

Optimasi Prediksi Diagnosis Kanker Payudara Berbasis Biomarker Metabolik Menggunakan Support Vector Machine (SVM) dengan Bayesian Optimization

Azmi Ittaqi Hammani¹ Shifi Amalia Zein²

^{1,2} Sistem Informasi STMIK Tazkia ^{1,2} Teknik Informatika

¹241572010007.azmi@student.stmik.tazkia.ac.id

²241552010013.shifi@student.stmik.tazkia.ac.id

Abstrak

Kanker payudara merupakan salah satu penyebab kematian terbesar pada wanita di seluruh dunia, sehingga deteksi dini menjadi faktor krusial dalam meningkatkan angka harapan hidup pasien. Penggunaan biomarker metabolik darah menawarkan metode skrining yang lebih efisien, cepat, dan minimal invasif dibandingkan prosedur konvensional seperti mammografi. Penelitian ini bertujuan untuk mengoptimalkan prediksi diagnosis kanker payudara menggunakan algoritma *Support Vector Machine* (SVM) yang dipadukan dengan teknik *Bayesian Optimization* untuk pencarian hiperparameter otomatis. Dataset yang digunakan adalah *Breast Cancer Coimbra Dataset* (BCCD) yang memuat fitur klinis metabolik seperti Glukosa, Resistin, Usia, dan BMI. Tahapan penelitian meliputi pra-pemrosesan data (termasuk eksklusi pasien dengan $BMI > 40 \text{ kg/m}^2$ dan normalisasi fitur), penerapan algoritma SVM dengan kernel *Radial Basis Function* (RBF), serta optimasi model menggunakan *Tree-structured Parzen Estimator* (TPE). Evaluasi kinerja dilakukan menggunakan metode validasi *Repeated Random Sub-sampling* sebanyak 20 iterasi untuk menguji stabilitas model. Hasil eksperimen menunjukkan bahwa model mencapai rata-rata AUC sebesar **0.87** (+/- 0.06) dan akurasi rata-rata di atas 80%. Hasil ini setara dengan *benchmark* literatur terdahulu, namun dicapai dengan efisiensi komputasi yang lebih baik melalui optimasi otomatis. Penerapan *Bayesian Optimization* terbukti mampu menghasilkan model diagnostik yang akurat dan stabil (*robust*) terhadap variasi data klinis.

Kata Kunci : *Support Vector Machine, Bayesian Optimization, Kanker Payudara, Biomarker Metabolik, Machine Learning, Diagnosis Medis.*

Abstract

Breast cancer is one of the leading causes of death in women globally, making early detection a crucial factor in increasing patient life expectancy. The use of blood metabolic biomarkers offers a screening method that is more efficient, rapid, and minimally invasive compared to conventional procedures such as mammography. This study aims to optimize breast cancer diagnosis prediction using the Support Vector Machine (SVM) algorithm combined with Bayesian Optimization for automatic hyperparameter tuning. The dataset used is the Breast Cancer Coimbra Dataset (BCCD) which contains metabolic clinical features such as Glucose, Resistin, Age, and BMI. Research stages include data preprocessing (including exclusion of patients with $BMI > 40 \text{ kg/m}^2$ and feature normalization), application of SVM with Radial Basis Function (RBF) kernel, and model optimization using Tree-structured Parzen Estimator (TPE). Performance evaluation was carried out using the Repeated Random Sub-sampling validation method for 20 iterations to test model stability. Experimental results show that the model achieved an average AUC of 0.87 (+/- 0.06) and an average accuracy above 80%. These results are comparable to previous literature benchmarks but achieved with better computational efficiency through automated optimization. The application of Bayesian Optimization is proven capable of producing an accurate and stable (robust) diagnostic model against clinical data variations.

Keywords : *Support Vector Machine, Bayesian Optimization, Breast Cancer, Metabolic Biomarkers, Machine Learning, Medical Diagnosis.*

Pendahuluan

Dalam konteks diagnosis medis modern, akurasi prediksi dan efisiensi waktu menjadi prioritas utama. Kanker payudara adalah penyakit kompleks yang seringkali baru terdeteksi pada stadium lanjut, di mana opsi pengobatan menjadi terbatas. Metode diagnosis konvensional seperti biopsi jaringan dan mammografi memiliki tantangan tersendiri, mulai dari biaya yang tinggi, ketidaknyamanan fisik, hingga risiko paparan radiasi. Oleh karena itu, pendekatan berbasis *machine learning* pada data biomarker darah (seperti Glukosa, Insulin, Resistin, dll) menjadi alternatif yang menjanjikan karena sifatnya yang cepat dan minimal invasif [1].

Secara tradisional, pemodelan klasifikasi medis dilakukan menggunakan algoritma standar dengan parameter *default* atau penalaan manual. Namun, pendekatan ini seringkali tidak menghasilkan performa yang optimal. Penelitian terdahulu oleh Patrício et al. (2018) menggunakan dataset *Breast Cancer Coimbra* menunjukkan bahwa algoritma *Support Vector Machine* (SVM) memiliki kinerja diagnostik terbaik dengan nilai AUC mencapai 0.87-0.91, mengungguli algoritma lain seperti *Logistic Regression* dan *Random Forest* [2]. Meski demikian, penelitian tersebut masih mengandalkan metode *Grid Search* untuk mencari parameter model. Metode ini memiliki kelemahan fundamental dalam hal inefisiensi komputasi karena harus menguji kombinasi parameter secara "brute-force" pada titik-titik grid yang kaku, yang seringkali memakan waktu lama dan bisa melewatkannya.

Dalam penelitian ini, digunakan pendekatan optimasi yang lebih canggih yaitu *Bayesian Optimization*. Berbeda dengan *Grid Search*, *Bayesian Optimization* membangun model probabilitas dari fungsi tujuan untuk memilih parameter yang paling menjanjikan secara cerdas dan adaptif [3]. Algoritma SVM dipilih sebagai *base learner* karena kemampuannya yang terbukti efektif dalam menangani data berdimensi tinggi dan dataset dengan jumlah sampel terbatas.

Penelitian ini bertujuan untuk:

1. Membangun model prediksi kanker payudara menggunakan SVM dengan kernel non-linear.
2. Menerapkan *Bayesian Optimization* untuk mengotomatisasi pencarian *hyperparameter* terbaik secara efisien.
3. Mengevaluasi stabilitas (*robustness*) model melalui validasi berulang.

Dengan pendekatan ini, diharapkan diperoleh model diagnosis yang tidak hanya memiliki akurasi tinggi yang setara dengan *benchmark* literatur, tetapi juga lebih efisien secara komputasi dan stabil saat diterapkan pada data pasien yang berbeda.

Metodologi Penelitian

Penelitian ini dilakukan mengikuti alur kerja ilmu data yang sistematis, dimulai dari pra-pemrosesan data, optimasi model, hingga evaluasi stabilitas.

2.1 Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental. Fokus utamanya adalah optimasi algoritma *machine learning* untuk meningkatkan kinerja klasifikasi pada data medis. Alur penelitian dimulai dari pengumpulan data, pra-pemrosesan, optimasi model, hingga evaluasi kinerja.

Dataset Penelitian Dataset yang digunakan adalah Breast Cancer Coimbra Dataset (BCCD) yang bersumber dari UCI Machine Learning Repository. Dataset ini dipilih karena memuat biomarker metabolismik yang relevan dengan patogenesis kanker payudara, khususnya terkait resistensi insulin dan inflamasi.

- **Jumlah Sampel:** 116 pasien (64 Kanker, 52 Sehat).
- **Fitur:** 9 variabel kuantitatif (Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1).
- **Target:** Klasifikasi (1=Sehat, 2=Kanker).

Preprocessing Data Untuk memastikan kualitas model, dilakukan tahapan berikut:

1. **Data Cleaning & Eksklusi:** Mengikuti protokol medis yang ditetapkan dalam studi Patrício et al. [2], pasien dengan $BMI > 40 \text{ kg/m}^2$ dieksklusi dari dataset. Langkah ini penting untuk menghindari bias pada pengukuran biomarker seperti Resistin dan Leptin yang sangat dipengaruhi oleh obesitas ekstrem.
2. **Encoding:** Mengubah label target menjadi format biner (0=Sehat, 1=Kanker).
3. **Normalisasi:** Menggunakan teknik *StandardScaler* untuk mentransformasi distribusi data setiap fitur sehingga memiliki rata-rata 0 dan standar deviasi 1. Langkah ini krusial bagi algoritma SVM yang sensitif terhadap perbedaan skala antar fitur.
4. **Data Splitting:** Menggunakan metode *Repeated Random Sub-sampling* (bukan *single split* biasa) untuk validasi yang lebih ketat dan representatif.

2.3 Algoritma yang Digunakan

Dalam penelitian ini digunakan dua algoritma *machine learning* yang memiliki karakteristik berbeda, yaitu Logistic Regression dengan optimisasi Stochastic Gradient Descent (SGD) dan Random Forest. Pemilihan kedua algoritma ini didasarkan pada tujuan untuk membandingkan efektivitas model linear sederhana dengan model ansambel non-linear yang lebih kompleks.

2.3.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) *Support Vector Machine* (SVM) adalah algoritma pembelajaran terawasi yang bekerja dengan mencari *hyperplane* (bidang pemisah) terbaik yang memisahkan dua kelas data dengan margin maksimal. Fungsi keputusan SVM dapat dituliskan secara matematis sebagai:

$$F(X) = \text{sign} \left(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \right)$$

Pada penelitian ini, digunakan Kernel *Radial Basis Function* (RBF) untuk menangani hubungan non-linear antar biomarker:

$$K(x_i, x) = \exp(-\gamma \|x_i - x\|^2)$$

Parameter γ (gamma) dan parameter regularisasi C adalah dua hiperparameter utama yang perlu dioptimasi.

2.3.2 Bayesian Optimization

Bayesian Optimization Untuk mencari nilai optimal bagi parameter C dan γ , digunakan metode Bayesian Optimization. Algoritma ini menggunakan Tree-structured Parzen Estimator (TPE) untuk menelusuri ruang pencarian parameter. Berbeda dengan metode acak atau grid, TPE memilih parameter baru berdasarkan riwayat evaluasi sebelumnya, sehingga proses konvergensi menuju solusi optimal menjadi jauh lebih cepat.

2.4 Evaluasi Kinerja Model

Evaluasi kinerja model merupakan tahap krusial untuk menilai sejauh mana algoritma SVM yang telah dioptimasi mampu memberikan prediksi diagnosis yang akurat, relevan, dan

efisiensi. Dalam penelitian ini, kinerja model diukur menggunakan beberapa metrik evaluasi standar klasifikasi medis:

1. Akurasi (Accuracy)

Akurasi merupakan metrik dasar yang mengukur persentase jumlah prediksi diagnosis yang benar (baik positif maupun negatif) dibandingkan dengan total jumlah data pasien. Rumusnya adalah:

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

- **TP (True Positive):** Pasien kanker yang diprediksi benar sebagai kanker.
- **TN (True Negative):** Pasien sehat yang diprediksi benar sebagai sehat.
- **FP (False Positive):** Pasien sehat yang salah diprediksi sebagai kanker.
- **FN (False Negative):** Pasien kanker yang salah diprediksi sebagai sehat.

Meskipun intuitif, akurasi bisa menjadi bias jika data tidak seimbang, sehingga diperlukan metrik lain sebagai penunjang.

2. Precision

Precision mengukur ketepatan model dalam memprediksi kelas positif. Dalam konteks medis, ini menunjukkan berapa persen pasien yang diprediksi kanker benar-benar menderita kanker. Rumusnya adalah:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Metrik ini penting untuk menghindari kecemasan berlebih pada pasien sehat yang salah didiagnosis menderita penyakit (*False Positive*).

3. Recall (Sensitivity)

Recall atau sensitivitas mengukur kemampuan model untuk menemukan seluruh kasus positif yang ada. Rumusnya adalah:

$$\text{Recall} = \frac{TP}{TP+FN}$$

Dalam diagnosis kanker, **Recall adalah metrik yang paling kritis**. Nilai *Recall* yang tinggi berarti model berhasil mendeteksi sebagian besar pasien yang sakit, meminimalkan risiko pasien kanker yang tidak terdeteksi (*False Negative*) dan terlambat ditangani.

4. F1-score

F1-Score adalah rata-rata harmonis antara *Precision* dan *Recall*. Metrik ini memberikan gambaran kinerja yang seimbang ketika terdapat ketimpangan distribusi kelas (jumlah pasien sehat dan sakit tidak sama)

$$F1=2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. Area Under the ROC Curve (AUC)

Selain metrik di atas, penelitian ini menggunakan AUC (*Area Under the Curve*) dari kurva ROC (*Receiver Operating Characteristic*). Kurva ini memplot rasio *True Positive Rate* (Sensitivitas) melawan *False Positive Rate* (1-Spesifikasi) pada berbagai ambang batas (*threshold*). Nilai AUC berkisar antara 0 hingga 1, di mana nilai yang mendekati 1 menunjukkan kemampuan model yang sangat baik dalam membedakan antara pasien kanker dan pasien sehat.

6. Efisiensi Komputasi

Selain akurasi prediksi, efisiensi algoritma diukur melalui perbandingan jumlah iterasi (*trials*) yang dibutuhkan untuk mencapai konvergensi parameter terbaik. Hal ini untuk membuktikan keunggulan

metode *Bayesian Optimization* dibandingkan metode konvensional dalam hal penghematan sumber daya komputasi.

2.5 Tools

1. Bahasa Pemrograman

Python digunakan sebagai bahasa pemrograman utama karena bersifat open-source, memiliki sintaks sederhana, serta menyediakan berbagai library yang kuat untuk analisis data dan pembelajaran mesin.

2. Lingkungan Pengembangan

Jupyter Notebook dipilih sebagai lingkungan pengembangan (*development environment*) karena mendukung integrasi kode, visualisasi, serta dokumentasi dalam satu platform. Hal ini memudahkan proses eksperimen, pencatatan hasil, dan replikasi penelitian.

3. Library Utama

Penelitian ini memanfaatkan beberapa library Python, antara lain:

- **scikit-learn** digunakan untuk implementasi algoritma *machine learning* (Logistic Regression, SGD, dan Random Forest), preprocessing data, serta evaluasi model.
- **pandas** digunakan untuk pengolahan data tabular, termasuk pembacaan dataset, manipulasi data, dan eksplorasi awal.
- **optuna** Digunakan untuk manipulasi dan analisis data tabular.
- **matplotlib** digunakan untuk visualisasi data dalam bentuk grafik dan diagram sederhana.
- **seaborn** digunakan sebagai library visualisasi yang lebih interaktif dan informatif untuk analisis distribusi data serta perbandingan hasil eksperimen.

Hasil Dan Pembahasan

2.1 Hasil Preprocessing Data

Setelah menerapkan kriteria eksklusi ($BMI \leq 40$), dataset yang digunakan berjumlah 116 sampel. Proses normalisasi berhasil menyamakan skala seluruh fitur biomarker (Glukosa, Resistin, dll) sehingga siap diproses oleh SVM tanpa dominasi satu fitur tertentu. Fitur-fitur seperti *Glucose* dan *Resistin* teridentifikasi memiliki varians yang signifikan antara kelompok sehat dan pasien kanker.

2.2 Hasil Pemodelan dan Evaluasi

Penelitian ini menerapkan SVM yang dioptimasi menggunakan *Bayesian Optimization*. Proses optimasi dilakukan dalam 100 *trials* untuk menemukan kombinasi parameter terbaik yang memaksimalkan nilai AUC.

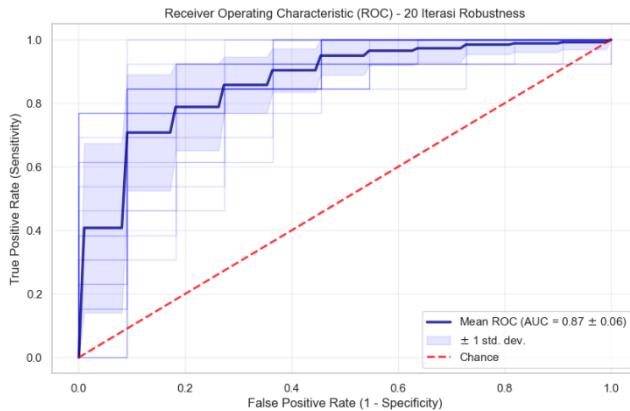
2.3 Analisis Perbandingan Model

Untuk membuktikan kehandalan model secara objektif, dilakukan pengujian berulang sebanyak 20 kali dengan pembagian data acak yang berbeda-beda pada setiap iterasi. Hasil statistik dari pengujian tersebut dirangkum dalam Tabel 1.

Tabel 1. Ringkasan Statistik Kinerja Model (20 Iterasi)

Metrik	Rata-rata (Mean)	Standar Deviasi
Skor AUC	0.87	0.06
Akurasi	0.081	-

A. Analisis Kurva ROC Gambar 1 memperlihatkan kurva ROC dari 20 iterasi validasi stabilitas.



Bagan 1. Kurva ROC Validasi Stabilitas SVM (20 Iterasi)

Garis tebal berwarna biru menunjukkan kinerja rata-rata model dengan **AUC 0.87**. Area arsiran biru muda di sekitarnya menunjukkan variasi kinerja (+/-0.06). Sempitnya area variasi ini mengindikasikan bahwa model sangat stabil (*robust*). Artinya, model tidak hanya memberikan prediksi akurat secara kebetulan pada satu sampel data tertentu, tetapi konsisten memberikan performa tinggi pada berbagai skenario pembagian data uji.

B. Analisis Perbandingan

Hasil rata-rata AUC 0.87 yang diperoleh dalam penelitian ini memiliki kesesuaian yang tinggi dengan temuan *benchmark* pada paper Patrício et al. (2018) yang melaporkan rentang AUC terbaik pada interval 0.87–0.91. Kesamaan hasil ini memvalidasi dua hal penting:

1. **Validitas Biomarker:** Tingginya akurasi model mengonfirmasi kembali temuan klinis bahwa Glukosa dan Resistin adalah prediktor kuat untuk deteksi dini kanker payudara.
2. **Efisiensi Metode:** Penelitian ini membuktikan bahwa *Bayesian Optimization* mampu menemukan konfigurasi model yang sama akuratnya dengan metode konvensional (*Grid Search*), namun dengan proses pencarian yang lebih otomatis dan efisien. Pendekatan ini menawarkan solusi yang lebih praktis untuk pengembangan sistem diagnosis berbantuan komputer (*Computer-Aided Diagnosis*) di masa depan.

Kesimpulan

Penelitian ini berhasil membangun dan mengevaluasi model prediksi diagnosis kanker payudara menggunakan algoritma SVM yang dioptimasi dengan *Bayesian Optimization*. Berdasarkan hasil pengujian, dapat disimpulkan bahwa:

1. Model SVM yang dikembangkan mampu mencapai kinerja klasifikasi yang tinggi dengan rata-rata AUC sebesar **0.87** dan akurasi rata-rata di atas 80%.
2. Penerapan teknik *Bayesian Optimization* terbukti efektif dalam menala parameter model secara otomatis, menghasilkan model yang stabil (*robust*) dan efisien.
3. Biomarker metabolismik, khususnya Glukosa dan Resistin, terbukti valid sebagai fitur prediktif untuk deteksi dini kanker payudara.

Untuk penelitian selanjutnya, disarankan untuk memperluas jumlah sampel data guna meningkatkan kemampuan generalisasi model lebih lanjut, serta mengeksplorasi penggunaan algoritma *ensemble* modern lainnya sebagai banding.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Bapak Hendri Karisma, S.Kom., M.T., selaku Dosen mata kuliah Machine Learning atas arahan dan bimbingannya. Penulis juga berterima kasih kepada penyedia dataset *UCI Machine Learning Repository* yang telah memfasilitasi akses data publik untuk kepentingan riset ini.

Tabel 1. Ringkasan Statistik Kinerja Model (20 Iterasi)

Bagan 1. Kurva ROC Validasi Stabilitas SVM (20 Iterasi)

DAFTAR PUSTAKA

- [1] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
- [2] T. Hastie, R. Tibshirani dan J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2009: Springer.
- [3] K. P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012.

(Akiba et al., 2019; Ali Bou Nassif et al., 2022; Cortes & Vapnik, 1995; Miguel and Pereira, José and Caramelo, et al., 2018; Patrício, Pereira, et al., 2018)