OPEN ACCESS
LEDGER, VOL. 3, NO. 2, MAY 2024
ISSN: 2963 – 8798 (ONLINE)

LEDGER: Journal Informatic and Information Technology

Penerapan *Recursive Feature Elimination* pada *Support Vector Machine* untuk Klasifikasi Kanker Payudara

Herlinda Sundari*, Muhammad Afrizal Amrustian, Aditya Dwi Putro Wicaksono

Fakultas Informatika, Institut Teknologi Telkom Purwokerto, Indonesia
*Corresponding Author: herlindasundari@gmail.com

Abstract

Breast cancer is a prevalent type of cancer among women worldwide and can have fatal consequences if not detected early. Errors in breast cancer diagnosis can occur due to the use of irrelevant features or attributes, leading to misclassification. To minimize this possibility, this study applies the Recursive Feature Elimination (RFE) feature selection method to the WDBC (Wisconsin Diagnostic Breast Cancer) dataset to select the most relevant features in distinguishing benign and malignant tumor classes. SVM (Support Vector Machine) algorithm was used as the classification model with a data sharing ratio of 90:10, resulting in an accuracy of 0.98, precision of 1.00, recall of 0.94, and F1-score of 0.97. The implementation of RFE successfully reduced 50% of the features without reducing the performance of the model compared to the use of all features.

Keywords: Breast Cancer, Recursive Feature Elimination (RFE), Support Vector Machine (SVM), WDBC Dataset

Abstrak

Kanker payudara adalah jenis kanker yang lazim terjadi pada wanita di seluruh dunia dan dapat berakibat fatal jika tidak terdeteksi secara dini. Kesalahan dalam diagnosis kanker payudara dapat terjadi karena penggunaan fitur atau atribut yang tidak relevan, yang menyebabkan kesalahan klasifikasi. Untuk memperkecil kemungkinan tersebut, penelitian ini menerapkan metode seleksi fitur *Recursive Feature Elimination* (RFE) pada dataset WDBC (*Wisconsin Diagnostic Breast Cancer*) untuk memilih fitur-fitur yang paling relevan dalam membedakan kelas tumor jinak dan ganas. Algoritma SVM (*Support Vector Machine*) digunakan sebagai model klasifikasi dengan rasio pembagian data 90:10, menghasilkan akurasi 0.98, presisi 1.00, *recall* 0.94, dan *F1-score* 0.97. Implementasi RFE berhasil mengurangi 50% fitur tanpa mengurangi kinerja model dibandingkan dengan penggunaan seluruh fitur.

Kata Kunci: Dataset WDBC, Kanker Payudara, *Recursive Feature Elimination* (RFE), *Support Vector Machine* (SVM)

I. INTRODUCTION

Organisasi Kesehatan Dunia (WHO) melaporkan bahwa pada tahun 2020, terdapat 2,3 juta kasus kanker payudara yang didiagnosis secara global, dengan angka kematian mencapai 685.000 jiwa. Pada akhir tahun tersebut, WHO mencatat bahwa terdapat 7,8 juta wanita yang masih hidup setelah didiagnosis kanker payudara dalam lima tahun terakhir [1]. Berdasarkan tren morbiditas dan mortalitas terkait kanker payudara saat ini, diperkirakan bahwa pada tahun 2030, jumlah kasus dan kematian akibat kanker payudara akan mencapai 2,64 juta dan 1,7 juta [2]. Peluang dan probabilitas kelangsungan hidup dapat meningkat secara signifikan melalui diagnosis dini kanker payudara karena hal ini memungkinkan pasien untuk menerima perawatan klinis tepat waktu. Salah satu metode untuk mendeteksi kanker payudara adalah dengan mengklasifikasikan tumor [3].

Dalam usaha untuk meningkatkan diagnosis kanker payudara dengan akurasi yang lebih tinggi, banyak penelitian telah dijalankan untuk mengembangkan model klasifikasi berdasarkan analisis data medis. Dataset WDBC (*Wisconsin Diagnostic Breast Cancer*) adalah salah satu sumber data yang digunakan dalam penelitian tersebut. Dataset ini berisi beragam fitur yang menggambarkan karakteristik

ISSN: 2963 – 8798 (ONLINE)

sel-sel yang ditemukan dalam sampel biopsi kanker payudara, digunakan untuk membedakan kelas antara kanker payudara yang bersifat jinak (*benign*) dan yang bersifat ganas (*malignant*) [4]. Dataset yang terdiri dari 569 sampel dengan 30 fitur memiliki dimensi yang cukup tinggi. Untuk mengatasi hal ini dan meningkatkan kinerja model, diperlukan seleksi fitur untuk mengidentifikasi fitur-fitur yang paling relevan dalam memprediksi kelas.

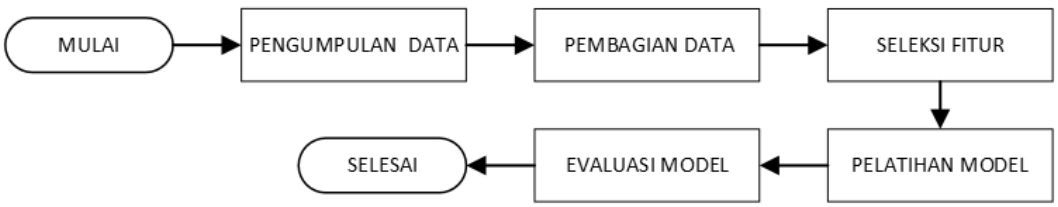
II. LITERATURE REVIEW

Recursive Feature Elimination (RFE) adalah salah satu teknik seleksi fitur yang menggunakan metode rekursif untuk menghilangkan fitur-fitur yang kurang relevan, sehingga diperoleh fitur terbaik untuk membangun model [5]. Penelitian [6] menunjukkan pengaruh RFE dalam meningkatkan akurasi model prediksi penyakit hati dengan menggunakan algoritma *Artificial Neural Network* (ANN) dan AdaBoost. Model AdaBoost tanpa RFE mencapai akurasi sebesar 86.74%, sementara dengan RFE, akurasinya meningkat menjadi 89.15%. Demikian pula, model ANN tanpa RFE memiliki akurasi sebesar 90.36%, namun dengan RFE, akurasinya meningkat menjadi 92.77%. Tidak hanya itu, penelitian [7] membandingkan metode seleksi fitur RFE dengan *Principal Component Analysis* (PCA) dan Kernel PCA (KPCA) yang diaplikasikan untuk memilih fitur TfEn (*Time-frequency Entropy*) yang paling optimal. Hasilnya menunjukkan keunggulan RFE dibandingkan dengan PCA dan KPCA. Terlihat dari nilai rata-rata akurasi total fitur TfEn mentah sebesar 98.80%, yang meningkat menjadi 100% dengan penerapan RFE. Sementara itu, PCA dan KPCA masing-masing hanya mencapai 88.80% dan 82.40%. Berdasarkan tinjauan literatur tersebut, penelitian ini memilih RFE sebagai metode seleksi fitur pada dataset WDBC.

Merujuk pada penelitian sebelumnya [8], dimana nilai akurasi klasifikasi dataset WDBC mencapai 92.228% menggunakan algoritma C4.5, penelitian ini mengeksplorasi metode alternatif dalam klasifikasi menggunakan *Support Vector Machine* (SVM). Hal ini didasarkan pada penelitian lain yang menunjukkan bahwa SVM memiliki kinerja lebih unggul dibandingkan dengan algoritma C4.5. Sebagai contoh, dalam penelitian [9] perbandingan algoritma C4.5 dan SVM untuk memprediksi ketepatan waktu kelulusan mahasiswa. Hasilnya menunjukkan tingkat presisi 81% dan tingkat akurasi 80% untuk Algoritma C4.5, sementara SVM lebih unggul dengan tingkat presisi 88% dan tingkat akurasi 85%. Penelitian [10] juga menunjukkan bahwa SVM mencapai performa terbaik dengan akurasi sebesar 90%, melebihi C4.5 yang hanya mencapai 85% dan KNN (*K-nearest neighbors*) dengan akurasi 88% dalam menentukan rata-rata kredit macet koperasi. Selanjutnya, penelitian [11] membandingkan kinerja algoritma SVM, C4.5, *Naive Bayes*, *Logistic Regression*, *Random Forest*, XGBoost, dan KNN dalam deteksi intrusi. Hasil eksperimen menunjukkan bahwa SVM secara signifikan melampaui algoritma-algoritma lainnya dalam hal akurasi, presisi, dan *recall*.

III. RESEARCH METHOD

Penelitian ini melibatkan beberapa tahapan yang dapat dilihat pada Gambar 1.



Gambar 1. Diagram Alir Penelitian

A. *Pengumpulan Data*

Penelitian ini menggunakan dataset kanker payudara WDBC (*Wisconsin Diagnostic Breast Cancer*) yang diperoleh dari UCI Machine Learning Repository. Informasi mengenai dataset dapat dilihat pada Tabel 1.

ISSN: 2963 – 8798 (ONLINE)

TABEL 1
INFORMASI DATASET WDBC

Nama Variabel	Fungsi	Tipe Data
ID	ID	int64
radius1, texture1, perimeter1, area1, smoothness1, compactness1, concavity1, concave_points1, symmetry1, fractal_dimension1, radius2, texture2, perimeter2, area2, smoothness2, compactness2, concavity2, concave_points2, symmetry2, fractal_dimension2, radius3, texture3, perimeter3, area3, smoothness3, compactness3, concavity3, concave_points3, symmetry3, fractal_dimension3	Fitur	float64
Diagnosis	Kelas	object

B. *Pembagian Data*

Dataset dibagi menjadi beberapa skenario untuk menentukan rasio terbaik antara data pelatihan dan data pengujian. Empat skenario yang diuji dapat dilihat pada Tabel 2.

TABEL 2
SKENARIO PEMBAGIAN DATASET

Skenario	Data Pelatihan	Data Pengujian
1	60%	40%
2	70%	30%
3	80%	20%
4	90%	10%

C. *Seleksi Fitur*

Tahap seleksi fitur menggunakan metode *Recursive Feature Elimination* (RFE) dengan *Support Vector Machine* (SVM) sebagai model dasar. RFE umumnya terdiri dari tiga tahap. Tahap pertama adalah melatih dataset untuk menghitung bobot setiap fitur. Bobot fitur ini diperoleh dari nilai α yang dihasilkan oleh model SVM. Persamaan untuk menghitung bobot fitur didefinisikan oleh (1) [12].

$$w = \sum_1^k \alpha_k y_k x_k \tag{1}$$

dimana x_k adalah data pelatihan dari fitur ke-k dan y_k adalah label kelas dari fitur ke-k. Kemudian tahap kedua yaitu menghitung kriteria peringkat untuk mengurutkan fitur berdasarkan bobotnya. Fungsi peringkat kriteria didefinisikan dengan persamaan (2) [12].

$$c_k = w_k^2, k = 1,2,\ldots,|S| \tag{2}$$

dimana nilai peringkat kriteria (c) adalah bobot (w) yang dikuadratkan untuk setiap fitur ke-k, mulai dari fitur pertama (k=1) hingga fitur terakhir (k=|S|). Tahap ketiga yaitu mengurutkan fitur berdasarkan nilai bobot dan mengeliminasi satu fitur dengan nilai bobot terkecil di setiap iterasi [12].

D. *Pelatihan Model*

Pada tahap ini, dilakukan pelatihan model klasifikasi menggunakan algoritma SVM. Untuk mengetahui pengaruh jumlah fitur terhadap kinerja model, dilakukan pengurangan fitur secara bertahap berdasarkan peringkat fitur yang paling rendah. Sebagai alternatif dari pengurangan satu per satu, fitur dikurangi dengan kelipatan 5. Dengan demikian, diperoleh model dengan 30, 25, 20, 15, 10, dan 5 fitur.

ISSN: 2963 – 8798 (ONLINE)

Pendekatan ini memungkinkan identifikasi lebih cepat terhadap titik di mana pengurangan fitur selanjutnya tidak lagi meningkatkan kinerja secara signifikan.

E. *Evaluasi Model*

Proses evaluasi berfokus pada metrik akurasi, presisi, *recall*, *f1-score* untuk menilai seberapa baik model klasifikasi yang dibangun dengan algoritma SVM dalam mengklasifikasikan data. Selain itu juga untuk mengetahui bagaimana dampak dari penghapusan fitur-fitur yang dianggap kurang relevan berdasarkan pemeringkatan oleh RFE terhadap model. Adapun perhitungan untuk setiap metrik adalah sebagai berikut:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

$$F_1 = 2^* \frac{precision^* recall}{precision + recall} = \frac{TP}{TP + \frac{FN + FP}{2}} \tag{6}$$

dimana TP (*True Positives*) adalah jumlah sampel dari kelas A yang terklasifikasi sebagai kelas A. FP (*False Positives*) adalah jumlah sampel dari kelas B yang salah terklasifikasi sebagai kelas A. TN (*True Negatives*) adalah jumlah sampel dari kelas B yang terklasifikasi sebagai kelas B. Sedangkan FN (*False Negatives*) adalah jumlah sampel dari kelas A yang salah terklasifikasi sebagai kelas B [13] [14].

IV. RESULTS AND DISCUSSION

Percobaan pembagian dataset menggunakan metode SVM dengan melibatkan seluruh fitur. Hasil dari percobaan ini menunjukkan bahwa skenario keempat memiliki performa yang lebih unggul dibandingkan dengan ketiga skenario lainnya. Analisis lebih rinci dapat ditemukan dalam Tabel 3.

TABEL 3
HASIL PERCOBAAN PEMBAGIAN DATASET

Skenario	Presisi	Recall	F1 Score	Akurasi
1	0.95	0.90	0.92	0.94
2	0.95	0.93	0.94	0.95
3	0.93	0.93	0.93	0.95
4	1.00	0.94	0.97	0.98

Berdasarkan hasil tersebut, maka penelitian ini menggunakan pembagian dataset dengan rasio pembagian data 90:10, dimana 90% dari data digunakan untuk pelatihan, terdiri dari 512 data, dan 10% sisanya digunakan untuk pengujian, terdiri dari 57 data. Kemudian memasuki proses seleksi fitur untuk mengetahui posisi peringkat setiap fitur. Hasil dari pemeringkatan seluruh fitur dari yang terbaik dapat dilihat dalam Tabel 4.

TABEL 4
HASIL PEMERINGKATAN RFE DARI YANG TERBAIK

Peringkat	Fitur	Peringkat	Fitur
1	*concave_points3*	16	*perimeter1*
2	*concavity3*	17	*fractal_dimension3*
3	*symmetry3*	18	*concavity2*
4	*compactness3*	19	*compactness2*
5	*smoothness3*	20	*concave_points2*
6	*texture2*	21	*texture1*
7	*radius3*	22	*area2*
8	*radius1*	23	*smoothness2*

ISSN: 2963 – 8798 (ONLINE)

Peringkat	Fitur	Peringkat	Fitur
9	*concavity1*	24	*perimeter3*
10	*concave_points1*	25	*radius2*
11	*symmetry1*	26	*fractal_dimension2*
12	*compactness1*	27	*fractal_dimension1*
13	*smoothness1*	28	*symmetry2*
14	*perimeter2*	29	*area3*
15	*texture3*	30	*area1*

Selanjutnya proses pelatihan model dengan skenario pengurangan fitur bertahap untuk mengamati pengaruh penghapusan fitur dengan peringkat rendah terhadap performa model. Hasil keseluruhan evaluasi model dengan pengurangan fitur secara bertahap dapat ditemukan dalam Tabel 5.

TABEL 5
HASIL EVALUASI MODEL DENGAN PENGURANGAN FITUR

Fitur yang Dihapus	Fitur yang Digunakan	Presisi	Recall	F1 Score	Akurasi
0	30	1.00	0.94	0.97	0.98
5	25	1.00	0.94	0.97	0.98
10	20	1.00	0.94	0.97	0.98
15	15	1.00	0.94	0.97	0.98
20	10	0.89	0.94	0.92	0.95

Pada tahap pengurangan fitur hingga 20 (tersisa 10 fitur), terjadi penurunan nilai akurasi, sehingga pengujian dihentikan pada tahap ini. Untuk menentukan batas perubahan yang tepat, dilakukan pengujian pada penggunaan 11 hingga 14 fitur. Hasil menunjukkan bahwa penurunan akurasi mulai terlihat pada penggunaan 14 fitur, sementara nilai akurasi, presisi, *recall*, dan *F1-score* tetap stabil untuk penggunaan 10 hingga 14 fitur. Lima belas fitur terbaik yang mampu mempertahankan performa evaluasi tertinggi adalah: *concave_points3*, *concavity3*, *symmetry3*, *compactness3*, *smoothness3*, *texture2*, *radius3*, *radius1*, *concavity1*, *concave_points1*, *symmetry1*, *compactness1*, *smoothness1*, *perimeter2*, dan *texture3*.

V. CONCLUSION

Pengurangan fitur secara bertahap menggunakan metode *Recursive Feature Elimination* (RFE) menunjukkan bahwa penggunaan 15 hingga 25 fitur menghasilkan nilai evaluasi yang konsisten dan sama seperti hasil pengujian pada 30 fitur, dengan nilai presisi mencapai 1.00, *recall* 0.94, *f1-score* 0.97, dan akurasi sebesar 0.98. Hal ini menunjukkan bahwa implementasi RFE berhasil mengurangi 50% fitur pada dataset WDBC tanpa mengurangi kinerja model dibandingkan dengan penggunaan seluruh fitur. Penelitian ini dapat dikembangkan lagi dengan membandingkan hasil pemeringkatan RFE dengan base model selain SVM atau melakukan observasi lebih lanjut mengapa fitur-fitur tertentu cenderung memiliki ranking lebih tinggi dari segi medis.

REFERENCES

[1] WHO, “Breast cancer,” 2023. https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed Oct. 13, 2023).
[2] C. Shang and D. Xu, “Epidemiology of Breast Cancer,” *Oncologie*, vol. 24, no. 4, pp. 649–663, 2022, doi: 10.32604/oncologie.2022.027640.
[3] S. Ara, A. Das, and A. Dey, “Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms,” in *2021 International Conference on Artificial Intelligence, ICAI 2021*, 2021, pp. 97–101, doi: 10.1109/ICAI52203.2021.9445249.
[4] W. Wolberg, O. Mangasarian, N. Street, and W. Street, “Breast Cancer Wisconsin (Diagnostic),” *UCI Machine Learning Repository*, 1995. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic (accessed Oct. 12, 2023).
[5] A. R. I. Pratama, S. A. Latipah, and B. N. Sari, “OPTIMASI KLASIFIKASI CURAH HUJAN MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM) DAN RECURSIVE FEATURE ELIMINATION (RFE),” vol. 7, no. 2, pp. 314–324, 2022.
[6] A. S. Jaddoa, S. J. Saba, and E. A. A. Al-Kareem, “Liver Disease Prediction Model Based on Oversampling Dataset with RFE Feature Selection using ANN and AdaBoost algorithms,” vol. 4, no. 2, pp. 85–93, 2023.
[7] X. Su, H. Liu, and L. Tao, “TF entropy and RFE based diagnosis for centrifugal pumps subject to the limitation of failure samples,” *Appl. Sci.*, vol. 10, no. 8, 2020, doi: 10.3390/APP10082932.
[8] F. A. Andoyo, “OPTIMASI AKURASI KLASIFIKASI MENGGUNAKAN K- MEANS DAN ALGORITMA GENETIKA DENGAN MENGINTEGRASIKAN ALGORITMA C4.5 UNTUK DIAGNOSIS KANKER PAYUDARA,” Universitas Negeri Semarang, 2020.

ISSN: 2963 – 8798 (ONLINE)

[9] A. Mailana, A. A. Putra, S. Hidayat, and A. Wibowo, "Comparison of C4.5 Algorithm and Support Vector Machine in Predicting the Student Graduation Timeliness," *J. Online Inform.*, vol. 6, no. 1, p. 11, 2021, doi: 10.15575/join.v6i1.608.
[10] Siswanto *et al.*, "Penerapan Algoritma C4.5, SVM, dan KNN Untuk Menentukan Rata-Rata Kredit Macet Koperasi," *Pros. Semin. Nas. SISFOTEK (Sistem Inf. dan Teknol.*, vol. 7, no. 1, pp. 274–281, 2023.
[11] J. Dhanke, R. N. Patil, I. Kumari, S. Gupta, S. Hans, and K. Kumar, "Comparative Study of Machine Learning Algorithms for Intrusion Detection," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 4s, pp. 647–653, 2023.
[12] A. Bustamam, A. Bachtiar, and D. Sarwinda, "Selecting features subsets based on support vector machine-recursive features elimination and one dimensional-naïve bayes classifier using support vector machines for classification of prostate and breast cancer," in *Procedia Computer Science*, 2019, vol. 157, pp. 450–458, doi: 10.1016/j.procs.2019.08.238.
[13] A. S. Jaddoa, S. J. Saba, and E. A. A. Al-Kareem, "Liver Disease Prediction Model Based on Oversampling Dataset with RFE Feature Selection using ANN and AdaBoost algorithms," vol. 4, no. 2, pp. 85–93, 2023.
[14] A. Geron, *Hands-On Machine Learning With Scikit-Learn, Keras & TensorFlow*, 2nd Editio. Gravenstein Highway North : O'Reilly Media., 2019.