SELEKSI FITUR *HYBRID GREY WOLF OPTIMIZATION* DAN *PARTICLE SWARM OPTIMIZATION* PADA *DISTANCE BIASED NAÏVE BAYES* UNTUK KLASIFIKASI KANKER PAYUDARA

Ratna Septia Devi[1], Triando Hamonangan Saragih[2], Mohammad Reza Faisal[3], Dwi Kartini[4], Irwan Budiman[5]

[1,2,3,4,5] Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat.
Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714
[1]ratnaseptiadevi@gmail.com, [2] triando.saragih@ulm.ac.id, [3]reza.faisal@ulm.ac.id, [4]dwikartini@ulm.ac.id, [5]irwan.budiman@ulm.ac.id
Corresponding author: triando.saragih@ulm.ac.id

Abstrak

Kanker payudara adalah penyebab utama kematian akibat kanker tertinggi kedua di dunia. Pasien Kanker payudara terus mengalami peningkatan dan menjadi masalah kesehatan yang cukup serius di seluruh dunia, termasuk juga di Indonesia. Diagnosis dini adalah salah satu pendekatan terbaik untuk mencegah penyakit ini semakin meningkat dan berkembang. *Machine learning* dapat melakukan penambangan data menggunakan serangkaian fitur pada sebuah data. Penelitian ini menggunakan *dataset public* dari UCI *machine learning repository* yaitu *Breast Cancer Wisconsin* (*Diagnostic*). Pada *dataset* ini memiliki atribut sebanyak 32 fitur, namun banyaknya fitur pada sebuah data juga akan memperlambat waktu komputasi dari metode klasifikasi yang digunakan. Pada penelian ini, akan dilakukan seleksi fitur menggunakan metode *Hybrid Grey Wolf Optimization* dan *Particle Swarm Optimization* (HGWOPSO) untuk memilih fitur yang paling informatif dan signifikan untuk digunakan pada klasifikasi. Metode HGWOPSO ini digunakan untuk mengoptimasi penyeleksian fitur untuk melihat kinerjanya pada data yang digunakan. Metode klasifikasi yang digunakan adalah *Distance Biased Naive Bayes* (DBNB) yang terdiri dari dua modul yaitu *Weighted Naïve Bayes Module* (WNBM) dan *Distance Reinforcement Module* (DRM). Dari penelitian ini, didapatkan performa akurasi tertinggi pada model DBNB tanpa seleksi fitur sebesar 94,90%, DBNB dengan GWO sebesar 95,08%, DBNB dengan PSO sebesar 95,25%, dan DBNB dengan HGWOPSO sebesar 96,13%. Dapat disimpulkan bahwa model DBNB dengan seleksi fitur HGWOPSO mengalami peningkatan dibandingkan dengan DBNB tanpa seleksi fitur maupun dengan seleksi fitur individualnya.

Kata kunci: Kanker Payudara, Seleksi Fitur, *Hybrid, Grey Wolf Optimization, Particle Swarm Optimization, Distance Biased Naive Bayes*

1. Pendahuluan

Breast cancer atau kanker payudara adalah jenis kanker yang paling umum terjadi pada wanita di usia berapa pun dengan risiko yang meningkat seiring bertambahnya usia. Menurut WHO (*World Health Organization*) pada tahun 2020, terdapat 2,3 juta wanita terdiagnosis kanker payudara dengan 685.000 kematian secara global (World health Organization, 2021). Diagnosis dini kanker payudara adalah salah satu pendekatan terbaik untuk mencegah penyakit ini semakin meluas. Di beberapa negara maju, tingkat kelangsungan hidup pasien kanker payudara relatif 5 tahun di atas 80% karena pencegahan dini (Sun et al., 2017). Pemerintah Indonesia melalui peraturan Menteri Kesehatan nomor 34 tahun 2015, menjelaskan pemeriksaan payudara klinis (SADANIS) merupakan pilihan untuk skrining Kanker Payudara. Berbagai upaya telah dilakukan oleh pemerintah melalui Kementerian Kesehatan. Salah satu upaya preventif yang telah dilakukan adalah mengkampanyekan kepada perempuan untuk melakukan SADARI (Pemeriksaan Payudara Sendiri) dan SADANIS (Pemeriksaan Payudara Klinis) secara berkala agar dapat dilakukan tindakan secepatnya (Kementerian Kesehatan RI, 2019). Apabila pada pemeriksaan SADANIS terdapat benjolan diperlukan pemeriksaan lanjutan dengan USG maupun Mammografi. Hasil dari pemeriksaan lanjutan akan diidentifikasi apakah benjolan bersifat jinak atau ganas (kanker). Proses identifikasi akan lebih baik jika dilakukan menggunakan *machine learning*, sehingga proses identifikasi bisa lebih cepat dan akurat.

Pada *machine learning*, metode klasifikasi dapat membantu pakar dalam mengidentifikasi penyakit lebih akurat, efektif dan cepat untuk memberikan status. Dalam klasifikasi, *dataset* sangat berperan penting dalam merancang pemodelan klasifikasi penyakit yang efisien. Namun, terdapat masalah yang sering terjadi pada saat mengklasifikasikan sejumlah data yaitu fitur yang banyak, tidak relevan dan berlebihan sehingga dapat mengurangi kinerja dari algoritma klasifikasi (Xue et al., 2013). Oleh karena itu, pemilihan fitur merupakan langkah penting dalam

merancang model pengetahuan menggunakan algoritma *machine learning*.

Beberapa penelitian terdahulu akan dibahas berikut ini. Penelitian Singh, N dan Singh, S. B. (Singh, 2020), melakukan penelitian untuk meningkatkan kinerja konvergensi menggunakan 2 varian algoritma meta-heuristik yaitu *Particle Swarm Optimization* dan *Grey Wolf Optimizer*. Secara sederhana konsepnya adalah menggabungkan masing-masing kekuatan dari kedua varian algoritma dan bekerja secara paralel. Meningkatkan kemampuan eksploitasi di *Particle Swarm Optimization* dengan kemampuan eksplorasi di *Grey Wolf Optimizer* untuk menghasilkan kekuatan kedua varian. Dua puluh tiga *dataset* digunakan untuk menguji kualitas varian hybrid PSOGWO dibandingkan dengan PSO dan GWO. Solusi eksperimental membuktikan bahwa varian hybrid lebih andal dalam memberikan kualitas solusi yang unggul dengan iterasi komputasi yang wajar dibandingkan dengan PSO dan GWO.

Al-Tashi dkk. (Al-Tashi et al., 2019), melakukan penelitian untuk memecahkan masalah pemilihan fitur dengan mengusulkan versi biner dari optimasi *Grey Wolf Optimization* (GWO) dan *Particle Swarm Optimization* (PSO). Untuk mengevaluasi keefektifan dan efisiensi metode yang diusulkan, digunakan 18 set data benchmark UCI standar. Hasil perbandingan Mean, Best, dan Worst Fitness membuktikan kinerja yang jauh lebih baik dibandingkan dengan metode *state-of-art* lainnya. Metode BGWOPSO menunjukkan kemampuannya untuk mengontrol *trade-off* antara perilaku eksplorasi dan eksploitasi selama iterasi.

Penelitian Sundaramurthy & Jayavel (Sundaramurthy & Jayavel, 2020) mengatasi permasalahan data pasien Rheumatoid Arthritis dengan fitur yang berjumalah 20 atribut dan 1 kelas. Pendekatan seleksi fitur yang digunakan adalah wrapper-based feature selection. Sebelum dilakukan klasifikasi data dinormalisasikan ke dalam rentang [0-1]. Kemudian data yang sudah dilakukan seleksi fitur, diklasifikasikan menggunakan algoritma C4.5 dan dievaluasi menggunakan 10-fold cross validation. Evaluasi dilakukan dengan membandingkan kinerja dari model individu, hybrid dan pendekatan canggih lainnya. Pada klasifikasi data uji menggunakan benchmark *dataset* diperoleh nilai balanced accuracy (BACC) dari HGWO-C4.5, PSO-C4.5 dan GWO-C4.5 berturut-turut sebesar 86,36%, 81,25%, dan 80,42%. Sedangkan, klasifikasi data uji menggunakan independen *dataset* diperoleh nilai balanced accuracy (BACC) dari HGWO-C4.5, PSO-C4.5 dan GWO-C4.5 berturut-turut sebesar 84%, 80,42%, dan 80,16%. Kemudian varian HGWO dibandingkan dengan pendekatan canggih lainnya yaitu CSBoost dan REACT, hasilnya menunjukan balanced accuracy (BACC) dari HGWO 86,36%, CSBoost 75,16%, dan REACT 77,75%. Hasilnya menunjukkan bahwa varian HGWO untuk seleksi fitur mampu melakukan eliminasi fitur namun tetap memberikan nilai akurasi yang baik.

Penelitian El-Kenawy & Eid (El-Kenawy & Eid, 2020) menilai efektivitas dan kualitas dari varian hybrid GWOPSO untuk melakukan seleksi fitur menggunakan 17 *dataset* dari UCI machine learning repository. Masalah pada seleksi fitur sangat unik karena ruang pencarian terbatas pada dua nilai biner 0 dan 1. Pada penelitian ini digunakan persamaan sigmoid untuk memodifikasi metode optimasi agar berfungsi dengan baik untuk masalah ini. Pendekatan hybrid GWOPSO dimulai dengan populasi awal vektor acak untuk memilih fitur dari fungsi fitness, KNN digunakan untuk klasifikasi setelah data dilakukan seleksi fitur. Metode evaluasi yang dilakukan menggunakan evaluation metrics yaitu Classification Average Error, Best Fitness, Worst Fitness, Average Fitness size, Mean, dan Std (Standard Deviation). Berdasarkan perhitungan Average Error, kesalahan terendah dicapai oleh hybrid GWOPSO terbukti saat menjelajahi ruang pencarian yang besar. Perhitungan Mean menghasilkan bahwa dari 17 *dataset* hybrid GWOPSO mampu memilih fitur terbaik di 12 *dataset*. Perhitungan Average Fitness, hybrid GWOPSO mampu menemukan kualitas fitness terendah untuk semua *dataset* yang berarti dapat memilih subset fitur optimal yang menawarkan kesalahan klasifikasi terendah. Berdasarkan perhitungan Best Fitness dan Worst Fitness, varian hybrid GWOPSO mampu menemukan fungsi fitness terbaik dan tidak menemukan fitness terburuk dibandingkan dengan pengoptimal lainnya. Hybrid GWOPSO memiliki nilai Standard Deviation terendah yang menunjukkan kekokohan dan keandalan model hybrid GWOPSO yang diusulkan.

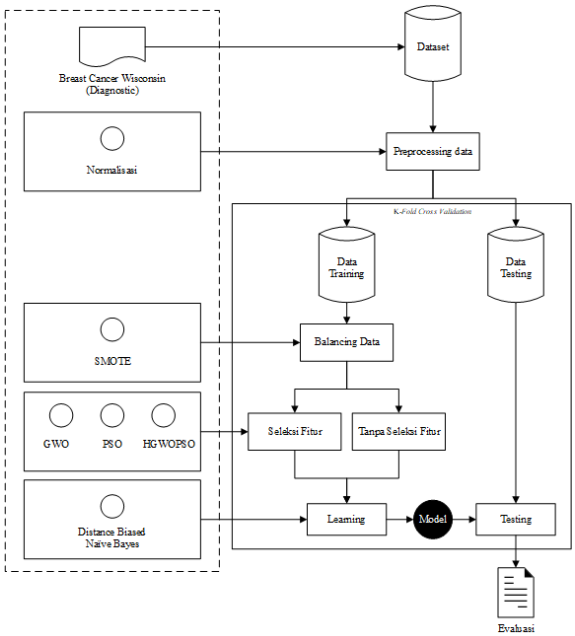
Shaban dkk. (Shaban et al., 2021), melakukan penelitian untuk mengatasi permasalahan pada algoritma *Naïve Bayes* tradisional pada data pasien diagnosis COVID-19. *Dataset* merepresentasikan catatan medis dari data yang dikumpulkan pada pasien dari Rumah Sakit Universitas Mansoura di Mesir. Pada penelitian ini terdapat dua fase, yaitu (i) Fase Seleksi Fitur dan (ii) Fase Klasifikasi. Pada fase pertama dilakukan seleksi fitur menggunakan algoritma APSO yang berbasis filter dan *wrapper*. Pada fase kedua menggunakan algoritma *Distance Biased Naïve Bayes* (DBNB) untuk mengatasi kelemahan NB tradisional dengan (i) memberikan bobot pada fitur yang dipilih, dan (ii) menyempurnakan keputusan dari WNB menggunakan bias berbasis jarak.Dihasilkan dua modul, yaitu *Weighted Naïve Bayes Module* (WBNM) dan (ii) *Distance Reinforcement Module* (DRM).

Oleh karena itu, pada penelitian ini dilakukan seleksi fitur menggunakan HGWOPSO dengan klasifikasi DBNB pada *dataset Breast Cancer Wisconsin* (*Diagnostic*). Penelitian ini juga akan membandingkan metode DBNB tanpa seleksi fitur

dan dengan seleksi fitur individual dari HGWOPSO agar dapat mengetahui perbandingan performa dari seleksi fitur *hybrid*.

2. Metode Penelitian

Kerangka kerja pada penelitian ini ditunjukkan pada Gambar 1. Variabel yang digunakan dalam penelitian ini yaitu pengujian berupa *confusion matrix (*akurasi, presisi, *recall, f1-score*), waktu komputasi (*time computation*) dan *Area Under the ROC Curve* (AUC) dari perbandingan tanpa seleksi fitur dan dengan seleksi GWO, PSO dan HGWOPSO pada *Distance Biased Naïve Bayes* untuk klasifikasi kanker payudara.



Gambar 1**.** Alur penelitian

2.1 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini merupakan *dataset Breast Cancer Wisconsin* (*Diagnostic*) dari UCI *machine learning repository*. *Dataset* ini memiliki 569 *instance* data dengan 32 atribut terdiri dari fitur id berisi id pasien, fitur diagnosis berisi label dengan 2 kelas yaitu *benign* (B) dan *malignant* (M), serta 30 fitur yang menggambarkan karakteristik inti sel yang ada dalam gambar digital *fine needle aspirate* (FNA) pada massa payudara. Fitur-fitur ini yang akan digunakan untuk mengklasifikasikan jenis kanker payudara pada pasien apakah termasuk kanker jinak (*benign*) atau kanker ganas (*maligant*).

2.2 Preprocessing Data

Pada tahap *preprocessing* data, *dataset* akan dinormalisasikan menggunakan metode *Min-Max Normalization* yang bertujuan untuk mengubah rentang nilai data pada setiap fitur agar berada pada rentang yang sama yaitu 0 sampai 1. Normalisasi data akan meminimalisir terjadinya noisedan relevansi yang rendah pada data, sehingga dapat memberikan kinerja yang baik pada proses klasifikasi (Gde Agung Brahmana Suryanegara et al., 2021).

Normalisasi data dilakukan karena range nilai *input* tidak sama. Nilai *input* akan diproses ke nilai *output* yang kecil sehingga data yang digunakan harus disesuaikan agar dapat diproses untuk mendapatkan nilai normalisasi yang kecil (Giusti et al., 2018). Persamaan 1 digunakan untuk mendapatkan nilai dari hasil normalisasi menggunakan *Min-Max Normalization* (Nishom, 2019).

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

Dimana x' merupakan data hasil normalisasi, x merupakan data yang akan dinormalisasi, $nilai_{min}$ merupakan nilai minimum data dan $nilai_{max}$ merupakan nilai maksimum data.

2.3 Pembagian Data

Pada tahap pembagian data, *dataset* akan dibagi menjadi data *training* (data latih) dan data *testing* (data uji). Pembagian data akan dilakukan menggunakan *K-Fold Cross Validation* menggunakan nilai k = 10 (Frank, 2005).

2.4 Penyeimbangan Data

Pada tahap ini akan dilakukan penyeimbangan data pada *dataset* yang telah dibagi sebelumnya. Pada penelitian ini, data akan diseimbangkan menggunakan metode *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE adalah strategi yang mengambil sampel objek kelas minoritas dengan menghasilkan sampel sintetis (Kumar et al., 2021). Tahap ini akan melalui beberapa skenario percobaan nilai k pada SMOTE menggunakan k=5 sampai dengan k=10 (Li et al., 2021). Skenario ini bertujuan untuk menemukan nilai k yang menghasilkan performa terbaik dan akan digunakan untuk tahap seleksi fitur dan klasifikasi.

Dataset dapat dikategorikan sebagai data tidak seimbang sesuai dengan nilai Imbalance Ratio (IR), yang didefinisikan sebagai proporsi jumlah sampel data di kelas mayoritas dengan minoritas. Persamaan ukuran IR ditunjukkan dengan menggunakan rumus berikut (Asniar et al., 2022):

$$IR = \frac{N^-}{N^+} \quad (2)$$

di mana N^- dan N^+ masing-masing adalah jumlah sampel dalam kelas mayoritas dan minoritas. Oleh karena itu, *dataset* tidak seimbang ketika $IR > 1$ (Asniar et al., 2022).

2.5 Seleksi Fitur

Tahap selanjutnya akan dilakukan seleksi fitur menggunakan *hybrid Grey Wolf Optimization dan Particle Swarm Optimization* (HGWOPSO). Singh, N & Singh, S.B. (Singh, 2020), menggunakan GWO dan PSO sebagai *hybrid* pada koevolusi tingkat rendah. Koevolusi ini menggabungkan fungsionalitas kedua varian bukan menggunakan kedua varian satu demi satu. Hybrid GWOPSO ini bekerja dengan meningkatkan kemampuan eksploitasi dari *Particle*

Swarm Optimization dan kemampuan eksplorasi dari *Grey Wolf Optimization* untuk menghasilkan kekuatan kedua varian.

Di HGWOPSO, posisi tiga agen pertama diperbarui di ruang pencarian dengan persamaan matematika. Fungsi fitness untuk kelompok-kelompok alpha (α), beta (β), dan delta (δ) dihitung secara terpisah dan nilai fitness dibandingkan. Perumusan matematis dari HGWOPSO diformulasikan sebagai berikut (Singh, 2020).

$$\begin{cases} \overrightarrow{d_\alpha} = \left|\overrightarrow{C_1} \cdot \overrightarrow{X_\alpha} - \omega\vec{x}\right| \\ \overrightarrow{d_\beta} = \left|\overrightarrow{C_2} \cdot \overrightarrow{X_\beta} - \omega\vec{x}\right| \\ \overrightarrow{d_\delta} = \left|\overrightarrow{C_3} \cdot \overrightarrow{X_\delta} - \omega\vec{x}\right| \end{cases} \quad (3)$$

dengan ω merupakan konstanta inersia, untuk menggabungkan varian PSO dan GWO *velocity* dan posisi baru direpresentasikan secara matematis sebagai berikut (Singh, 2020).

$$\begin{cases} v_i^{k+1} = \omega + c_1 r_1 (x_1 - x_i^k) + c_2 r_2 (x_2 - x_i^k) + c_3 r_3 (x_3 - x_i^k) \\ x_i^{k+1} = x_i^k + v_i^{k+1} \end{cases} \quad (4)$$

Masalah pemilihan fitur pada dasarnya bersifat bi-objektif. Satu tujuannya adalah untuk menemukan jumlah fitur minimum, dan memaksimalkan akurasi klasifikasi. Untuk mempertimbangkan keduanya, persamaan berikut digunakan sebagai fungsi fitness (Al-Tashi et al., 2019):

$$fitness = \propto \rho_R(D) + \beta \frac{|S|}{|T|} \quad (5)$$

dimana $\rho_R(D)$ merupakan nilai error rate dari model klasifikasi (NB tradisional), $|S|$ adalah jumlah fitur terpilih dan $|T|$ adalah jumlah seluruh fitur pada *dataset*.

Pada tahap ini akan dilakukan dengan beberapa skenario percobaan pada parameter populasi, kombinasi c1, c2, c3 dan iterasi dapat dilihat pada Tabel 1. Setiap skenario percobaan akan menghasilkan fitur-fitur yang dievaluasi menggunakan DBNB untuk menemukan kombinasi parameter terbaik.

Tabel 1. Parameter yang digunakan pada HGWOPSO

Parameter	**Nilai**
c1	[0,5 , 1]
c2	[0,5 , 1]
c3	[0,5 , 1]
ω	0,5 + rand()/2
populasi	[5, 15, 30, 50]
iterasi	[10, 50, 100, 250, 500]
α pada *fitness function*	0.99
β pada *fitness function*	0,01

2.6 Klasifikasi

Tahap klasifikasi akan dilakukan menggunakan *Distance Biased Naïve Bayes* (DBNB). DBNB merupakan metode yang yang diperkenalkan oleh Shaban dkk. (Shaban et al., 2021), untuk meningkatkan performa dan mengatasi kelemahan NB tradisional dengan (i) memberikan bobot pada fitur dan (ii) menyempurnakan keputusan WNB menggunakan bias berbasis jarak antara input dengan pusat kelas.

Teorema Bayes memiliki bentuk persamaan umum berbasis probabilitas sebagai berikut(Annur, 2018):

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (6)$$

Metode DBNB terdiri dari dua modul, modul pertama yaitu *Weighted Naïve Bayes Module* (WNBM) yang mana pengklasifikasi WNB digunakan untuk mengambil keputusan awal terkait tingkat kepemilikan input yang akan diklasifikasikan ke masing-masing kelas yang dipertimbangkan. Pada modul ini akan dilakukan perhitungan bobot fitur menggunakan persamaan berikut (Shaban et al., 2021):

$$w_y = \text{akurasi}(+f_y) - \text{akurasi}(-f_y) \quad (7)$$

dengan w_y merupakan bobot (dampak) dari fitur f_y, akurasi$(+f_y)$ merupakan akurasi dari model ketika fitur f_y dimasukkan dalam himpunan fitur, dan akurasi$(-f_y)$ merupakan akurasi dari model ketika fitur f_y dihapus. Bobot normal dari masing-masing fitur dihitung menggunakan persamaan 1:

Untuk dapat mengklasifikasikan data (sementara) pada modul ini dilakukan perhitungan *Belonging Score* (BS) dari input I_x ke kelas c_i dapat dihitung menggunakan (Shaban et al., 2021):

$$BS(I_x, c_i) = P(c_i) * \prod_{j=1}^{n} P(f_j|c_i)^{Nw_j} \text{ dengan } w_j \in \mathbb{R}^+ \quad (8)$$

dengan $BS(I_x, c_i)$ merupakan skor kepemilikan dari input I_x diberikan label kelas c_i. Kemudian, $P(c_i)$ merupakan peluang sebelumnya dari kelas c_i, Nw_j merupakan bobot normal dari fitur ke-j, dan $P(f_j|c_i)$ merupakan peluang bersyarat dari fitur f_j diberikan kelas c_i.

Pada modul kedua yaitu *Distance Reinforcement Module* (DRM) yang mana tingkat kepemilikan yang diperkirakan oleh WNBM disetel dengan baik untuk mengambil keputusan akhir. Untuk mengambil keputusan akhir, input harus diklasifikasikan ke salah satu kelas target. Untuk mencapai tujuan tersebut, awalnya, semua item (dari kelas target yang berbeda) diproyeksikan ke dalam ruang fitur n-dimensi yang dipertimbangkan. Pusat setiap kelas yang berisi t contoh dalam ruang fitur n-dimensi dapat diselesaikan menggunakan (Shaban et al., 2021):

$$C = \left\{ \frac{\sum_{q=1}^{t} V_q^1}{t}, \frac{\sum_{q=1}^{t} V_q^2}{t}, \dots, \frac{\sum_{q=1}^{t} V_q^n}{t} \right\} \quad (9)$$

dengan C merupakan pusat kelas dalam ruang fitur n-dimensi yang dipertimbangkan, t merupakan banyaknya contoh dalam kelas, dan V_q^i merupakan nilai dimensi ke-i dari contoh ke-q. Input yang akan diklasifikasikan (I_x) juga diproyeksikan pada ruang fitur n-dimensi. Kemudian, Derajat Afiliasi (*Affiliation Degree*) dari input diberikan masing-

masing kelas target ditentukan menggunakan (Shaban et al., 2021):

$$AD(I_x, c_i) = \frac{BS(I_x, c_i)}{\text{Dis}(I_x,\text{Center}\langle c_i\rangle)} \quad (10)$$

dengan $AD(I_x, c_i)$ merupakan derajat afiliasi dari input I_x diberikan kelas c_i, $BS(I_x, c_i)$ merupakan skor kepemilikan untuk input I_x diberikan label kelas c_i, $\text{Dis}(I_x,\text{Center}\langle c_i\rangle)$ merupakan jarak Euclidian di antara input I_x dan pusat kelas c_i dalam ruang fitur. Menghitung jarak antara dua titik p_x dan p_y dalam ruang fitur n-dimensi dapat dihitung menggunakan (Shaban et al., 2021):

$$\text{Dis}(p_x, p_y) = \sqrt{\sum_{i=1}^{n}\left(p_x^i - p_y^i\right)^2} \quad (11)$$

dengan p_x^i dan p_y^i masing-masing merupakan nilai dimensi ke-i dari titik p_x dan p_y dalam rung fitur n-dimensi. Terakhir kelas target dari input I_x, dinotasikan oleh $\text{Target}(I_x)$, dapat diperoleh menggunakan (Shaban et al., 2021):

$$\text{Target}(I_x) = \text{argmax}_{c_i \in C}\left[\frac{P(c_i) * \prod_{j=1}^{n} P(f_j|c_i)^{Nw_j}\ (\text{dengan } w_j \in \mathbb{R}^+)}{\text{Dis}(I_x,\text{Center}\langle c_i\rangle)}\right] \quad (12)$$

dengan $P(c_i)$ merupakan peluang sebelumnya dari kelas c_i, Nw_j merupakan bobot normal dari fitur ke-j, $P(f_j|c_i)$ merupakan peluang bersyarat dari fitur f_j diberikan kelas c_i.

2.7 Evaluasi

Tahap evaluasi akan dilakukan menggunakan *Confussion Matrix* dan akan di evaluasi berdasarkan nilai-nilai akurasi, presisi, recall, f1-score, dan AUC. Kemudian dilakukan perhitungan waktu komputasi (*time computation*) pada tahap seleksi fitur dan klasifikasi untuk melihat percobaan mana yang paling efisien.

3. Hasil dan Pembahasan

Pada penelitian ini terdapat beberapa skenario percobaan pada tahap penyeimbangan data dan pada tahap seleksi fitur. Skenario percobaan yang akan dilakukan pada tahap penyeimbangan data menggunakan SMOTE adalah percobaan pada nilai parameter k dan skenario percobaan pada tahap seleksi fitur baik seleksi fitur GWO, PSO ataupun HGWOPSO akan dilakukan percobaan pada nilai parameter populasi, kombinasi c1, c2, dan c3, serta nilai iterasi maksimal.

3.1 Skenario Penyeimbangan Data

Skenario percobaan ini dilakukan pada *dataset* yang telah dibagi menjadi data latih dan uji. Tahap penyeimbangan data hanya dilakukan pada data latih menggunakan SMOTE dengan skenario percobaan parameter k sebesar k = 5 sampai dengan k = 10 ditunjukan pada Tabel 2. Skenario ini bertujuan untuk menemukan parameter k optimal pada SMOTE yang dapat menghasilkan performa terbaik pada pengklasifikasi.

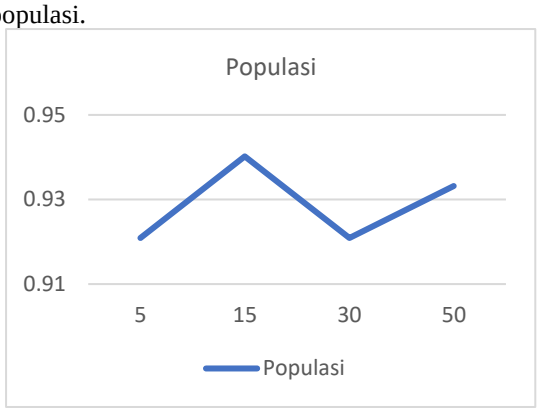
Tabel 2. Hasil pengujian parameter k pada SMOTE

k-SMOTE	Performa Klasifikasi			
	Tanpa seleksi fitur	BGWO	BPSO	BHGWOPSO
5	0,9279	0,9332	0,9279	0,9350
6	0,9315	0,9244	0,9315	0,9315
7	0,9490	0,9332	0,9227	0,9385
8	0,9350	0,9332	0,9367	0,9332
9	0,9279	0,9385	0,9350	0,9332
10	0,9262	0,9350	0,9192	0,9244

Dari **Error! Reference source not found.** diatas dapat dilihat bahwa akurasi tertinggi pada model DBNB tanpa seleksi fitur dihasilkan pada nilai k-SMOTE = 7 dengan nilai akurasi sebesar 94,90%, model BGWO-DBNB dihasilkan pada nilai k-SMOTE = 9 dengan nilai akurasi sebesar 93,85%, model BPSO-DBNB dihasilkan pada nilai k-SMOTE = 8 dengan nilai akurasi sebesar 93,67%, model BHGWOPSO-DBNB dihasilkan pada nilai k-SMOTE = 7 dengan nilai akurasi sebesar 93,85%. Nilai k dengan performa akurasi tertinggi pada setiap model akan digunakan pada tahap seleksi fitur dan klasifikasi.

3.2 Skenario Parameter Populasi pada Seleksi Fitur

Pada skenario ini, akan dilakukan percobaan pada parameter populasi untuk menentukan nilai parameter yang menghasilkan performa terbaik. Fitur yang dihasilkan pada setiap percobaan akan dihitung performanya menggunakan *Distance Biased Naïve Bayes* (DBNB). Pada Gambar 2 menunjukan performa yang dihasilkan dari setiap uji coba nilai populasi.



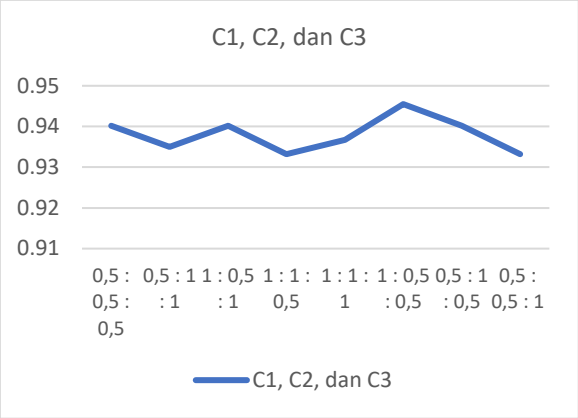
Gambar 2. Grafik Percobaan Parameter Populasi

Skenario ini menghasilkan kombinasi fitur terbaik pada nilai populasi = 15. Nilai populasi = 15 menghasilkan performa akurasi sebesar 94,02%, presisi sebesar 92,79%, *recall* sebesar 91,04%, *f1-score* sebesar 91,90%, AUC sebesar 93,42% dan rata-rata waktu komputasi sebesar 2,0434 detik. Nilai parameter populasi terbaik ini akan digunakan pada skenario percobaan berikutnya.

3.3 Skenario Parameter C1, C2, dan C3 pada Seleksi Fitur

Pada skenario ini, akan dilakukan percobaan pada kombinasi parameter c1, c2 dan c3 untuk menentukan nilai parameter yang menghasilkan

performa terbaik. Fitur yang dihasilkan pada setiap percobaan akan dihitung performanya menggunakan *Distance Biased Naïve Bayes* (DBNB). Pada Gambar 3 menunjukan performa yang dihasilkan dari setiap uji coba nilai c1, c2 dan c3.

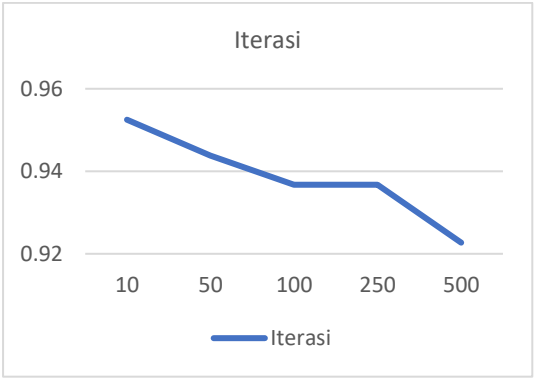


Gambar 3. Grafik Percobaan Parameter c1, c2 dan c3

Skenario ini menghasilkan kombinasi fitur terbaik pada nilai kombinasi c1 = 1, c2 = 0,5 dan c3 = 0,5. Nilai c1 = 1, c2 = 0,5 dan c3 = 0,5 menghasilkan performa akurasi sebesar 94,55%, presisi sebesar 93,30%, *recall* sebesar 91,98%, *f1-score* sebesar 92,64%, AUC sebesar 94,03% dan rata-rata waktu komputasi sebesar 1,4121 detik. Nilai kombinasi parameter c1, c2 dan c3 terbaik ini akan digunakan pada skenario percobaan berikutnya.

3.4 Skenario Parameter Iterasi pada Seleksi Fitur

Skenario percobaan ini merupakan skenario terakhir pada tahap seleksi fitur. Percobaan ini akan melakukan pengujian pada parameter iterasi untuk menentukan nilai parameter yang menghasilkan performa terbaik. Fitur yang dihasilkan pada setiap percobaan akan dihitung performanya menggunakan *Distance Biased Naïve Bayes* (DBNB). Fitur terpilih dengan akurasi tertinggi pada skenario ini akan digunakan pada tahap klasifikasi. Pada Gambar 4 menunjukan performa yang dihasilkan dari setiap uji coba nilai iterasi.



Gambar 4. Grafik Percobaan Parameter Iterasi

Skenario ini menghasilkan kombinasi fitur terbaik pada nilai iterasi = 10. Nilai iterasi = 10 menghasilkan performa akurasi sebesar 95,25%, presisi sebesar 94,69%, *recall* sebesar 92,45%, *f1-score* sebesar 93,56%, AUC sebesar 94,69% dan rata-rata waktu komputasi sebesar 0,6127 detik.

Ditampilkan pada tabel 3 percobaan parameter juga dilakukan pada seleksi fitur individualnya yaitu GWO dan PSO. Observasi parameter pada metode seleksi fitur GWO menghasilkan parameter terbaik pada nilai *agent* = 30 dan iterasi = 50 dengan performa akurasi sebesar 94,55%, presisi sebesar 94,15%, *recall* sebesar 91,04%, *f1-score* sebesar 92,57%, AUC sebesar 93,84% dan rata-rata waktu komputasi sebesar 1,0572 detik. Seleksi fitur PSO menghasilkan parameter terbaik pada nilai partikel = 15, c1 = 1, c2 = 2 dan iterasi = 100 dengan performa akurasi sebesar 94,73%, presisi sebesar 94,17%, *recall* sebesar 91,51%, *f1-score* sebesar 92,82%, AUC sebesar 94,90% dan rata-rata waktu komputasi sebesar 1,5380 detik. Fitur terpilih dari setiap metode seleksi fitur akan digunakan pada klasifikasi.

3.5 Klasifikasi

Setelah mendapatkan subset fitur terpilih pada skenario percobaan terakhir pada tahap seleksi fitur, tahap selanjutnya akan dilakukan klasifikasi menggunakan model DBNB. Pada tahap ini juga akan dilakukan perbandingan dengan model WNBM tanpa seleksi fitur serta dengan seleksi fitur individual dari HGWOPSO seperti pada Tabel 3 dan pada

Tabel 4 untuk model DBNB.

Tabel 3. Perbandingan Performa Klasifikasi menggunakan WNBM

Metode	Akurasi	Presisi	*Recall*	*F1-Score*	AUC	Waktu Komputasi Klasifikasi (detik)
WNBM (tanpa seleksi fitur)	0,9473	0,9417	0,9151	0,9282	0,9407	1,6754
BGWO-WNBM	0,9490	0,9463	0,9151	0,9305	0,9421	0,7095
BPSO-WNBM	0,9473	0,9292	0,9292	0,9292	0,9436	0,2874
BHGWOPSO-WNBM	0,9525	0,9469	0,9245	0,9356	0,9469	0,7110

Tabel 4. Perbandingan Performa Klasifikasi menggunakan DBNB

Metode	Akurasi	Presisi	*Recall*	*F1-Score*	AUC	Waktu Komputasi Klasifikasi (detik)
DBNB (tanpa seleksi fitur)	0,9490	0,9463	0,9151	0,9305	0,9421	1,8133
BGWO-DBNB	0,9508	0,9510	0,9151	0,9327	0,9435	0,7689
BPSO-DBNB	0,9525	0,9384	0,9340	0,9362	0,9488	0,3128
BHGWOPSO-DBNB	0,9613	0,9612	0,9340	0,9474	0,9558	0,7703

Pada klasifikasi modul pertama WNBM, model BHGWOPSO-WNBM menghasilkan performa tertinggi dibandingkan dengan model lainnya dengan nilai akurasi sebesar 95,25%, presisi sebesar 94,69%, recall sebesar 92,45%, f1-score sebesar 93,56%, AUC sebesar 94,69% dan rata-rata waktu komputasi sebesar 0,7110 detik. Sedangkan pada klasifikasi modul kedua yaitu DRM atau bisa juga disebut DBNB, model BHGWOPSO-DBNB menghasilkan performa tertinggi dibandingkan dengan model lainnya dengan nilai akurasi sebesar 96,13%, presisi sebesar 96,12%, *recall* sebesar 93,40%, *f1-score* sebesar 94,74%, AUC sebesar 95,58% dan rata-rata waktu komputasi sebesar 0,7703 detik.

4. Kesimpulan dan Saran

Berdasarkan penelitian yang telah dilakukan, seleksi fitur pada metode klasifikasi DBNB mampu meningkatkan akurasi, presisi, *recall*, *f1-score*, dan AUC dibandingkan DBNB tanpa seleksi fitur. Model BGWO-DBNB dibandingkan dengan DBNB, mengalami peningkatan akurasi sebesar 0,18%, peningkatan presisi sebesar 0,47%, peningkatan *f1-score* sebesar 0,22%, dan peningkatan AUC sebesar 0,14%, namun pada nilai evaluasi *recall* tidak mengalami perubahan baik peningkatan ataupun penurunan. Model BPSO-DBNB dibandingkan dengan DBNB, mengalami peningkatan akurasi sebesar 0,35%, peningkatan *recall* sebesar 1,89%, peningkatan *f1-score* sebesar 0,57%, dan peningkatan AUC sebesar 0,67%, namun terjadi penurunan presisi sebesar 0,79%. Model BHGWOPSO-DBNB dibandingkan dengan DBNB, mengalami peningkatan akurasi sebesar 1,23%, peningkatan presisi sebesar 1,49%, peningkatan *recall* sebesar 1,89%, peningkatan *f1-score* sebesar 1,69%, dan peningkatan AUC sebesar 1,37%. Berdasarkan Performa akurasi maka model BHGWOPSO-DBNB lebih baik dibandingkan model lainnya, walaupun dari segi waktu komputasi model BPSO-DBNB lebih unggul. Peningkatan akurasi dari setiap uji coba metode seleksi fitur yang digunakan memang tidak terlalu siginifikan dikarenakan data yang digunakan sudah sangat layak digunakan tanpa dilakukannya seleksi fitur. Penelitian selanjutnya dapat mencoba menggunakan *dataset* lain untuk membandingkan kinerja metode dari hasil penelitian ini.

Klasifikasi DBNB mampu meningkatkan hasil klasifikasi sementara dari WNBM pada model dengan seleksi fitur, dimana tingkat kepemilikan yang diperkirakan oleh WNBM disetel lebih baik lagi untuk pengambilan keputusan akhir. Pada model BGWO-DBNB dibandingkan dengan BGWO-WNBM, mengalami peningkatan akurasi sebesar 0,18%. Model BPSO-DBNB dibandingkan dengan BPSO-WNBM, mengalami peningkatan akurasi sebesar 0,52. Model BHGWOPSO-DBNB dibandingkan dengan BHGWOPSO-WNBM, mengalami peningkatan akurasi sebesar 0,88%. Seleksi fitur juga mampu meningkatkan kecepatan waktu komputasi dari model BGWO-DBNB, BPSO-DBNB, dan BHGWOPSO-DBNB berturut-turut sebesar 0,7689 detik, 0,3128 detik, dan 0,7703 detik. Berdasarkan Performa akurasi maka model BHGWOPSO-DBNB lebih baik dibandingkan model lainnya, walaupun dari segi waktu komputasi model BPSO-DBNB lebih unggul.

Pada penelitian selanjutnya juga disarankan untuk menggunakan metode seleksi fitur lain dari pendekatan *wrapper*, *filter* ataupun *embedded*, untuk mengetahui bagaimana kinerja dan perbandingan akurasi dari metode seleksi fitur lain pada metode klasifikasi DBNB.

Daftar Pustaka :

Al-Tashi, Q., Abdul Kadir, S. J., Rais, H. M., Mirjalili, S., & Alhussian, H. (2019). Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection. *IEEE Access*, *7*, 39496–39508. https://doi.org/10.1109/ACCESS.2019.2906757

Al-Tashi, Q., Abdul Kadir, S. J., Rais, H. M., Mirjalili, S., & Alhussian, H. (2019). Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection. *IEEE Access*, *7*, 39496–39508. https://doi.org/10.1109/ACCESS.2019.2906757

Annur, H. (2018). KLASIFIKASI MASYARAKAT MISKIN MENGGUNAKAN METODE NAÏVE BAYES. In *Agustus* (Vol. 10, Issue 2).

Asniar, Maulidevi, N. U., & Surendro, K. (2022). SMOTE-LOF for noise identification in imbalanced data classification. *Journal of King Saud University - Computer and Information Sciences*, *34*(6), 3413–3423. https://doi.org/10.1016/j.jksuci.2021.01.014

El-Kenawy, E. S., & Eid, M. (2020). Hybrid gray wolf and particle swarm optimization for feature selection. *International Journal of Innovative Computing, Information and Control*, *16*(3), 831–844. https://doi.org/10.24507/ijicic.16.03.831

Frank, I. H. W. & E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques - 2nd ed*.

Gde Agung Brahmana Suryanegara, Adiwijaya, & Mahendra Dwifebri Purbolaksono. (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. https://doi.org/10.29207/resti.v5i1.2880

Giusti, A., Widodo, A. W., & Adinugroho, S. (2018). *Prediksi Penjualan Mi Menggunakan Metode Extreme Learning Machine (ELM) di Kober Mie Setan Cabang Soekarno Hatta* (Vol. 2, Issue 8). http://j-ptiik.ub.ac.id

Kementerian Kesehatan RI. (2019). *BEBAN KANKER DI INDONESIA*.

Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. (2021). Classification of Imbalanced Data:Review of Methods and Applications. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012077. https://doi.org/10.1088/1757-899x/1099/1/012077

Li, J., Zhu, Q., Wu, Q., & Fan, Z. (2021). A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. *Information Sciences*, *565*, 438–455. https://doi.org/10.1016/j.ins.2021.03.041

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

Shaban, W. M., Rabie, A. H., Saleh, A. I., & Abo-Elsoud, M. A. (2021). Accurate detection of COVID-19 patients based on distance biased Naïve Bayes (DBNB) classification strategy. *Pattern Recognition*, *119*, 108110. https://doi.org/10.1016/j.patcog.2021.108110

Singh, N. S. and S. B. (2020). Hybrid Algorithm of Particle Swarm Optimization and Grey Wolf Optimizer for Reservoir Operation Management. *Water Resources Management*, *34*(15), 4545–4560. https://doi.org/10.1007/s11269-020-02656-8

Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., Shi, W., Jiang, J., Yao, P. P., & Zhu, H. P. (2017). Risk factors and preventions of breast cancer. *International Journal of Biological Sciences*, *13*(11), 1387–1397. https://doi.org/10.7150/ijbs.21635

Sundaramurthy, S., & Jayavel, P. (2020). A hybrid Grey Wolf Optimization and Particle Swarm Optimization with C4.5 approach for prediction of Rheumatoid Arthritis. *Applied Soft Computing Journal*, *94*. https://doi.org/10.1016/j.asoc.2020.106500

World health Organization. (2021). *Breast cancer*. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Xue, B., Zhang, M., & Browne, W. N. (2013). Particle swarm optimization for feature selection in classification: A multi-objective approach. *IEEE Transactions on Cybernetics*, *43*(6), 1656–1671. https://doi.org/10.1109/TSMCB.2012.2227469