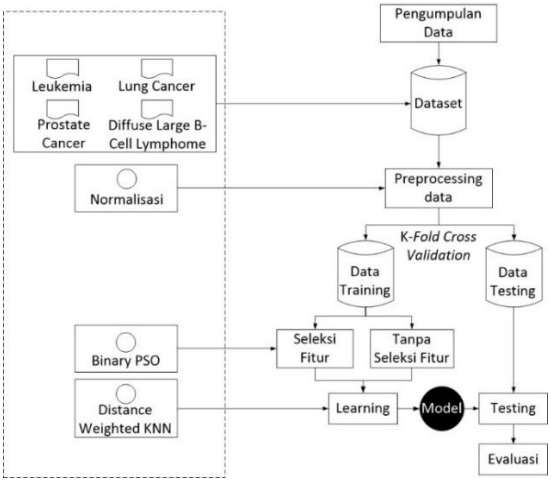
Berdasarkan penelitian sebelumnya, diketahui bahwa dimensi data yang tinggi pada data *microarray* menyebabkan data memiliki fitur-fitur yang tidak relevan dan mengurangi kinerja dari *classifier*, sehingga perlu dilakukan seleksi fitur untuk memilih fitur-fitur yang lebih relevan. Pada penelitian ini, algoritma BPSO digunakan untuk melakukan seleksi fitur. Pemilihan metode ini didasari oleh penelitian yang dilakukan oleh Amrullah et al. (2020) dimana BPSO mampu mereduksi dimensi data yang tinggi dan mampu meningkatkan kinerja dari metode klasifikasi yang digunakan. Untuk proses klasifikasi, penulis akan menggunakan algoritma DWKNN yang didasari oleh penelitian terdahulu yang dilakukan oleh Hardianti et al. (2017) dan Chrismanto et al. (2020) untuk memperbaiki kelemahan dari metode KNN. Analisis juga dilakukan pada penerapan seleksi fitur BPSO terhadap algoritma DWKNN dalam mengklasifikasikan data *microarray*. Analisis dilakukan untuk mengetahui pengaruh yang diberikan oleh seleksi fitur BPSO terhadap performa dari algoritma DWKNN serta melihat bagaimana seleksi fitur BPSO dalam mengatasi permasalahan dimensi data yang tinggi pada data *microarray*. Analisis dilakukan dengan membandingkan performa antara algoritma DWKNN dengan penerapan seleksi fitur BPSO dan DWKNN tanpa penerapan seleksi fitur BPSO.

2. Metode Penelitian

Prosedur penelitian yang digunakan pada penelitian ini dapat dilihat pada Gambar 1. Berikut merupakan penjelasan dari setiap prosedur penelitian yang dilakukan.

2.1 Pengumpulan Data

Data yang dikumpulkan dan akan digunakan dalam penelitian ini bersumber dari *Kent Ridge Biomedical Dataset Repository* berupa 4 jenis penyakit kanker yaitu *Leukemia*, *Lung Cancer*, *Prostate Cancer*, *Diffuse Large B-Cell Lymphoma* (DLBCL). Deskripsi dari keempat *dataset* yang digunakan pada penelitian ini ditampilkan pada Tabel 1.



Gambar 1. Diagram alur prosedur penelitian

Tabel 1. Detail data yang digunakan

Dataset	Jumlah Fitur	Jumlah Sampel	Kelas (Jumlah)
Leukemia	7129	72	ALL (47), AML (25)
Lung Cancer	12533	181	ADCA (150), *Mesothelioma* (31)
Prostate Cancer	12600	136	Tumor (77), Normal (59)
DLBCL	7129	77	DLBCL (58), FL (19)

2.2 *Preprocessing* Data

Pada tahap preprocessing, akan dilakukan normalisasi menggunakan *Min Max Normalization* terhadap data yang digunakan. Normalisasi data merupakan tahap yang penting untuk dilakukan karena biasanya *dataset* yang digunakan memiliki perbedaan nilai rentang data yang akan mempengaruhi hasil analisis data. Normalisasi adalah sebuah proses untuk mengubah rentang nilai atribut yang berbeda hingga memiliki bobot yang sama dan berada pada rentang yang lebih kecil. Normalisasi data akan meminimalisir terjadinya *noise* dan relevansi yang rendah pada data, sehingga dapat memberikan kinerja yang baik pada proses klasifikasi (Suryanegara et al., 2021).

Normalisasi bertujuan untuk mengurangi perbedaan rentang nilai yang besar antar data. Data yang digunakan akan dikonversi nilainya hingga berada pada rentang 0 sampai 1. *Min Max normalization* dihitung menggunakan persamaan (1) (Nishom, 2019).

$$x' = \frac{x - nilai_{min}}{nilai_{max} - nilai_{min}} \quad (1)$$

Dengan x adalah data yang akan dinormalisasi, $nilai_{min}$ adalah nilai minimum data, dan $nilai_{max}$ adalah nilai maksimum data.

2.3 Pembagian Data

K-Fold Cross Validation adalah suatu cara yang bisa digunakan untuk mengevaluasi kinerja dari sebuah metode. *K-Fold Cross Validation* bekerja dengan membagi data menjadi data *training* dan data *testing* secara acak sebanyak nilai k (Cahyanti et al., 2020). Dalam *K-Fold Cross Validation*, nilai k = 5, k = 10, dan k = 20 adalah nilai k yang paling sering digunakan. Nilai k yang digunakan mempengaruhi stabilitas dari *K-Fold Cross Validation* (Zhou, 2021).

Pembagian data dalam penelitian ini dilakukan menggunakan *K-Fold Cross Validation* melalui beberapa skenario pembagian data dengan menggunakan nilai k = 2 sampai dengan k = 10. Skenario ini bertujuan untuk menemukan nilai k yang menghasilkan performa terbaik pada proses seleksi fitur dan juga proses klasifikasi.

2.4 Seleksi Fitur

Selanjutnya akan dilakukan seleksi fitur menggunakan BPSO. BPSO merupakan salah satu algoritma yang dapat digunakan untuk menangani masalah seleksi fitur. BPSO memiliki mekanisme yang mirip dengan PSO, terutama untuk memperbarui nilai *velocity* (v). Perbedaan antara BPSO dan PSO terletak pada variabel x_{id}, p_{id}, dan p_{gd} yang hanya memiliki nilai 0 atau 1 (Mafarja et al., 2018). Langkah awal BPSO dalam seleksi fitur adalah menginisialisasi jumlah partikel yang akan digunakan. Partikel-partikel tersebut memiliki posisi yang akan diinisialisasi secara acak dan akan bergerak pada ruang pencarian dengan *velocity* (kecepatan) tertentu. Tiap partikel akan terus bergerak selama proses iterasi berlangsung untuk menemukan solusi terbaik hingga kondisi iterasi terpenuhi. Tiap posisi memiliki nilai antara 0 dan 1 dimana nilai 0 menandakan fitur yang tidak terpilih dan nilai 1 menandakan fitur yang terpilih. Kecepatan partikel akan diperbarui tiap iterasi dengan persamaan (2) (Gohzali et al., 2019).

$$v_{pd}^{new} = \omega * v_{pd}^{old} + c_1 * rand1 * \left(pbest_{pd} - x_{pd}^{old}\right) + c_2 * rand2 * \left(gbest_{pd} - x_{pd}^{old}\right), \quad (2)$$

Dimana ω adalah *inertia weight*, c_1 dan c_2 adalah parameter percepatan, rand1 dan rand2 adalah nilai acak antara 0 sampai 1, v_{pd}^{new} adalah kecepatan partikel setelah diperbarui, v_{pd}^{old} adalah kecepatan partikel sebelum diperbarui dan x_{pd}^{old} adalah posisi partikel saat ini.

Pada penelitian ini BPSO dilakukan dengan beberapa skenario percobaan pada parameter *cognitive learning* (c_1), *social learning* (c_2), dan *inertia weights* (ω) dengan kombinasi nilai yang dapat dilihat pada Tabel 2. Skenario kedua dilakukan pada parameter *population* (p) dengan nilai p = 10, p = 20, p = 30, dan p = 50. Setiap skenario percobaan akan menghasilkan fitur-fitur yang dievaluasi menggunakan DWKNN untuk menemukan kombinasi parameter terbaik.

Tabel 2. Kombinasi parameter c_1 dan c_2

[c_1, c_2]	**1**	**1,25**	**1,5**	**1,75**	**2**
1	[1, 1]	[1, 1,25]	[1, 1,5]	[1, 1,75]	[1, 2]
1,25	[1,25, 1]	[1,25, 1,25]	[1,25, 1,5]	[1,25, 1,75]	[1,25, 2]
1,5	[1,5, 1]	[1,5, 1,25]	[1,5, 1,5]	[1,5, 1,75]	[1,5, 2]
1,75	[1,75, 1]	[1,75, 1,25]	[1,75, 1,5]	[1,75, 1,75]	[1,75, 2]
2	[2, 1]	[2, 1,25]	[2, 1,5]	[2, 1,75]	[2, 2]

2.5 Klasifikasi

Tahapan selanjutnya adalah melakukan klasifikasi dengan menggunakan DWKNN. DWKNN merupakan metode yang digunakan untuk mengatasi masalah pada KNN dan *Weighted K-Nearest Neighbors* (WKNN) dengan menggunakan pembobotan ganda (Tamrakar & Ibrahim, 2021). Pembobotan ganda pada DWKNN dilakukan untuk memberikan bobot yang berbeda pada setiap tetangga terdekat. Setelah dilakukan pembobotan ganda, tetangga dengan jarak terdekat memiliki bobot 1 sedangkan tetangga dengan jarak terjauh memiliki bobot 0 dan tetangga lain yang memiliki jarak diantara tetangga terdekat dan terjauh memiliki bobot diantara 0 dan 1. Pembobotan ganda dihitung menggunakan persamaan (3) (Chrismanto et al., 2020).

$$w_i = \begin{cases} \dfrac{d(x,x_k^{NN}) - d(x,x_i^{NN})}{d(x,x_k^{NN}) - d(x,x_1^{NN})} \times \dfrac{d(x,x_k^{NN}) + d(x,x_1^{NN})}{d(x,x_k^{NN}) + d(x,x_i^{NN})} & ,if\ d(x,x_k^{NN}) \neq d(x,x_1^{NN}) \\ 1 & ,if\ d(x,x_k^{NN}) = d(x,x_1^{NN}) \end{cases} \quad (3)$$

Dimana $d(x,x_k^{NN})$ adalah nilai tetangga terbesar, $d(x,x_1^{NN})$ adalah nilai tetangga terkecil, dan (x,x_i^{NN}) adalah nilai tetangga pada iterasi ke-i.

Pada penelitian ini klasifikasi dilakukan melalui dua pemodelan, yaitu DWKNN tanpa menggunakan BPSO dan DWKNN dengan menggunakan BPSO. Nilai parameter k pada DWKNN yang akan digunakan pada kedua model adalah k = 1 sampai k = 20 yang mengacu pada percobaan yang dilakukan dalam penelitian ini. Ketika nilai k lebih besar dari 20 akurasi yang dihasilkan menurun secara perlahan sehingga nilai k yang digunakan hanya sampai k = 20.

2.6 Evaluasi

Salah satu metode yang digunakan untuk evaluasi performa klasifikasi dalam *Machine Learning* adalah *Confusion Matrix*. Pada umumnya *Confusion Matrix* berbentuk tabel, dimana baris pada tabel mewakili kelas aktual dan kolom pada tabel mewakili kelas yang diprediksi (Mir & Dhage, 2018).

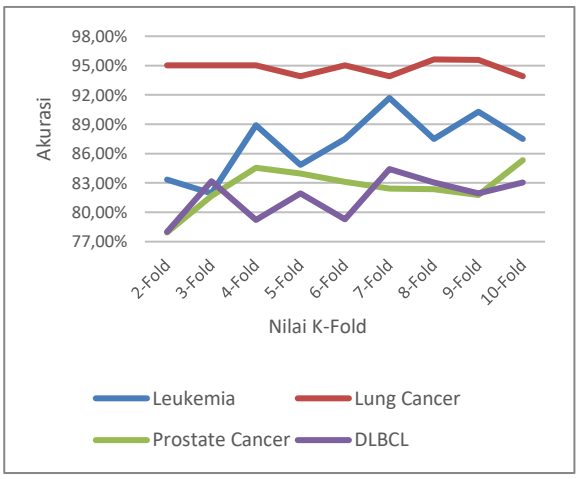
Kedua model yang telah dibuat akan memasuki tahap evaluasi menggunakan *Confusion Matrix* dan akan menghasilkan nilai-nilai berupa akurasi, presisi, *recall* dan *f1-score*. Selanjutnya, hasil pegujian pada model DWKNN dan BPSO-DWKNN tersebut akan dibandingkan.

3. Hasil dan Pembahasan

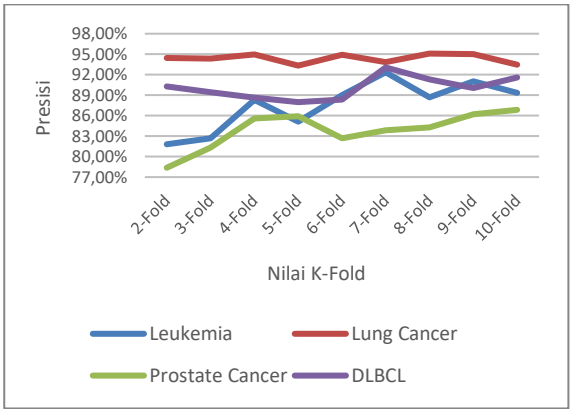
Dalam penelitian ini, seleksi fitur dan klasifikasi dilakukan dalam beberapa skenario percobaan. Skenario tersebut dilakukan terhadap pembagian data, parameter-parameter BPSO (c_1, c_2, ω dan p), dan parameter k pada DWKNN. Setiap fitur yang dihasilkan pada skenario percobaan parameter-parameter BPSO performanya akan dihitung melalui klasifikasi DWKNN dengan nilai k pada DWKNN sebesar k = 1. Skenario ini dilakukan untuk menemukan kombinasi nilai parameter yang dapat menghasilkan performa terbaik.

3.1 Skenario Pembagian Data Seleksi Fitur

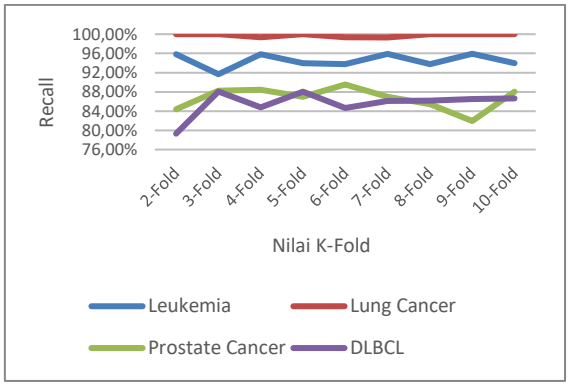
Pada tahapan ini, seleksi fitur dilakukan pada *dataset* yang telah dibagi menjadi data latih dan data uji menggunakan *K-Fold Cross Validation* dengan nilai k = 2 sampai dengan k = 10. Tiap pembagian data akan menghasilkan kombinasi fitur yang berbeda dengan performa yang berbeda juga. Tahapan ini dilakukan untuk mengetahui nilai pembagian k yang menghasilkan kombinasi fitur dengan performa terbaik. Grafik performa berupa akurasi, presisi, *recall*, dan *f1-score* dari kombinasi fitur hasil skenario pembagian data seleksi fitur pada keempat *dataset* berturut-turut dapat dilihat pada Gambar 2, Gambar 3, Gambar 4, dan Gambar 5.



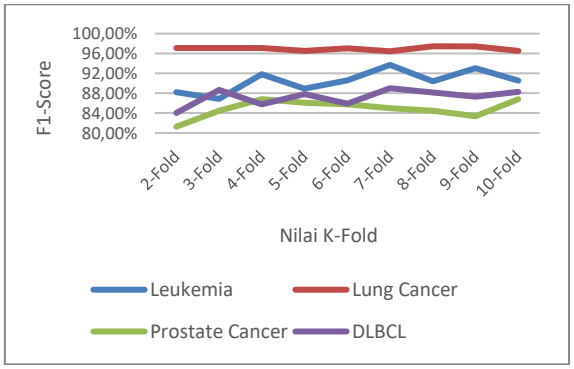
Gambar 2. Akurasi seleksi fitur tiap *fold*



Gambar 3. Presisi seleksi fitur tiap *fold*



Gambar 4. *Recall* seleksi fitur tiap *fold*



Gambar 5. *F1-score* seleksi fitur tiap *fold*

Berdasarkan keempat grafik tersebut, dapat dilihat bahwa pada *dataset Leukemia* pembagian k = 7 menghasilkan performa terbaik dengan akurasi, presisi, *recall* dan *f1-score* berturut-turut sebesar 91,69%, 92,35%, 95,92%, dan 93,7%. Pada *dataset Lung Cancer* diperoleh performa terbaik pada pembagian k = 8 dengan akurasi, presisi, *recall* dan *f1-score* berturut-turut sebesar 95,63%, 95,09%, 100%, dan 97,45%. Pada *dataset Prostate Cancer* diperoleh performa terbaik pada pembagian k = 10 dengan akurasi, presisi, *recall* dan *f1-score* berturut-turut sebesar 85,33, 86,85%, 88,04%, dan 86,8%. Pada *dataset* DLBCL diperoleh performa terbaik pada pembagian k = 7 dengan akurasi, presisi, *recall* dan *f1-score* berturut-turut sebesar 84,42%, 93,06%, 86,11%, dan 89,01%.

3.1 Skenario Pengaturan Parameter c_1, c_2 dan ω

Pada skenario ini, seleksi fitur dilakukan untuk menemukan kombinasi terbaik pada parameter c_1, c_2, dan ω pada BPSO. Kombinasi nilai parameter c_1 dan c_2 yang akan digunakan dapat dilihat pada Tabel 2. Setiap kombinasi nilai c_1 dan c_2 pada Tabel 2 akan dikombinasikan dengan nilai ω = 0,9 dan ω = 1. Kombinasi terbaik hasil skenario pengaturan parameter c_1, c_2, dan ω ditampilkan pada Tabel 3. Berdasarkan Tabel 3, pada *dataset Leukemia* diperoleh kombinasi terbaik pada $c_1 = 2$, $c_2 = 1{,}5$, ω = 1 dengan akurasi sebesar 93,12%, presisi sebesar 93%, *recall* sebesar 97,96%, *f1-score* sebesar 96,06%, dan fitur sebanyak 3726 fitur.

Pada *dataset Lung Cancer* akurasi terbaik diperoleh pada $c_1 = 1$, $c_2 = 1{,}5$, ω = 1 dan $c_1 = 1$, $c_2 = 2$, ω = 1 yaitu sebesar 96,71%. Akan tetapi, presisi, *recall*, *f1-score* yang dihasilkan oleh kombinasi pertama lebih besar, fitur yang dihasilkan juga lebih sedikit dibandingkan dengan kombinasi kedua. Dengan demikian kombinasi terbaik untuk *dataset Lung Cancer* adalah kombinasi pertama dengan akurasi sebesar 96,71%, presisi sebesar 96,28%, *recall* sebesar 100%, *f1-score* sebesar 98,08%, dan fitur sebanyak 6491 fitur.

Pada *dataset Prostate Cancer* diperoleh kombinasi terbaik pada $c_1 = 1{,}25$, $c_2 = 1{,}75$ dan ω = 0,9 dengan akurasi sebesar 86,76%, presisi sebesar 90,92%, *recall* sebesar 85,54%, *f1-score* sebesar 87,53%, dan fitur sebanyak 6507 fitur.

Pada *dataset* DLBCL akurasi terbaik diperoleh pada $c_2 = 1{,}25$, $c_2 = 1$, ω = 0,9; $c_1 = 1$, $c_2 = 1$, ω = 1; dan $c_1 = 2$, $c_2 = 2$, ω = 1 dengan akurasi sebesar 85,71%. Akan tetapi *recall* dan *f1-score* pada kombinasi pertama lebih besar dibandingkan dengan kombinasi kedua. Jumlah fitur yang dihasilkan pada kombinasi pertama juga lebih kecil dibandingkan dengan kombinasi ketiga. Dengan demikian kombinasi terbaik untuk *dataset* DLBCL adalah kombinasi pertama dengan akurasi sebesar 85,71%, presisi sebesar 93,57%, *recall* sebesar 87,9%, *f1-score* sebesar 90,2%, dan fitur sebanyak 3695 fitur.

Berdasarkan hasil percobaan pada *dataset Leukemia* dan *Lung Cancer*, diperoleh bahwa nilai parameter ω = 0,9 cenderung akan menghasilkan performa terbaik apabila dikombinasikan dengan $c_2 = 1{,}5$. Kemudian, berdasarkan hasil perobaan pada *dataset Prostate cancer* dan DLBCL, dapat dilihat bahwa nilai parameter ω = 1 cenderung akan menghasilkan performa terbaik apabila dikombinasikan dengan $c_1 = 1{,}25$.

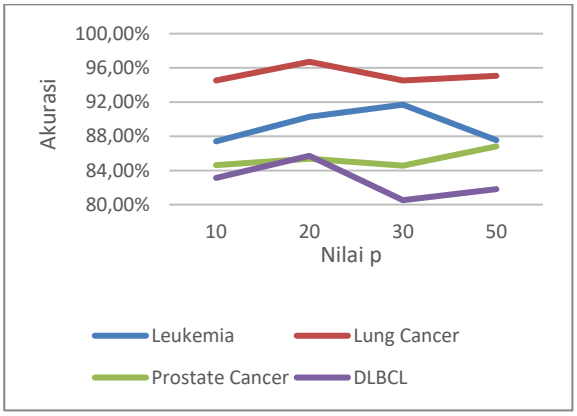
Tabel 3. Kombinasi c_1, c_2, dan ω yang menghasilkan performa terbaik

Dataset	c_1	c_2	ω	Akurasi	Presisi	*Recall*	*F1-Score*	Jumlah Fitur
Leukemia	2	1,5	1	93,12%	93%	97,96%	96,06%	3726
Lung Cancer	1	1,5	1	96,71%	96,28%	100%	98,08%	6491
	1	2	1	96,71%	96,22%	100%	98,06%	6522
Prostate Cancer	1,25	1,75	0,9	86,76%	90,92%	85,54%	87,53%	6507
DLBCL	1,25	1	0,9	85,71%	93,57%	87,9%	90,2%	3695
	1	1	1	85,71%	94,33%	86,11%	89,75%	3685
	2	2	1	85,71%	93,57%	87,9%	90,2%	3728

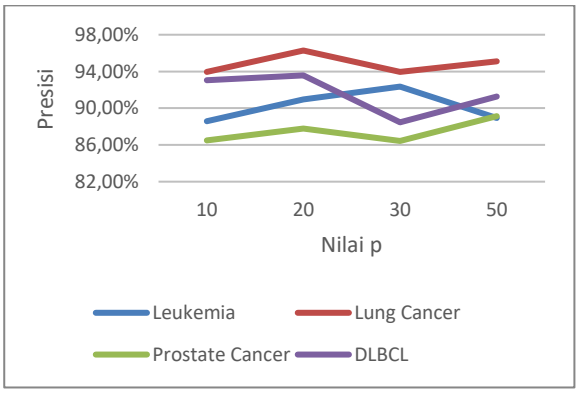
3.2 Skenario Pengaturan Parameter p

Nilai parameter p yang akan digunakan pada skenario ini adalah p = 10, p = 20, p = 30, dan p = 50. Fitur dengan performa terbaik pada skenario ini akan menjadi hasil akhir proses seleksi fitur. Fitur-fitur tersebut kemudian akan diterapkan pada data latih dan data uji, kemudian akan dilakukan proses klasifikasi menggunakan DWKNN

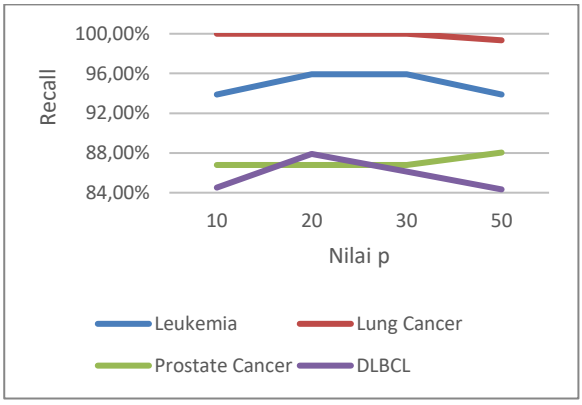
Grafik performa berupa akurasi, presisi, *recall*, dan *f1-score* dari kombinasi fitur hasil skenario pengaturan parameter p pada keempat *dataset* berturut-turut dapat dilihat pada Gambar 6, Gambar 7, Gambar 8, dan Gambar 9.



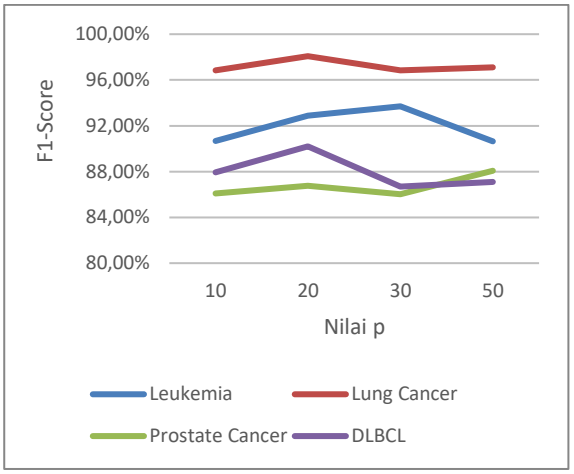
Gambar 6. Akurasi tiap nilai p



Gambar 7. Presisi tiap nilai p



Gambar 8. *Recall* tiap nilai p



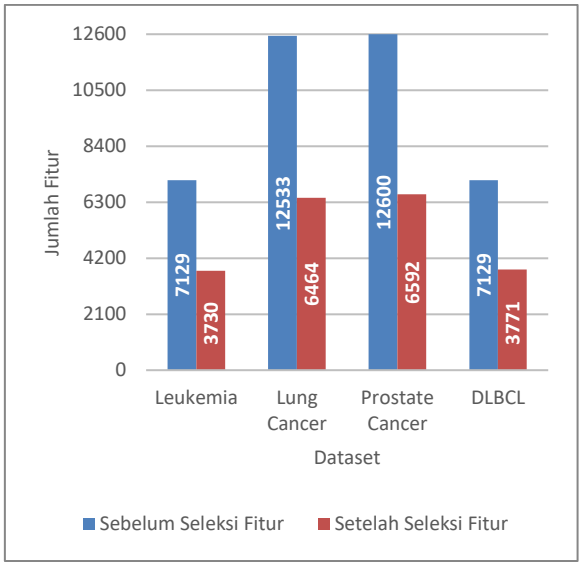
Gambar 9. *F1-score* tiap nilai p

Berdasarkan keempat grafik tersebut, pada *dataset Leukemia* performa terbaik diperoleh pada nilai p = 30 dengan akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 91,69%, 92,35%, 95,92%, 93,7% dengan jumlah fitur sebanyak 3730 fitur. Pada *dataset Lung Cancer* performa terbaik diperoleh pada nilai p = 20 dengan akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 96,71%, 96,28%, 100%, 98,08% dengan jumlah fitur sebanyak 6464 fitur. Pada *dataset Prostate Cancer* performa terbaik diperoleh pada nilai p = 50 dengan akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 86,81%, 89,13%, 88,04%, 88,07% dengan jumlah fitur sebanyak 6592 fitur. Pada *dataset* DLBCL performa terbaik diperoleh pada nilai p = 20 dengan akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 85,71%, 93,57%, 87,9%, 90,2% dengan jumlah fitur sebanyak 3771 fitur.

Berdasarkan data tersebut diperoleh bahwa fitur dengan performa terbaik dari setiap dataset dihasilkan dari jumlah populasi yang berbeda-beda. Pada *dataset Lung Cancer* dan DLBCL fitur dengan performa terbaik diperoleh pada p = 20. Performa terkecil pada *dataset Lung Cancer* diperoleh pada p = 30 dan p = 10. Sedangkan pada *dataset* DLBCL diperoleh pada p = 30. Pada *dataset Leukemia* dan *Prostate Cancer* fitur dengan performa terbaik diperoleh pada populasi dengan jumlah yang lebih besar yaitu p = 30 dan p = 50. Sedangkan fitur dengan performa terkecil pada *dataset Leukemia* dan *Prostate Cancer* diperoleh pada p = 10 dan p = 30. Hal ini menunjukkan bahwa jumlah populasi yang besar tidak selalu dapat menghasilkan fitur dengan performa terbaik, beberapa *dataset* menghasilkan fitur dengan performa terbaik pada jumlah populasi yang kecil dan tidak terlalu besar.

Berdasarkan tiga skenario percobaan yang dilakukan pada proses seleksi fitur, diperoleh bahwa rata-rata fitur tidak relevan yang dapat dihapus oleh seleksi fitur BPSO pada penelitian ini mencapai 47,72% fitur. Perbandingan jumlah fitur sebelum dan sesudah dilakukan seleksi fitur pada keempat

dataset yang digunakan dapat dilihat pada Gambar 10.



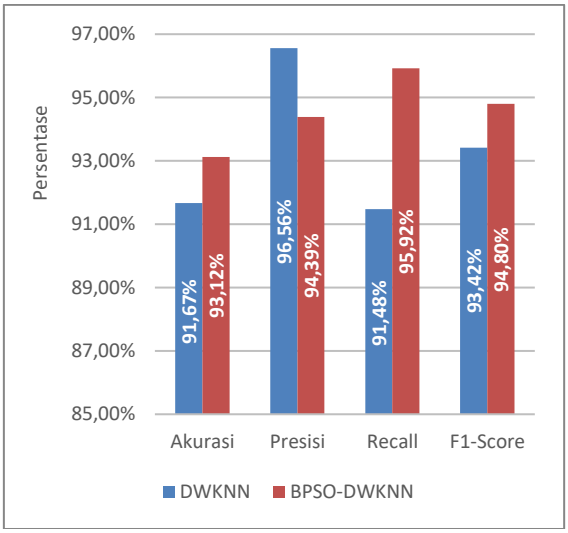
Gambar 10. Perbedaan jumlah fitur

3.3 Klasifikasi

Proses yang akan dilakukan selanjutnya adalah proses klasifikasi pada model DWKNN tanpa seleksi fitur dan model DWKNN yang menggunakan seleksi fitur (BPSO-DWKNN). Proses klasifikasi dilakukan pada pembagian data dengan nilai k = 2 sampai dengan k = 10 menggunakan *K-Fold Cross Validation*. Nilai k yang digunakan pada DWKNN adalah k = 1 sampai dengan k = 20. Kemudian, setiap model akan dievaluasi menggunakan *confusion matrix*. Nilai k beserta pembagian data yang menghasilkan performa terbaik dari kedua model ditampilkan pada Tabel 4.

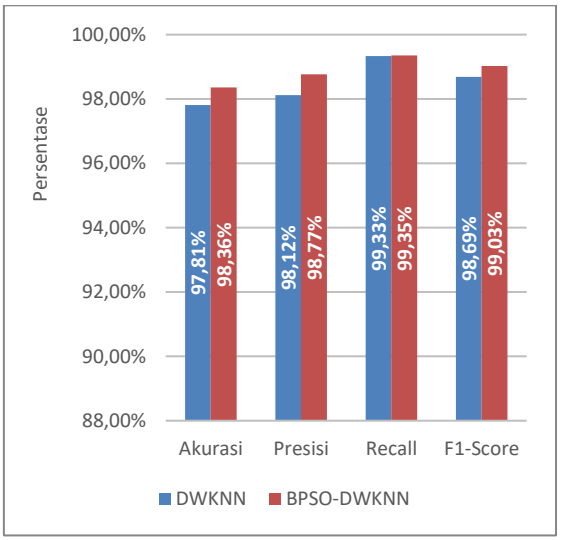
Berdasarkan hasil percobaan yang dilakukan pada *dataset Leukemia* dengan model DWKNN menghasilkan performa terbaik pada pembagian data k = 9 dan nilai k-DWKNN k = 13 dengan nilai akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 91,67%, 96,56%, 91,48%, dan 93,42%. Sedangkan dengan model BPSO-DWKNN menghasilkan performa terbaik pada pembagian data k = 7 dan nilai k-DWKNN k = 3 dengan nilai akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 93,12%, 94,39%, 95,92%, dan 94,8%. Grafik perbandingan performa klasifikasi yang dihasilkan pada *dataset Leukemia* dapat dilihat pada Gambar 11.

Pada *dataset Lung Cancer* dengan model DWKNN menghasilkan performa terbaik pada pembagian data k = 10 dan nilai k-DWKNN k = 18 dengan nilai akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 97,81%, 98,12%, 99,33%, dan 98,69.



Gambar 11. Perbandingan performa pada *dataset Leukemia*

Sedangkan dengan model BPSO-DWKNN menghasilkan performa terbaik pada pembagian data k = 9 dan nilai k-DWKNN k = 17 dengan nilai akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 98,36%, 98,77%, 99,35%, dan 99,03%. Grafik perbandingan performa klasifikasi pada *dataset Lung Cancer* dapat dilihat pada Gambar 12.

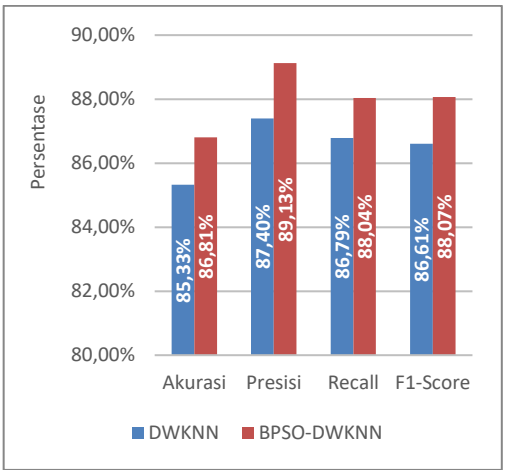


Gambar 12. Perbandingan performa pada *dataset Lung Cancer*

Pada *dataset Prostate Cancer* dengan model DWKNN menghasilkan performa terbaik pada pembagian data k = 10 dan nilai k-DWKNN k = 1 dan k = 2 dengan nilai akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 85,33%, 87,4%, 86,79%, dan 86,61. Sedangkan dengan model BPSO-DWKNN menghasilkan performa terbaik pada pembagian data k = 10 dan nilai k-DWKNN k = 1 dan k = 2 dengan nilai akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 86,81%, 89,13%, 88,04%, dan 88,07. Grafik perbandingan performa klasifikasi pada *dataset Prostate Cancer* dapat dilihat pada Gambar 13.

Tabel 4. Hasil performa terbaik klasifikasi pada keempat *dataset*

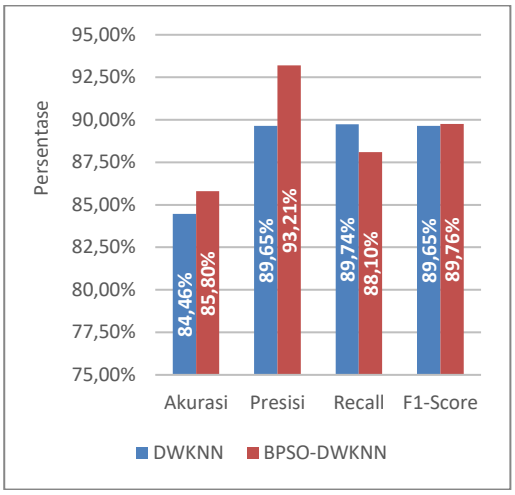
Dataset	**Model**	**K-Fold**	**K-DWKNN**	**Akurasi**	**Presisi**	***Recall***	***F1-Score***	**Jumlah Fitur**
Leukemia	DWKNN	9	13	91,67%	96,56%	91,48%	93,42%	7129
	BPSO-DWKNN	7	3	93,12%	94,39%	95,92%	94,80%	3730
Lung Cancer	DWKNN	10	18	97,81%	98,12%	99,33%	98,69%	12533
	BPSO-DWKNN	9	17	98,36%	98,77%	99,35%	99,03%	6464
Prostate Cancer	DWKNN	10	1 dan 2	85,33%	87,40%	86,79%	86,61%	12600
	BPSO-DWKNN	10	1 dan 2	86,81%	89,13%	88,04%	88,07%	6592
DLBCL	DWKNN	3	1 dan 2	84,46%	89,65%	89,74%	89,65%	7129
	BPSO-DWKNN	9	1 dan 2	85,80%	93,21%	88,10%	89,76%	3771



Gambar 13. Perbandingan performa pada *dataset Prostate Cancer*

Pada *dataset* DLBCL dengan model DWKNN menghasilkan performa terbaik pada pembagian data k = 3 dan nilai k-DWKNN k = 1 dan k = 2 dengan nilai akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 84,46%, 89,65%, 89,74%, dan 89,65%. Sedangkan dengan model BPSO-DWKNN menghasilkan performa terbaik pada pembagian data k = 9 dan nilai k-DWKNN k = 1 dan k = 2 dengan nilai akurasi, presisi, *recall*, dan *f1-score* berturut-turut sebesar 85,8%, 93,21%, 88,1%, dan 89,76%. Grafik perbandingan performa klasifikasi pada *dataset* DLBCL dapat dilihat pada Gambar 14.

Berdasarkan proses klasifikasi dan evaluasi yang telah dilakukan sebelumnya, dapat diketahui bahwa seleksi fitur BPSO memberikan pengaruh terhadap performa algoritma DWKNN. Seleksi fitur BPSO mampu meningkatkan akurasi, presisi, *recall*, dan *f1-score* dari algoritma DWKNN, meskipun terdapat penurunan presisi pada *dataset Leukemia* dan *recall* pada *dataset* DLBCL.



Gambar 14. Perbandingan performa pada *dataset* DLBCL

Penurunan pada beberapa performa seperti presisi dan *recall* ini terjadi karena masih adanya beberapa fitur tidak relevan yang masih tersisa atau ikut terseleksi. Fitur-fitur tersebut mempengaruhi proses klasifikasi sehingga presisi dan *recall* yang dihasilkan menjadi turun. Dari segi waktu komputasi seleksi fitur BPSO mampu mengurangi waktu komputasi dari algoritma DWKNN untuk melakukan proses klasifikasi. Selisih waktu komputasi yang dihasilkan pada proses klasifikasi pada *dataset Leukemia*, *Lung Cancer*, *Prostate Cancer*, dan DLBCL berturut-turut sebesar 2,91 detik, 11,04 detik, 11,12 detik, dan 4,13 detik. Hasil komputasi ini didapatkan dengan menghitung lama waktu komputasi pada saat proses klasifikasi antara model dengan seleksi fitur dan model tanpa seleksi fitur. Kemudian dihitung selisih waktu komputasi antara kedua model tersebut. Proses perhitungan tersebut dilakukan menggunakan *library time* pada pemrograman *python*.

4. Kesimpulan dan Saran

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa seleksi fitur BPSO mampu menghapus fitur yang tidak relevan hingga mencapai rata-rata sebesar 47,72% fitur. Penerapan seleksi fitur ini mampu meningkatkan akurasi, presisi, *recall*, dan *f1-score* pada algoritma DWKNN walaupun terjadi penurunan presisi pada *dataset Leukemia* dan *recall* pada *dataset* DLBCL. Dari hasil percobaan yang dilakukan, *dataset Leukemia* mengalami peningkatan akurasi, *recall*, *f1-score* berturut-turut sebesar 1,45%, 4,44%, 1,38% dan penurunan presisi sebesar 2.17%. *Dataset Lung Cancer* mengalami peningkatan akurasi, presisi, *recall*, dan *f1-score* beturut-turut sebesar 0,55%, 0,65%, 0,02%, dan 0,34%. *Dataset Prostate Cancer* mengalami peningkatan akurasi, presisi, *recall*, dan *f1-score* beturut-turut sebesar 1,48%, 1,73%, 1,25%, dan 1,46%. *Dataset* DLBCL mengalami peningkatan akurasi, presisi, *f1-score* beturut-turut sebesar 1,34%, 3,56%, 0,11%, dan penurunan *recall* sebesar 1,64%. Rata-rata peningkatan yang dihasilkan dari penerapan seleksi fitur terhadap DWKNN berturut-turut sebesar 1,21%, 0,94%, 1,02% dan 0,82%. Selain meningkatkan performa, penerapan seleksi fitur BPSO juga mengurangi waktu komputasi algoritma DWKNN untuk melakukan klasifikasi.

Daftar Pustaka:

Adiwijaya, A. (2018). Deteksi Kanker Berdasarkan Klasifikasi Microarray Data. *Jurnal Media Informatika Budidarma*, *2*(4), 181. https://doi.org/10.30865/mib.v2i4.1043

Amrullah, M. N. M., Adiwijaya, & Astuti, W. (2020). Implementation of Modified Backpropagation with Conjugate Gradient as Microarray Data Classifier with Binary Particle Swarm Optimization as Feature Selection for Cancer Detection. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *9*(3), 339–349. https://doi.org/10.32736/sisfokom.v9i3.978

Cahyanti, D., Rahmayani, A., & Husniar, S. A. (2020). Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara. *Indonesian Journal of Data and Science*, *1*(2), 39–43. https://doi.org/10.33096/ijodas.v1i2.13

Chrismanto, A. R., Lukito, Y., & Susilo, A. (2020). Implementasi Distance Weighted K-Nearest Neighbor Untuk Klasifikasi Spam & Non-Spam Pada Komentar Instagram. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *6*(2), 236–244. https://doi.org/10.26418/jp.v6i2.39996

Gohzali, H., Megawan, S., & Onggo, J. (2019). Rekomendasi Buku Menggunakan K-Nearest Neighbor (KNN) dan Binary Particle Swarm Optimization (BPSO). *Jurnal SIFO Mikroskil*, *20*(1), 1–5. https://www.ranks.nl/stopwords.

Hardianti, W., Indriani, F., & Nugroho, R. A. (2017). Analisis Perbandingan Algoritma Distance-Weighted KNN dan Algoritma KNN pada Prediksi Masa Studi Mahasiswa. *Seminar Nasional Ilmu Komputer (SOLITER)*, *1*, 108–117.

Ma'wa, S., Adiwijaya, & Rohmawati, A. A. (2019). Klasifikasi K-Nearest Neighbor untuk Data Microarray dengan Seleksi Genetic Algorithm. *E-Proceeding of Engineering*, *6*(2), 9838–9847.

Mafarja, M., Jarrar, R., Ahmad, S., & Abusnaina, A. A. (2018). Feature selection using Binary Particle Swarm optimization with time varying inertia weight strategies. *ACM International Conference Proceeding Series*, *18*, 1–9. https://doi.org/10.1145/3231053.3231071

Manikandan, G., Susi, E., & Abirami, S. (2017). Feature selection on high dimensional data using wrapper based subset selection. *2017 2nd International Conference on Recent Trends and Challenges in Computational Models, ICRTCCM 2017*, 320–325. https://doi.org/10.1109/ICRTCCM.2017.58

Mir, A., & Dhage, S. N. (2018). Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare. *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697439

Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. *Jurnal Informatika: Jurnal Pengembangan IT*, *4*(1), 20–24. https://doi.org/10.30591/jpit.v4i1.1253

Rani, R. R., & Ramyachitra, D. (2018). Microarray Cancer Gene Feature Selection using Spider Monkey Optimization Algorithm and Cancer Classification using SVM. *Procedia Computer Science*, *143*(2018), 108–116. https://doi.org/10.1016/j.procs.2018.10.358

Suryanegara, G. A. B., Adiwijaya, & Purbolaksono,

M. D. (2021). Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(1), 114–122.

Tamrakar, P., & Ibrahim, S. P. S. (2021). Lazy Learning Associative Classification with WkNN and DWkNN Algorithm. *ITM Web of Conferences*, *37*. https://doi.org/10.1051/itmconf/20213701023

Zhou, Z.-H. (2021). Ensemble Learning. In *Machine Learning*. https://doi.org/10.1007/978-981-15-1967-3_8