Optimasi Algoritma K-*Nearest Neighbor* (k-NN) dengan *Wrapper Forward Selection* untuk Deteksi Penderita *Breast Cancer*

Oei Joviano Matthew Wijaya[1], Siska Narulita[2*]

[1,2]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi,
Universitas Nasional Karangturi, Semarang, Indonesia
Email: [2]siskanarulita84@gmail.com

ABSTRACT

Breast cancer is a disorder in which abnormal body cells proliferate and replace healthy cells in the breast area. Breast cancer requires prompt treatment to prevent cancer cells from spreading widely, as the disease can be lethal. Data mining could be one of the options to aid with breast cancer diagnosis. Data mining can improve decision-making by allowing processed data to be analyzed before reaching a decision. The data analysis in this work makes use of the k-Nearest Neighbor (k-NN) classification algorithm, which has been refined using a feature selection technique called as wrapper forward selection. The study's findings show that data mining is extremely useful and advantageous in the analysis and diagnosis of breast cancer. The research results show that the percentage values of accuracy, precision, and recall in the model using forward selection are higher than those that do not use forward selection, with 96.19%. Meanwhile, the model without a forward selection approach achieves an accuracy of 84.16%. As a result, in this study, the forward selection technique has a significant impact on the accuracy of the produced model.

Keywords*:Data Mining; Breast Cancer; Feature Selection; Wrapper; Forward Selection*
.

ABSTRAK

Breast cancer atau kanker payudara adalah penyakit yang disebabkan karena adanya pertumbuhan sel-sel tubuh yang tidak normal dan mengambil alih sel yang masih sehat pada daerah payudara. *Breast cancer* sangat diperlukan penanganan dini agar sel kanker pada payudara tidak menyebar secara luas, karena *breast cancer* dapat menyebabkan kematian. *Data mining* dapat menjadi salah satu opsi solusi dalam membantu diagnosis kanker payudara. *Data mining* dapat berperan dalam membantu pengambilan keputusan, karena data yang sudah diolah dapat digunakan dalam analisis sebelum pengambilan keputusan. Analisis data dalam penelitian ini menggunakan algoritma klasifikasi k-*Nearest Neighbor* (k-NN) yang dioptimasi menggunakan teknik *feature selection*, yaitu *wrapper forward selection*. Hasil penelitian menunjukkan bahwa *data mining* sangat berguna dan bermanfaat dalam menganalisis dan mendiagnosis penyakit *breast cancer*. Hasil penelitian menunjukkan bahwa nilai persentase akurasi, presisi, dan *recall* pada model yang menggunakan *forward selection* menghasilkan persentase yang lebih tinggi daripada yang tidak menggunakan *forward selection*, yaitu sebesar 96,19%. Sedangkan model yang tidak menggunakan teknik *forward selection* menghasilkan tingkat akurasi sebesar 84,16%. Sehingga dalam penelitian ini, teknik *forward selection* sangat berpengaruh dalam meningkatkan akurasi pada model yang terbentuk.

Katakunci: *Data Mining; Kanker Payudara; Feature Selection; Wrapper; Forward Selection*

PENDAHULUAN

Kanker payudara atau *breast cancer*adalah perkembangan sel yang mempunyai karakteristikyang tidak dapat dikendalikan pada jaringan payudara terutama pada kelenjar penghasil susu, *ductus*, pembuluh limfa, dan tidak termasuk pada kulit payudara.(1)

Kanker payudara pada umumnya dapat diklasifikasikan menjadi dua, yaitu *benign*(jinak) dan *malignant*(ganas), pada kanker payudara *benign* biasanya ditandai dengan adanya benjolan kecil, sedangkan pada kanker payudara *malignant* biasanya ditandai dengan bentuk yang tidak simetris, terasa nyeri, dan lainnya.[2]

Penyakit kanker menempati peringkat kedua penyakit terbesar di dunia.[3] Pada tahun 2022, *International Agency for Research on Cancer* bersama*World Health Organization* (WHO) menyatakan sebanyak 2.296.607 jiwa di seluruh dunia mengidap penyakit kanker pada rentang usia 15 tahun hingga 85 tahun.[4] Indonesia memiliki kasus kanker payudara yang terus meningkat dan menjadi ancaman kesehatan yang cukup serius.[5]

Kanker payudara baik yang bersifat jinak maupun ganas harus segera ditangani dan diobati, hal ini dikarenakan apabila kanker payudara tidak segera ditangani akan berakibat buruk bagi penderita dan bahkan dapat menyebabkan kematian. Perlunya deteksi sejak dini, sehingga penderita dapat melakukan pemeriksaan dan memperoleh penanganan untuk keselamatan penderita.[6]

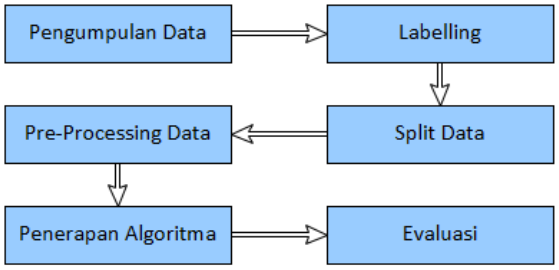
Salah satu metode yang dapat digunakan untuk melakukan prediksi atau deteksi terhadap suatu penyakit menggunakan data-data yang sudah ada adalah *data mining*. *Data mining* merupakan proses penggalian pengetahuan dari sekumpulan data sebagai upaya ilmiah masa lalu.[7] Penelitian ini memanfaatkan metode *data mining* untuk proses peninjauan atau diagnosis pada *breast cancer*. Algoritma *data mining* yang digunakan pada penelitian ini adalah k-NN. k-NN adalah algoritma generalisasi aturan tetangga terdekat. *Offset* induktifnya adalah label kelas k dengan label kelas yang akan diuji kemiripannya dengan tetangga terdekat. Algortima ini memperluas kawasan tetangga terdekatnya dalam fase pengambilan keputusan. Perluasan ini menghasilkan algoritma k-NN dalam mengambil banyak informasi.[8] Beberapa penelitian terdahulu yang menggunakan algoritma k-NN untuk deteksi *breast cancer*, antara lain dilakukan oleh Henderi *et al*., (2021) dalam penelitiannya yang berjudul *Comparison of Min-Max Normalization and Z-Score Normalization in the K-Nearest Neighbor* (kNN) *Algorithm to Test the Accuracy of Types of Breast Cancer*.[9] Dalam penelitiannya, dengannilaik = 5 dan k = 21 mendapatkan akurasi 98% dan hasil terendah adalah dengan nilai k = 1, k = 7, k = 9, dan k = 27 dengan akurasi 96%.[9] Penelitian kedua oleh Enriko *et al*., (2021) berjudul *Breast Cancer Recurrence Prediction System using k-Nearest Neighbor, Naive-Bayes, and Support Vector Machine Algorithm*. Dalam penelitiannya tersebut mendapatkan hasil dari nilai k= 1 sampai 9 hasil akurasi terbaik adalah pada k = 7 karena memiliki hasil akurasi 77,98%.[10] Penelitian ketiga oleh Atthalla*etal*., (2018) berjudul Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K-*NearestNeighbors* (KNN). Dalam penelitiannya mendapat hasil penelitian dengan nilaik = 3 menghasilkan akurasi tertinggi sebesar 93%.[11] Penelitian keempat oleh Ganjar *et al*., (2022) berjudul *Predicting Breast Cancer from Risk Factors using SVM and Extra-Trees-Based Feature Selection Method*. Dalam penelitiannya mendapatkan hasil bahwa model yang digunakan dalam penelitiannya menggunakan penggabungan SVM dan *extra-tree* mendapatkan nilai akurasi sebesar 80,23%.[12] Penelitian kelima oleh Mahesh *et al*., (2022) berjudul *An Efficient Ensemble Method using K-Fold Cross Validation for the Early Detection of Benign and Malignant Breast Cancer*. Dalam penelitiannya mendapatkan hasil penelitian *true positive* dengan hasil 36,50%, *true negative*dengan hasil 61,31%, *false positives* dengan hasil "*type* I *error*", *false negative* dengan "*type* II *mistake*".[13] Dari beberapa penelitian terdahulu tersebut di atas, dapat disimpulkan bahwa penelitian untuk memprediksi *breast cancer* menggunakan *data mining* masih dilakukan, meskipun menggunakan beberapa algoritma *data mining* yang berbeda. Namun, dalam penelitian terdahulu tersebut, penggunaan algoritma *data mining* k-NN pada deteksi *breast cancer* mempunyai hasil akurasi yang cukup bagus.

Berdasarkan latar belakang di atas, maka dalam penelitian ini, peneliti melakukan deteksi atau prediksi penderita *breast cancer* menggunakan algoritma k-NN. Untuk mengoptimalkan algoritma k-NN yang digunakan, peneliti menambahkan *wrapper feature selection* pada *pre-processing* data. Metode evaluasi yang digunakan untuk menguji performansi algoritma k-NN adalah *confusion matrix*. Hasil dari penelitian ini diharapkan dapat digunakan untuk meningkatkan diagnosis dari penyakit *breast cancer*.

METODE

Penelitian ini dilakukan dalam beberapa tahap yaitu, pengumpulan data, *labeling*, *split*

data, *pre-processing* data, penerapan algoritma, dan evaluasi. Gambar 1 berikut merupakan alur dari metode penelitian ini:



Gambar 1. Metode Penelitian

1. Pengumpulan Data

   Pada penelitian ini, *dataset* menggunakan data *public* yang diambil dari *platform* terbuka/*open source*, yaitu UC *Irvine* (UCI) *Machine Learning Repository* (https://archive.ics.uci.edu). *Dataset* yang digunakan adalah *breast cancer dataset*.

2. *Labelling*

   Pada tahap *labelling* dilakukan penentuan *class*/label pada *dataset breast cancer* untuk proses klasifikasi.

3. *Split* Data

   Split data pada penelitian ini bertujuan untuk membagi *data training* dan *data testing* dengan rasio 60% banding 40%.

4. *Pre-Processing* Data

   Pre-processing data pada penelitian ini dilakukan dengan menerapkan *feature selection* untuk mengoptimasi kinerja algoritma *machine learning*. Metode *feature selection* yang digunakan adalah *wrapper forward selection*.

5. Penerapan Algoritma

   Algoritma *data mining* yang diterapkan pada penelitian ini adalah algoritma k-NN, yaitu algortima generalisasi aturan tetangga terdekat. *Offset* induktifnya adalah label *class* k dengan label *class* yang akan diuji kemiripannya dengan tetangga terdekat. Algortima ini memperluas kawasan tetangga terdekatnya dalam fase pengambilan keputusan. Perluasan ini menghasilkan algoritma k-NN dalam mengambil banyak informasi.[8]

6. Evaluasi

   Tahap evaluasi digunakan untuk mengukur performa model yang dihasilkan dari penelitian. Di sini, tahap evaluasi digunakan untuk melihat tingkat akurasi dari algoritma k-NN.[14] Pada tahap evaluasi ini digunakan *confusion matrix*, yaitu *tools* untuk mengevaluasi model dan menampilkan hasil evaluasi dalam bentuk table.[14] Beberapa indikator yang dapat diketahui pada *confusion matrix*, yaitu:[15]

   a. *Accuracy*, merupakan metrik evaluasi yang mengukur tingkat prediksi yang benar (*true*) dari keseluruhan prediksi yang sudah ada di model.[16]
   b. *Precision*, merupakan gambaran dari tingkat akurasi antara data yang diminta dengan data hasil prediksi yang diberikan oleh model.[17]
   c. *Recall*, merupakan gambaran keberhasilan model dalam mencari sebuah Informasi.[17]

HASIL

Pengumpulan Data

Hasil dari tahap pengumpulan data yaitu *dataset breast cancer* yang mempunyai 31 atribut seperti ditunjukkan pada Tabel 1, yaitu:

Tabel 1. Atribut

No.	Atribut	No.	Atribut
1.	*diagnosis*	17.	*compactness_se*
2.	*radius_mean*	18.	*concavity_se*
3.	*texture_mean*	19.	*concave_points_se*
4.	*perimeter_mean*	20.	*symmetry_se*
5.	*area_mean*	21.	*fractal_dimension_se*
6.	*smoothness_mean*	22.	*radius_worst*
7.	*compactness_mean*	23.	*texture_worst*
8.	*concavity_mean*	24.	*perimeter_worst*
9.	*concave_points_mean*	25.	*area_worst*
10.	*symmetry_mean*	26.	*smoothness_worst*
11.	*fractal_dimension_mean*	27.	*compactness_worst*
12.	*radius_se*	28.	*concavity_worst*
13.	*texture_se*	29.	*concave_points_worst*
14.	*perimeter_se*	30.	*symmetry_worst*
15.	*area_se*	31.	*fractal_dimension_worst*
16.	*smoothness_se*		

Dataset ini mempunyai 569 *record* data dan tidak mempunyai *missing value*.

Labelling

Peneliti menentukan *class* (label) pada *dataset breast cancer*, yaitu *diagnosis_result*.

Split Data

Hasil dari proses *split* data dengan rasio *data training*:*data testing*, yaitu sebesar 60%:40% diperoleh jumlah *record* data pada *data training*

sebesar 341, sedangkan pada *data testing* sebesar 228*record* data.

***Pre-Processing* Data**

Pada tahap *pre-processing* data, dilakukan proses *feature selection* untuk mengoptimalkan algoritma k-NN menggunakan teknik *wrapper forward selection*. Hasil dari tahap ini, terjadi reduksi jumlah atribut yang semula berjumlah 31 menjadi 4 atribut, yaitu *diagnosis_result*, *perimeter_worst*, *texture_worst*, dan *radius_mean*.

Penerapan Algoritma

Pada penelitian ini, algoritma k-NN diimplementasikan sebelum dan setelah *dataset* melalui *pre-processing* data (teknik *wrapper forward selection*). Tabel 2 berikut ini menunjukkan hasil *accuracy* algoritma k-NN sebelum penggunaan teknik *wrapper forward selection*.

Tabel 2. *Confusion Matrix* (*Accuracy*: 84,16%)

	***True* M**	***True* B**	***Class Precision***
Pred. M	83	10	89,25%
Pred. B	44	204	82,26%
Class Recall	65,35%	95,33%	

Tabel 3 menunjukkan nilai *precision* dari algoritma k-NN.

Tabel 3. *Confusion Matrix* (*Precision*: 82,26% | *Positive Class*: B)

	***True* M**	***True* B**	***Class Precision***
Pred. M	83	10	89,25%
Pred. B	44	204	82,26%
Class Recall	65,35%	95,33%	

Sedangkan Tabel 4 menunjukkan nilai *recall* dari algoritma k-NN.

Tabel 4. *Confusion Matrix* (*Recall*: 95,33% | *Positive Class*: B)

	***True* M**	***True* B**	***Class Precision***
Pred. M	83	10	89,25%
Pred. B	44	204	82,26%
Class Recall	65,35%	95,33%	

Evaluasi

Setelah implementasi teknik *wrapper forward selection*, hasil *accuracy* algoritma k-NN ditunjukkan pada Tabel 5.

Tabel 5. *Confusion Matrix* + *Forward Selection* (*Accuracy*: 96,19%)

	***True* M**	***True* B**	***Class Precision***
Pred. M	118	4	96,72%
Pred. B	9	210	95,89%
Class Recall	92,91%	98,13%	

Tabel 6 menunjukkan nilai *precision* dari algoritma k-NN dengan penambahan teknik *wrapper forward selection*.

Tabel 6. *Confusion Matrix* + *Forward Selection* (*Precision*: 95,89% | *Positive Class*: B)

	***True* M**	***True* B**	***Class Precision***
Pred. M	118	4	96,72%
Pred. B	9	210	95,89%
Class Recall	92,91%	98,13%	

Sedangkan Tabel 7 menunjukkan nilai *recall* dari algoritma k-NN dengan penambahan teknik *wrapper forward selection*.

Tabel 7. *Confusion Matrix* + *Forward Selection* (*Recall*: 98,13% | *Positive Class*: B)

	***True* M**	***True* B**	***Class Precision***
Pred. M	118	4	96,72%
Pred. B	9	210	95,89%
Class Recall	92,91%	98,13%	

PEMBAHASAN

Setelah dilakukan uji coba,terdapat perbedaan nilai *accuracy*, *precision*, dan *recall*pada implementasi algoritma k-NN sebelum dan sesudahpenambahan *feature selection wrapper forward selection*. Perbandingan hasilnya ditunjukkan pada Tabel 8 berikut ini:

Tabel 8. Perbandingan Model Sebelum dan Sesudah Penambahan *Wrapper Forward Selection*

Parameter	**Tanpa *Forward Selection***	**Dengan *Forward Selection***
Accuracy	84,16%	96,19%
Precision	82,26%	95,89%
Recall	95,33%	98,13%

Pada Tabel 8 dapat dilihat bahwa adanya perbandingan dan perbedaan hasil dalam model yang menggunakan *forward selection*dengan yang tanpa menggunakan *forward selection*. Hal ini dapat terjadi dikarenakan adanya seleksi fitur (*feature selection*). *Feature selection* dalam hal

ini memilih atribut yang relevan untuk digunakan dan menghilangkan atribut yang tidak terlalu berpengaruh. Oleh karena itu, seleksi fitur berperan sangat besar pada peningkatan dan memaksimalkan hasil perhitungan pada *data mining*, sehingga *forward selection* dapat membantu dalam meningkatkan kinerja model.[18] Model yang terbentuk menggunakan *wrapper forward selection* mempunyai nilai *accuracy* yang lebih tinggi dibandingkan dengan model tanpa *forward selection*. Namun, pada kasus ini, nilai *precision* yang lebih rendah justru yang lebih bagus (model tanpa *forward selection*). Hal ini dikarenakan persentase yang benar-benar penderita *breast cancer* dibandingkan dengan keseluruhan prediksi penderita lebih kecil. Pada kasus ini, nilai *recall* lebih baik yang bernilai tinggi, karena lebih baik algoritma memprediksi penderita *breast cancer*, tetapi sebenarnya tidak terkena breast cancer, daripada algoritma salah memprediksi sebaliknya.

SIMPULAN

Hasil penelitian menunjukkan bahwa penerapan algoritma*data mining*k-NN dan *feature selection wrapper forward selection* sangat berguna dan bermanfaat untuk membantu prediksi dan diagnosis pada penyakit *breast cancer*. Analisis pada penelitian penderita *breast cance*r menggunakan algoritma k-NN dengan metode *forward selection* menghasilkan beberapa kesimpulan, yaitu penggunaan *forward selection* dapat membantu dalam melakukan seleksi fitur dan meningkatkan kinerja model yang lebih akurat. Hasil dari penelitian ini menunjukkan bahwa penambahan*forward selection* sangat berpengaruh.Penggunaannya dalam penelitian ini menyebabkan adanya selisihyang cukup besar pada nilai akurasi antara implementasi algoritma k-NN sebelum dan sesudah penambahan *forward selection*, yaitu sebesar 9,86%. Sehingga dapat disimpulkan bahwa penggunaan *forward selection* pada algoritma *data mining*k-NN sangat bermanfaat dan dapat membantu dalam melakukan prediksi pada penderita *breast cancer*.

DAFTARPUSTAKA

1. Putra SR. Buku Lengkap Kanker Payudara. Yogyakarta: Laksana; 2015.
2. Rejani YIA, Selvi ST. Early Detection of Breast Cancer using SVM Classifier Technique. IJCSE Int J Comput Sci Eng. 2009;1(3):127–30.
3. Chimed T, Sandagdorj T, Znaor A, Laversanne M, Tseveen B, Genden P, et al. Cancer Incidence and Cancer Control in Mongolia: Results from the National Cancer Registry 2008-12. IJC Int J Cancer. 2016;140(2):302–9.
4. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. Am Cancer Soc J. 2024;74(3):229–63.
5. Alfalah R. Jenis Histopatologi Berdasarkan Stadium pada Pasien Kanker Payudara di RSUCM Aceh Utara Tahun 2020. Matriks J Sos dan Sains. 2022;4(1):21–300.
6. Farahdiba BA, Nugroho YS. Klasifikasi Kanker Payudara Menggunakan Algoritma Gain Ratio. J Tek Elektro. 2016;8(2):43–6.
7. Narulita S, Prihati, Oktaga AT, Widyantoro AE. Performansi Algoritma Clustering K-Means untuk Penentuan Status Malnutrisi pada Balita. ISAINTEK J Informasi, Sains, dan Teknol. 2023;6(1):188–202.
8. Gbenga DE, Christopher N, Yetunde DC. Performance Comparison of Machine Learning Techniques for Breast Cancer Detection. Nov J Eng Appl Sci. 2017;6(1):1–8.
9. Henderi, Wahyuningsih T, Rahwanto E. Comparison of Min-Max Normalization and Z-Score Normalization in the K-Nearest Neighbor (KNN) Algorithm to Test the Accuracy of Types of Breast Cancer. Bright Int J Informatics Inf Syst. 2021;4(1):13–20.
10. Enriko IKA, Melinda, Sulyani AC, Astawa IGB. Breast Cancer Recurrence Prediction System using k-Nearest Neighbor, Naive-Bayes, and Support Vector Machine Algorithm. J Infotel Informatics, Telecommun Electron. 2021;13(4):185–8.
11. Atthalla IN, Jovandy A, Habibie H. Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K Nearest Neighbor (KNN). In: Prosiding Annual Research Seminar 2018. Palembang: Fakultas Ilmu Komputer UNSRI; 2018. p. 148–51.
12. Alfian G, Syafrudin M, Fahrurrozi I, Fitriyani NL, Atmaji FTD, Widodo T, et al. Predicting Breast Cancer from Risk Factors using SVM and Extra-Trees-Based Feature Selection Method. Computers. 2022;11(9):1–14.

13. R MT, Kaladevi AC, M BJ, Vivek V, Prabu M, Muthukumaran V. An Efficient Ensemble Method using K-Fold Cross Validation for the Early Detection of Benign and Malignant Breast Cancer. Int J Integr Eng. 2022;14(7):204–16.
14. Fitrianingsih, Zuraeni B. Analisis Ramalan Cuaca di Sekupang, Kota Batam Menggunakan Algoritma Decision Tree dan Confusion Matrix. Ekosph J Ekon Pembang dan Manaj. 2024;1(3):15–26.
15. Narulita S, Prihati P, Priyambodo A. Analisis dan Komparasi Algoritma Klasifikasi untuk Prediksi Kerugian Tower Provider Akibat Penalti yang Diberikan oleh Operator Telekomunikasi karena Keterlambatan Penyelesaian Pekerjaan oleh Tower Provider. Cakrawala Inf. 2022;2(2):1–14.
16. Yulianto LD, Triayudi A, Sholihati ID. Implementation Educational Data Mining for Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4.5. J Mantik. 2020;4(1):441–51.
17. Sitompul N. Rapid Miner Testing with the KNN Algorithm. J Data Sci. 2023;1(2):30–6.
18. Tarigan LRA, Dahlan. Optimalisasi Fitur dengan Forward Selection pada Estimasi Tingkat Penyakit Paru-Paru Menggunakan Algoritma Klasifikasi Random Forest. JATI J Mhs Tek Inform. 2024;8(5):10341–8.