DATAMATIKA

Tersedia secara daring di:
https://conference.ipb.ac.id/datamatika

ANALYSIS OF THE EFFECT OF HIGHLY CORRELATED FEATURE REDUCTION AND RECURSIVE FEATURE ELIMINATION ON BREAST CANCER CLASSIFICATION USING SVM AND RANDOM FOREST

ANALISIS PENGARUH PENGURANGAN FITUR BERKORELASI TINGGI SERTA *RECURSIVE FEATURE ELIMINATION* TERHADAP KLASIFIKASI KANKER PAYUDARA DENGAN SVM DAN *RANDOM FOREST*

Adinda Rachmalia[1], Mochamad Tito Julianto[2‡], and Elis Khatizah[3‡]

[1,2,3]Department of Mathematics, IPB University, Indonesia
‡corresponding author: adindarachmalia@apps.ipb.ac.id

Abstract

Breast cancer is one of the most common cancers among women and requires accurate classification methods to support early diagnosis. This study aims to analyze the effect of removing highly correlated features and applying Recursive Feature Elimination (RFE) on breast cancer classification using Support Vector Machine (SVM) and Random Forest. The dataset used in this study is the Wisconsin Breast Cancer Diagnostic Dataset obtained from the UCI Machine Learning Repository. The research employed four classification scenarios, namely baseline, Variance Inflation Factor (VIF) based feature reduction, RFE with 15 features and a combination of VIF and RFE. The results showed that the combination of VIF and RFE produced the best performance for the SVM model using 15 selected features, while the Random Forest model achieved the best performance using RFE without VIF. In general, feature selection methods were able to reduce the number of features from 30 to 15 features while maintaining high classification performance.

Keywords: Breast Cancer Classification, Random Forest, Recursive Feature Elimination, Support Vector Machine, Variance Inflation Factor.

Abstrak

Kanker payudara merupakan salah satu jenis kanker dengan tingkat kejadian yang tinggi pada perempuan sehingga diperlukan metode klasifikasi yang akurat untuk mendukung diagnosis dini. Penelitian ini bertujuan menganalisis pengaruh pengurangan fitur berkorelasi tinggi serta penggunaan Recursive Feature Elimination (RFE) terhadap klasifikasi kanker payudara menggunakan algoritma Support Vector Machine (SVM) dan Random Forest. Dataset yang digunakan adalah Wisconsin Breast Cancer Diagnostic Dataset dari UCI Machine Learning Repository. Penelitian dilakukan menggunakan empat skenario, yaitu baseline, RFE 15 fitur, pengurangan fitur menggunakan Variance Inflation Factor (VIF), serta kombinasi VIF dan RFE. Hasil penelitian menunjukkan bahwa kombinasi VIF dan RFE menghasilkan performa terbaik pada model SVM menggunakan 15 fitur terpilih, sedangkan model Random Forest menghasilkan performa terbaik pada skenario RFE tanpa VIF. Secara umum, metode seleksi fitur mampu mengurangi jumlah fitur dari 30 fitur menjadi 15 fitur tanpa menurunkan performa klasifikasi secara signifikan.

Kata Kunci: Klasifikasi Kanker Payudara, Random Forest, Recursive Feature Elimination*, Support Vector Machine*, *Variance Inflation Factor.*

1. Pendahuluan

Kanker merupakan salah satu penyebab utama kematian di dunia, dengan kanker payudara sebagai jenis kanker dengan kasus tertinggi pada perempuan. Berdasarkan laporan World Health Organization (2022), kanker payudara menyerang sekitar 2,3 juta perempuan di dunia dengan angka kematian mencapai 670.000 jiwa. Sementara itu, data GLOBOCAN (2022), memperkirakan terdapat 66.271 kasus baru kanker payudara di Indonesia dengan angka kematian mencapai 22.598 jiwa (Ferlay *et al*., 2024). Tingginya angka tersebut menunjukkan bahwa deteksi dini kanker payudara masih menjadi permasalahan penting dalam bidang kesehatan (Rambe *et al*., 2025).

Perkembangan *machine learning* mendorong pemanfaatannya dalam bidang kesehatan, khususnya pada proses klasifikasi penyakit. Data medis yang dihasilkan umumnya memiliki jumlah dan kompleksitas yang tinggi sehingga memerlukan metode analisis yang tepat (Dhinakaran *et al*., 2025). Algoritma *Support Vector Machine* (SVM) dan *Random Forest* banyak digunakan pada klasifikasi kanker payudara karena memiliki performa yang baik pada data berdimensi tinggi. Namun, data medis umumnya memiliki banyak fitur dengan korelasi tinggi sehingga dapat menyebabkan multikolinearitas dan redundansi informasi (Yildirim, 2024). Oleh karena itu, penelitian ini menerapkan *Variance Inflation Factor* (VIF) untuk mengurangi multikolinearitas serta *Recursive Feature Elimination* (RFE) untuk memilih fitur yang paling relevan terhadap proses klasifikasi.

Beberapa penelitian terdahulu menunjukkan bahwa seleksi fitur mampu meningkatkan performa klasifikasi kanker payudara. Penelitian Tinaliah & Elizabeth (2024) memperoleh nilai *recall* sebesar 97,67% menggunakan 10 fitur terpilih, sedangkan penelitian Juarto (2023) menunjukkan bahwa penggunaan VIF mampu meningkatkan performa klasifikasi dengan mengurangi multikolinearitas. Oleh karena itu, penelitian ini bertujuan menganalisis pengaruh pengurangan fitur berkorelasi tinggi menggunakan VIF dan penerapan RFE terhadap performa klasifikasi kanker payudara menggunakan algoritma SVM dan *Random Forest*.

2. Metodologi

2.1 Bahan dan Data

Penelitian ini menggunakan Wisconsin Breast Cancer Diagnostic (WBCD) Dataset yang diperoleh dari UCI Machine Learning Repository. Dataset ini berasal dari hasil pemeriksaan *Fine Needle Aspiration* (FNA) pada pasien di University of Wisconsin Hospitals. Dataset terdiri atas 569 data tumor dan 30 fitur numerik yang berasal dari 10 karakteristik inti sel, yaitu *radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry,* dan *fractal dimension.* Setiap karakteristik terdiri atas nilai *mean, standard error* (SE), dan *worst*. Variabel target pada dataset berupa diagnosis tumor, yaitu jinak atau ganas.

2.2 Metode Penelitian

Penelitian ini dilakukan menggunakan perangkat lunak Python untuk proses pengolahan dan analisis data. Tahapan penelitian yang dilakukan adalah sebagai berikut.

1) Pre-processing data

Melakukan transformasi label target ke bentuk numerik, deteksi *missing value*, serta penghapusan kolom yang tidak relevan.

2) *Exploratory Data Analysis* (EDA)

a. Analisis korelasi fitur menggunakan *heatmap* korelasi Pearson.
Nilai korelasi Pearson berada pada rentang -1 hingga +1, di mana nilai yang semakin mendekati nilai +1 atau -1 menunjukkan hubungan linear yang semakin kuat antara dua fitur, sedangkan nilai yang mendekati 0 menunjukkan hubungan linear yang lemah antara kedua fitur tersebut (Kumar dan Rani 2024). Secara matematis, koefisien korelasi Pearson dirumuskan sebagai berikut (Shrestha, 2020).

$$r_{xy} = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}, \quad (1)$$

dengan r menyatakan korelasi antara X dan Y, n menyatakan banyaknya sampel, X menyatakan variabel pertama, dan Y menyatakan variabel kedua.

b. Deteksi multikolinearitas menggunakan *Variance Inflation Factor* (VIF).
Multikolinearitas merupakan kondisi ketika terdapat hubungan linear yang tinggi antarfitur, sehingga menyebabkan redundansi informasi. Pada penelitian ini, multikolinearitas dideteksi menggunakan *Variance Inflation Factor* (VIF). Secara matematis, VIF dirumuskan sebagai berikut.

$$VIF_j = \frac{1}{1 - R_j^2} = \frac{1}{Tolerance}, \quad (2)$$

dengan VIF_j menyatakan nilai VIF fitur ke-j, dan R_j^2 menyatakan koefisien determinasi yang menunjukkan seberapa baik suatu fitur dapat dijelaskan oleh fitur-fitur independen lainnya. Nilai VIF yang semakin besar menunjukkan tingkat multikolinearitas yang semakin tinggi. Nilai $VIF = 1$ menunjukkan tidak terdapat korelasi antar fitur, nilai $1 < VIF < 5$ menunjukkan korelasi sedang, sedangkan nilai $VIF > 10$ mengindikasikan adanya multikolinearitas yang tinggi pada model (Shrestha, 2020).

3) Penghapusan fitur dengan nilai VIF>10.
4) Pembagian data
Digunakan rasio 70:30, yaitu 70% data digunakan untuk pelatihan model (data *training)* dan 30% digunakan untuk pengujian model (data *testing*).
5) Pembentukan Skenario dataset
Pada penelitian ini dibentuk empat skenario dataset, yaitu
   a. Dataset sebelum pengurangan fitur berkorelasi tinggi (*baseline*).
   b. Dataset setelah pengurangan fitur berkorelasi tinggi yang memiliki nilai VIF>10.
   c. Dataset setelah seleksi fitur menggunakan RFE dengan 15 fitur terpilih.
   d. Dataset setelah pengurangan fitur berkorelasi tinggi (VIF>10) yang selanjutnya diseleksi fitur menggunakan RFE hingga diperoleh 15 fitur terpilih.

   Keempat skenario digunakan untuk membandingkan pengaruh pengurangan fitur berkorelasi tinggi terhadap performa model klasifikasi.
6) Pembangunan model SVM
Support Vector Machine (SVM) merupakan algoritma *supervised learning* yang digunakan untuk proses klasifikasi dengan mencari *hyperplane* optimal sebagai pemisah antar kelas. Pada penelitian ini, model SVM dibangun menggunakan *default hyperparameter* dengan

kernel linear. Secara matematis, fungsi keputusan SVM dirumuskan sebagai berikut.

$$f(x) = \boldsymbol{w}^T \boldsymbol{x} + b, \quad (3)$$

dengan $\boldsymbol{w}$ menyatakan vektor bobot, $\boldsymbol{x}$ menyatakan vektor fitur, dan b menyatakan bias (Géron, 2019).

a. Standarisasi data

Standarisasi data dilakukan menggunakan *StandardScaler* untuk menyamakan skala fitur pada algoritma SVM. Secara matematis, standarisasi data dirumuskan sebagai berikut.

$$z = \frac{x - \mu}{\sigma}, \quad (4)$$

dengan z menyatakan hasil standarisasi, x menyatakan nilai yang di amati, μ menyatakan rata-rata data, dan σ menyatakan standar deviasi data (Phudjashakty, 2026).

7) Pembangunan model *Random Forest*

Random Forest merupakan algoritma *ensemble learning* yang bekerja dengan membangun banyak *decision tree* menggunakan subset data dan subset fitur yang dipilih secara acak (Asri *et al*., 2026). Pada penelitian ini, *random forest* dibangun menggunakan *default hyperparameter.* Hasil prediksi akhir diperoleh melalui mekanisme *majority voting* sehingga model menjadi lebih stabil dan mampu mengurangi *overfitting* dibandingkan penggunaan satu *decision tree* saja (Nugroho, 2025).

8) Seleksi fitur menggunakan *Recursive Feature Elimination* (RFE)

Recursive Feature Elimination (RFE) merupakan metode seleksi fitur yang bekerja dengan menghapus fitur secara bertahap berdasarkan tingkat kepentingannya hingga diperoleh subset fitur optimal. Pada penelitian ini, jumlah fitur hasil seleksi RFE ditetapkan sebanyak 15 fitur.

a. Seleksi fitur RFE model SVM

Proses seleksi fitur pada model SVM dengan kernel linear dilakukan menggunakan nilai bobot pada vektor $\boldsymbol{w}$ yang dirumuskan sebagai berikut.

$$\boldsymbol{w} = \sum_{k=1}^{m} \propto_k y_k \boldsymbol{x}_k, \quad (5)$$

dengan α_k menyatakan koefisien *classifier* yang dihasilkan dari proses pelatihan SVM, $\boldsymbol{x}_k$ menyatakan vektor fitur data ke-k, dan y_k menyatakan label kelas data ke-k. Selanjutnya, dilakukan perhitungan kriteria peringkat fitur berdasarkan kuadrat nilai bobot fitur sebagai berikut.

$$c_k = w_k^2, \quad (6)$$

nilai c_k digunakan untuk menentukan tingkat kepentingan setiap

fitur. Semakin besar nilai c_k, maka semakin besar kontribusi fitur terhadap proses klasifikasi. Fitur kemudian diurutkan berdasarkan nilai bobotnya, dan fitur dengan nilai terkecil dieliminasi secara bertahap pada setiap iterasi hingga diperoleh fitur terbaik (Bustamam *et al*., 2019).

b. Seleksi fitur RFE model *Random Forest*
Pada algoritma *Random Forest,* proses RFE dilakukan menggunakan nilai *feature importance.* Fitur dengan nilai *importance* terendah akan dieliminasi secara bertahap pada setiap iterasi hingga diperoleh subset fitur terbaik sesuai jumlah fitur yang ditentukan.

9) Melakukan evaluasi model pada masing-masing skenario.
Evaluasi model klasifikasi dilakukan untuk mengukur kemampuan model dalam melakukan prediksi terhadap data. Proses evaluasi dilakukan menggunakan *confusion matrix* yang terdiri dari *True Positive* (TP), *True Negative* (TN)*, False Positive* (FP) dan *False Negative* (FN)*.* Berdasarkan *confusion matrix*, metrik evaluasi yang digunakan meliputi *accuracy, precision, recall,* dan *F1-Score* (Hidayat *et al*., 2026). Metrik *recall* digunakan sebagai indikator utama karena penting untuk meminimalkan kesalahan prediksi pada kasus kanker ganas (Chen *et al*., 2023).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$\mathrm{F1} - Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{10}$$

Selain itu, digunakan kurva *Receiver Operating Characteristic* (ROC) yang dibentuk berdasarkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR), yang masing-masing didefinisikan sebagai berikut.

$$\mathrm{TPR} = \frac{TP}{TP + FN}, \tag{11}$$

$$\mathrm{FPR} = \frac{FP}{FP + TN}. \tag{12}$$

Berdasarkan kurva ROC tersebut, diperoleh nilai *Area Under Curve* (AUC), yaitu luas area di bawah kurva ROC yang digunakan untuk

mengukur kemampuan model dalam membedakan kelas positif dan negatif. Nilai AUC berada pada rentang 0–1, di mana nilai 0,5–0,6 menunjukkan performa yang gagal, nilai 0,6–0,7 menunjukkan performa yang buruk, nilai 0,7–0,8 menunjukkan performa yang cukup baik, nilai 0,8–0,9 menunjukkan performa yang baik, dan nilai 0,9–1,0 menunjukkan performa yang sangat baik (Nahm, 2022). ROC-AUC digunakan untuk mengevaluasi kemampuan diskriminasi model dalam membedakan dua kelas target secara keseluruhan (Solang *et al*., 2026).

10) Membandingkan performa model pada setiap skenario penelitian dan membuat kesimpulan.

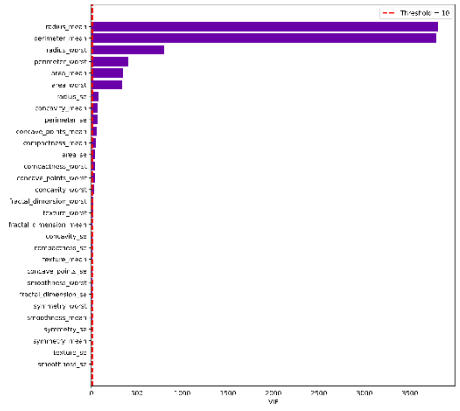
3. Hasil dan Pembahasan

3.1 *Pre-Processing*

Pada tahap *pre-processing,* dilakukan transformasi label target dengan mengubah kelas diagnosis *malignant* menjadi 1 dan *benign* menjadi 0 agar dapat diproses secara numerik oleh algoritma *machine learning*. Selanjutnya, dilakukan deteksi *missing value* dan diperoleh hasil bahwa dataset tidak memiliki *missing value*. Selain itu, kolom "*ID*" dihapus karena tidak berkontribusi terhadap proses klasifikasi kanker payudara.

3.2 Exploratory Data Analysis (EDA)

Hasil *Exploratory Data Analysis* (EDA) menunjukkan terdapat beberapa pasangan fitur dengan tingkat korelasi yang tinggi berdasarkan *heatmap* korelasi Pearson dan dilakukan pengujian menggunakan VIF untuk mengukur tingkat multikolinearitas pada setiap fitur.



Gambar 1: Nilai *variance inflation factor* (VIF) pada 30 fitur dataset.

Berdasarkan Gambar 1, terlihat bahwa pada kondisi awal banyak fitur memiliki nilai VIF di atas *threshold* 10 yang menunjukkan adanya

multikolinearitas tinggi antar fitur. Nilai VIF tertinggi terdapat pada *radius_mean* dan *perimeter_mean* dengan nilai yang sangat besar dibandingkan fitur lainnya.

3.3 Pengurangan Fitur Berkorelasi Tinggi

Berdasarkan hasil pengujian VIF, fitur dengan nilai VIF > 10 dihapus secara iteratif, yaitu dengan menghapus fitur yang memiliki nilai VIF tertinggi kemudian menghitung kembali nilai VIF pada fitur yang tersisa hingga seluruh fitur memiliki nilai VIF ≤ 10. Setelah proses tersebut dilakukan, jumlah fitur berkurang dari 30 fitur menjadi 17 fitur, sehingga redundansi informasi akibat multikolinearitas dapat dikurangi dan diharapkan menghasilkan model klasifikasi yang lebih optimal.

3.4 Pembagian Data

Dataset yang terdiri dari 569 data dibagi menjadi 398 data *training* (70%) dan 171 data *testing* (30%). Proporsi pembagian ini digunakan agar model dapat mempelajari pola dari data *training* dan kemudian diuji pada data *testing* untuk mengukur kemampuan generalisasi model terhadap data baru.

3.5 Pemodelan *Support Vector Machine* (SVM)

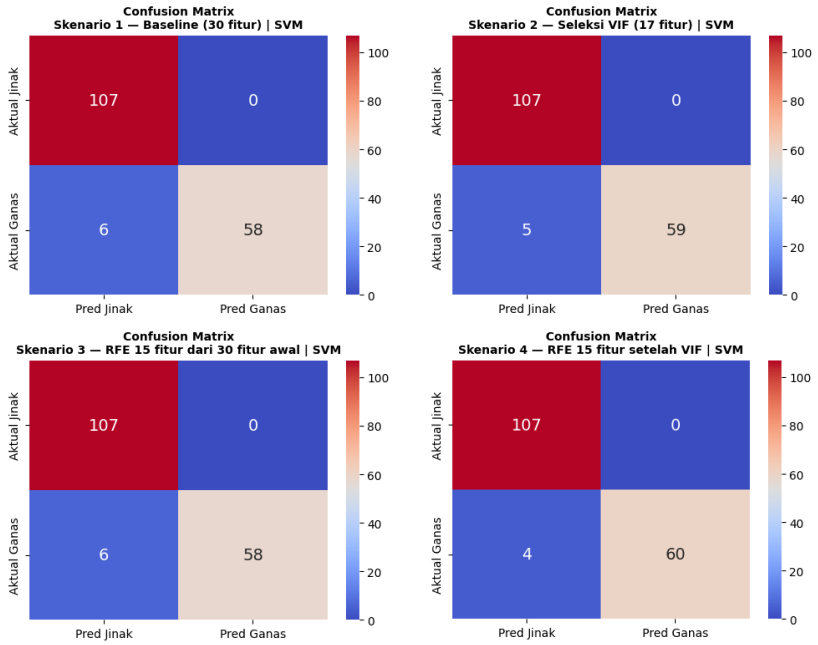
Hasil evaluasi model SVM pada setiap skenario dataset ditunjukkan pada Tabel 1.

Tabel 1: Evaluasi model SVM.

Skenario	Jumlah Fitur	*Accuracy*	*Precision*	*Recall*	*F1-Score*	AUC
1 (*Baseline*)	30	0,9649	1,0000	0,9062	0,9508	0,9945
2 (VIF)	17	0,9474	0,9825	0,8750	0,9256	0,9930
3 (RFE)	15	0,9649	1,0000	0,9062	0,9508	0,9968
4 (VIF+RFE)	**15**	**0,9766**	**1,0000**	**0,9375**	**0,9677**	**0,9955**

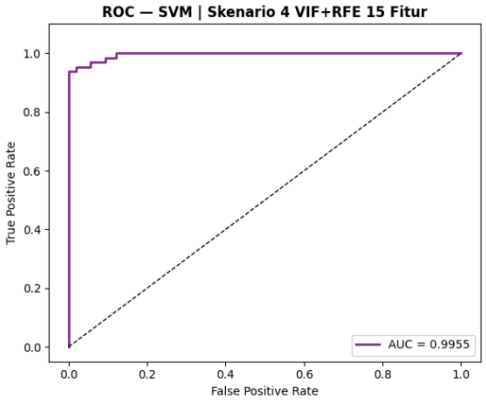
Berdasarkan hasil pengujian, performa terbaik model SVM diperoleh pada skenario kombinasi VIF dan RFE dengan nilai *accuracy* sebesar 0,9766, *precision* sebesar 1,0000, *recall* sebesar 0,9375, dan *F1-score* sebesar 0,9677. Hasil tersebut menunjukkan bahwa pengurangan multikolinearitas menggunakan VIF dan seleksi fitur menggunakan RFE mampu mengurangi redundansi informasi antarfitur serta mempertahankan fitur-fitur yang paling relevan terhadap proses klasifikasi. Sementara itu, skenario *baseline* dan RFE tanpa VIF menghasilkan nilai *accuracy,*

precision, recall, dan *f1-score* yang sama. Hal ini menunjukkan bahwa penerapan RFE tanpa pengurangan multikolinearitas belum memberikan peningkatan performa klasifikasi yang signifikan.



Gambar 2: *Confusion matrix* model SVM.

Hasil *confusion matrix* model SVM pada setiap skenario penelitian ditunjukkan pada Gambar 2. Berdasarkan hasil klasifikasi, skenario kombinasi VIF dan RFE menghasilkan jumlah *false negative* paling sedikit, sehingga lebih baik dalam mendeteksi kasus kanker payudara ganas. Selanjutnya, kurva ROC model SVM pada setiap skenario penelitian ditunjukkan pada Gambar 3. Nilai AUC yang mendekati 1 menunjukkan bahwa model memiliki kemampuan diskriminasi kelas yang sangat baik dalam membedakan kelas jinak dan ganas.



Gambar 3: Kurva ROC model terbaik SVM.

3.6 Pemodelan *Random Forest*

Hasil evaluasi model *Random Forest* pada setiap skenario dataset ditunjukkan pada Tabel 2.

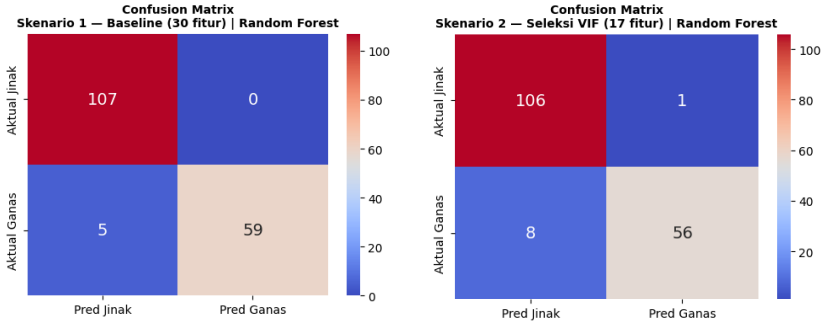
Tabel 2: Evaluasi model *random forest.*

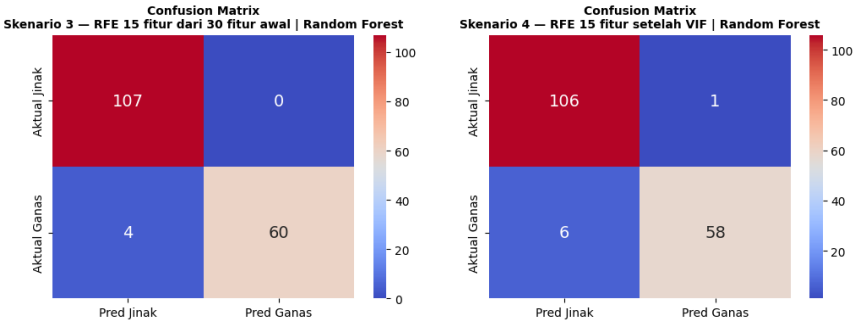
Skenario	Jumlah Fitur	*Accuracy*	*Precision*	*Recall*	*F1-Score*	AUC
1 (*Baseline*)	30	0,9708	1,0000	0,9219	0,9593	0,9966
2 (VIF)	17	0,9474	0,9825	0,8750	0,9256	0,9930
3 (RFE)	**15**	**0,9766**	**1,0000**	**0,9375**	**0,9677**	**0,9966**
4 (VIF+RFE)	15	0,9591	0,9831	0,9062	0,9431	0,9964

Berdasarkan hasil pengujian, performa terbaik model *Random Forest* diperoleh pada skenario RFE menggunakan 15 fitur dengan nilai *accuracy* sebesar 0,9766, *recall* sebesar 0,9375, dan *F1-score* sebesar 0,9677. Hasil tersebut menunjukkan bahwa RFE mampu mempertahankan fitur-fitur paling informatif dalam pembentukan *decision tree* meskipun jumlah fitur berkurang sebesar 50% dari dataset awal.

Berbeda dengan SVM, performa *Random Forest* mengalami penurunan pada skenario VIF serta kombinasi VIF dan RFE. Hasil tersebut menunjukkan bahwa beberapa fitur yang dihapus masih memiliki kontribusi terhadap proses pembentukan *decision tree*, sehingga *Random Forest* cenderung lebih toleran terhadap multikolinearitas dibandingkan SVM.

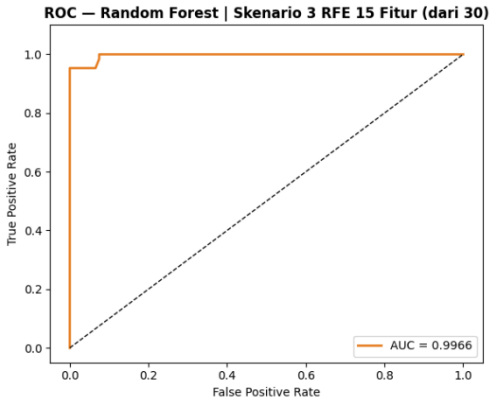
Hasil *confusion matrix* model *Random Forest* pada setiap skenario penelitian ditunjukkan pada Gambar 4.





Gambar 4: *Confusion matrix* model *random forest.*

Berdasarkan hasil *confusion matrix*, jumlah *false negative* pada skenario seleksi fitur menggunakan RFE lebih sedikit dibandingkan skenario lainnya, sehingga model lebih optimal dalam mendeteksi keganasan kanker payudara. Selanjutnya, kurva ROC model *Random Forest* pada setiap skenario ditunjukkan pada Gambar 5. Nilai AUC yang mendekati 1 menunjukkan bahwa model memiliki kemampuan klasifikasi yang sangat baik dalam membedakan kelas jinak dan ganas.



Gambar 5: Kurva ROC model terbaik *random forest.*

3.7 Perbandingan Performa Model

Berdasarkan hasil pengujian, model SVM menunjukkan peningkatan performa setelah dilakukan pengurangan multikolinearitas menggunakan VIF. Hal tersebut menunjukkan bahwa SVM sensitif terhadap redundansi informasi antar fitur karena proses klasifikasi dilakukan dengan membentuk *hyperplane* optimal pada ruang fitur berdimensi tinggi. Sementara itu, model *Random Forest* menghasilkan performa terbaik pada skenario RFE tanpa VIF. Hal tersebut menunjukkan bahwa *Random Forest* lebih efektif dalam memilih fitur-fitur penting berdasarkan proses pembentukan *decision tree* tanpa memerlukan eliminasi multikolinearitas secara eksplisit.

Secara keseluruhan, hasil penelitian menunjukkan bahwa kombinasi seleksi fitur menggunakan VIF dan RFE memberikan performa terbaik pada model SVM, sedangkan metode RFE tanpa pengurangan VIF

menghasilkan performa terbaik pada model *Random Forest*. Selain mampu mengurangi dimensi fitur, kedua metode seleksi fitur tersebut tetap menghasilkan performa klasifikasi yang sangat baik dengan nilai AUC mendekati 1 pada seluruh skenario penelitian.

4. Simpulan dan Saran

Berdasarkan hasil penelitian, metode seleksi fitur menggunakan *Variance Inflation Factor* (VIF) mampu mengurangi multikolinearitas antarfitur, sedangkan *Recursive Feature Elimination* (RFE) mampu mempertahankan fitur-fitur yang memiliki kontribusi penting terhadap proses klasifikasi. Hasil penelitian menunjukkan bahwa pengurangan multikolinearitas menggunakan VIF memberikan hasil yang lebih baik pada model *Support Vector Machine* (SVM), sementara seleksi fitur menggunakan RFE lebih optimal pada model *Random Forest*. Secara keseluruhan, metode seleksi fitur mampu mengurangi jumlah fitur tanpa menurunkan performa klasifikasi secara signifikan.

Penelitian selanjutnya disarankan melakukan optimasi *hyperparameter*, menggunakan dataset dengan jumlah data yang lebih besar, serta memanfaatkan dataset kanker payudara dari rumah sakit atau institusi kesehatan di Indonesia agar hasil penelitian lebih representatif terhadap karakteristik data pasien di Indonesia.

Daftar Pustaka

Asri, J. S., Aryani, D., Fannya, P., & Dewi, R. (2026). Penerapan random forest dan conten-based filtering pada alokasi tenaga kesehatan hipertensi. *Journal of Information System Research.* 7(2): 447–459.

Bustamam, A., Bachtiar, A., & Sarwinda, D. (2019). Selecting features subsets based on support vector machinerecursive features elimination and one dimensional-naïve bayes classifier using support vector machines for classification of prostate and breast cancer. *Procedia Computer Science.*

Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. (2023). Classification prediction of breast cancer based on machine learning. *Computational Inteligencel Neuroscience.*

Dhinakaran, A., Srinivasan, A., Raja, S. E., Valarmathi, K., & Nayagam, M. G. (2025). Synergistic feature selection and distributed classification framework for high-dimensional medical data analysis. *MethodX. 14*.

Ferlay, J., Ervil, M., Lam, F., Laversanne, M., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., & Bray, F. (2024). Global cancer observatory: Cancer today. Prancis: International Agency for Research on Cancer. Tersedia pada: GLOBOCAN Indonesia Fact Sheet.

Géron, A. (2019). Hands-on machine learning with scikit-learn, keras, and tensorflow concepts, tools, and techniques to build intelligent systems second edition.

Hidayat, F., Sholihin, M., & Budi, A. S. (2026). Sistem pengenalan wajah siswa dengan metode you only look once version 3. *Jurnal Publikasi Ilmu Komputer dan Multimedia. 5*(1).

Juarto, B. (2023). Breast cancer classification using outlier detection and variance inflation factor. *Jurnal EMACS: Engineering Mathematics and Computer Science. 5*(1):17–23.

Kumar, S., & Rani, P. (2024). An AI-based liver disease prediction model based on pearson correlation feature selection method. *Biomedical and Pharmacology Journal. 17*(4).

Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practicfal use for clinicians. *Korean Journal of Anesthesiology*. *75*(1). 25–36.

Nugroho, W. M. (2025). Analisis performa algoritma random forest dalam mengatasi overfitting pada model prediksi. *Jurnal Teknologi Informasi dan Komunikasi. 9*(4).

Phudjashakty, N. T., Firmansyah, H., Asriyani, W., & Sofyan, A. (2026). Segmentasi pelanggan grosir menggunakan k-means: analisis outlier dan ketidakseimbangan data. *Jurnal Pengabdian Masyarakat dan Riset Pendidikan. 4*(3).

Rambe, R. Y., Harahap, F. Y., Bihar, S., & Panggabean, Y. C. (2025). Gambaran penderita kanker payudara di bawah usia 40 tahun di RSHAM tahun 2020-2023. *Scripta Score Scientific Medical Journal. 7*(1): 36–41.

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statictics. 8*(2):39–42.

Solang, E. W., Adu, F. X., Dharma, A., & Gunantara, N. (2026). Evaluasi machine learning untuk prediksi pembatalan hotel dengan threshold adjustment dan cost-based evaluation. *Indonesian Journal of Machine Learning and Computer Science. 6*(1).

Tinaliah & Elizabeth, T. (2024). Prediksi jenis kanker payudara menggunakan metode support vector machine berbasis recursive feature elimination. *Jurnal Teknik Informatika dan Sistem Informasi*.

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995) *Breast Cancer Wisconsin* (*Diagnostic*) *Data Set.* UCI Machine Learning Repository.

World Health Organization. 2026. *Breast cancer*. Tersedia pada: WHO Breast Cancer.

Yildirim, H. (2024). The multicollinearity effect on the performance of machine learning algorithms : case examples in healthcare modelling. *Acad Platform Journal Engineering and Smart System.*