OPTIMASI K-NN DENGAN *BAYES SEARCH* CV UNTUK KLASIFIKASI KANKER PAYUDARA

Toni Arifin[1], Ilham Rachmat Wibowo[2], Ignatius Wiseto Prasetyo Agung[3], Erfian Juniation[4]

[1,2,3,4]ARS Digital Research and Innovation (ADRI)
[1,2,3] Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Adhirajasa Reswara Sanjaya, Jl. Sekolah Internasional No.1-2, Antapani, Kota Bandung, Jawa Barat, Indonesia.
[4] Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Adhirajasa Reswara Sanjaya, Jl. Sekolah Internasional No.1-2, Antapani, Kota Bandung, Jawa Barat, Indonesia.

[1] toni.arifin@ars.ac.id, [2] ihamwibowo125@gmail.com, [3] wiseto.agung@ars.ac.id, [4] erfian.ejn@ars.ac.id

Abstrak

Kanker payudara merupakan salah satu penyebab utama mortalitas global pada wanita dan menjadi tantangan serius dalam bidang kesehatan. Sifat penyakit yang heterogen menuntut adanya metode klasifikasi yang akurat dan andal guna mendukung proses diagnosis serta penentuan strategi terapi yang tepat. Penelitian ini bertujuan untuk mengoptimalkan kinerja algoritma *K-Nearest Neighbors* (KNN) melalui penerapan metode *Bayes Search Cross Validation* (CV) dalam meningkatkan akurasi klasifikasi berbasis data ekspresi gen messenger *Ribonucleic Acid* (mRNA). Metodologi penelitian meliputi proses pemodelan KNN yang diintegrasikan dengan optimasi hyperparameter menggunakan *Bayes Search* CV pada dataset *Breast Cancer Gene Expression* Profiles (METABRIC) yang terdiri dari 1.904 sampel dan 692 atribut, mencakup data ekspresi gen dan karakteristik klinis pasien. Tahapan penelitian mencakup pengolahan data awal, pembagian data latih dan uji, proses optimasi, serta evaluasi model menggunakan metrik akurasi. Hasil penelitian menunjukkan bahwa penerapan *Bayes Search* CV mampu meningkatkan akurasi klasifikasi secara signifikan menjadi 78,68%, dibandingkan dengan model dasar tanpa optimasi yang hanya mencapai 66,81%. Temuan ini mengindikasikan bahwa pemilihan hyperparameter yang optimal berkontribusi besar terhadap peningkatan performa model. Dengan demikian, dapat disimpulkan bahwa pendekatan optimasi berbasis *Bayes Search* CV efektif dalam meningkatkan kinerja algoritma KNN pada data genetik yang kompleks serta berpotensi mendukung pengembangan sistem pendukung keputusan klinis yang lebih akurat, personal, efisien, adaptif, dan relevan dalam mendukung implementasi *Precision* medicine pada diagnosis kanker payudara modern.

Kata kunci: kanker payudara, *k-nearest neighbors*, *bayes search cross validation*, ekspresi gen mRNA, klasifikasi, optimasi hyperparameter.

1. Pendahuluan

Kanker payudara merupakan salah satu penyakit paling umum pada wanita di seluruh dunia yang telah menyebabkan kematian sekitar 570.000 jiwa pada tahun 2015. Setiap tahunnya, lebih dari 1,5 juta wanita atau setara dengan 25% dari total kasus kanker pada perempuan didiagnosis mengidap penyakit ini. Laporan statistik dari *World Health Organization* (WHO) menunjukkan bahwa prevalensi kanker payudara sangat tinggi, terutama pada perempuan di negara-negara maju (Sun et al., 2017). Tingginya angka kasus tersebut menjadikan kanker payudara sebagai salah satu penyakit yang memerlukan penanganan serta deteksi dini secara serius guna menekan tingkat kematian pasien. Secara klinis, sel kanker payudara umumnya membentuk massa tumor yang dapat diidentifikasi melalui pemeriksaan sinar-X atau teraba sebagai benjolan (Rajaguru & Chakravarthy, 2019). Sel-sel ganas tersebut dapat bermetastasis ke bagian tubuh lainnya melalui sirkulasi darah atau sistem limfatik akibat adanya mutasi dan perubahan pada struktur DNA (Rawal, 2020).

Sebagai upaya untuk menurunkan angka mortalitas akibat kanker payudara, berbagai penelitian telah membuktikan bahwa implementasi metode machine learning mampu mengoptimalkan efektivitas diagnosis dan deteksi dini (Abunasser et al., 2023). Pemanfaatan teknologi ini menjadi penting karena metode *machine learning* mampu membantu proses analisis data medis secara lebih cepat dan akurat dibandingkan pendekatan konvensional. Teknik ini memiliki keunggulan dalam mempelajari model data secara otomatis tanpa memerlukan asumsi implisit, serta mampu menangani interdependensi dan hubungan *non-linear* yang kompleks antar variabel (Takeshita et al., 2023). Kemampuan tersebut menjadikan machine learning efektif dalam menemukan pola tersembunyi pada data medis yang sulit diidentifikasi secara manual. Dengan memanfaatkan pola tersebut, *machine learning* dapat digunakan untuk memprediksi probabilitas

keberhasilan penanganan kasus baru secara lebih akurat (Boeri et al., 2020).

Machine learning merupakan cabang dari kecerdasan buatan yang berfokus pada pengembangan algoritma serta metode untuk keperluan klasifikasi, prediksi, dan analisis citra (Ranti et al., 2022). Teknologi ini banyak digunakan karena mampu membantu proses pengolahan data secara otomatis dan efisien dalam berbagai bidang, termasuk kesehatan. Algoritma *machine learning* dirancang agar mampu mengekstraksi pengetahuan dari dataset historis melalui pemrosesan data dalam jumlah besar sehingga model dapat melakukan analisis mendalam untuk menghasilkan prediksi yang akurat (Fatima et al., 2020).

Salah satu teknik klasifikasi fundamental yang sering diimplementasikan dalam diagnosis kanker karena kesederhanaannya adalah *K-Nearest Neighbors* (KNN) (Khorshid & Abdulazeez, 2021). Sebagai teknik non-parametrik, KNN bekerja dengan mengkalkulasi metrik jarak antara data input dan data pengujian guna menentukan prediksi kelas yang tepat berdasarkan kedekatan antar-instansi data (Houfani et al., 2020). Namun, kinerja KNN sangat dipengaruhi oleh pemilihan parameter, seperti nilai *k* dan fungsi jarak yang digunakan (Ubaidillah et al., 2022). Selain itu, keberadaan data yang bersifat *noisy* menjadi keterbatasan yang signifikan, karena metode ini hanya mempertimbangkan k tetangga terdekat dari data uji. Kondisi tersebut dapat memengaruhi pembentukan konsep *neighborhood* sehingga berdampak pada penurunan akurasi model (Ozturk Kiyak et al., 2023).

Untuk mengatasi permasalahan tersebut, dapat diterapkan teknik hyperparameter tuning yang bertujuan memperoleh kombinasi parameter optimal pada model (Rahman et al., 2023). Teknik ini merupakan proses optimisasi untuk meminimalkan kesalahan model *machine learning* melalui penyesuaian hyperparameter, dengan membandingkan kinerja model menggunakan parameter default dan hasil tuning. Berbagai penelitian menunjukkan bahwa pendekatan ini mampu meningkatkan akurasi pada kasus klasifikasi (Elgeldawi et al., 2021). Salah satu metode yang umum digunakan dalam optimasi hyperparameter adalah *Bayes Search* CV, yaitu teknik optimasi berbasis probabilistik yang memodelkan ruang pencarian untuk memperoleh parameter optimal secara efisien (Nisanova et al., 2024).

Bayes Search CV merupakan metode yang menggabungkan cross-validation dan *Bayesian Optimization* untuk memilih kombinasi hyperparameter yang optimal dari ruang pencarian, kemudian memperbarui pemilihan parameter secara iteratif berdasarkan hasil evaluasi sebelumnya (Zhao et al., 2024). *Bayesian Optimization* digunakan karena mampu mempercepat proses pelatihan sekaligus meningkatkan kualitas model (Zhao et al., 2024). Dalam penelitian ini, metode *Bayes Search* CV diterapkan untuk mengatasi permasalahan dalam penentuan hyperparameter optimal pada algoritma. Nilai fungsi objektif digunakan sebagai dasar dalam menentukan kombinasi hyperparameter terbaik, dengan memanfaatkan prinsip optimasi maksimum. Pendekatan ini digunakan untuk mengestimasi nilai optimal hyperparameter berdasarkan data ekspresi gen pada pasien kanker payudara (Babichev et al., 2024).

Identifikasi kanker melalui analisis ekspresi gen yang dikombinasikan dengan model *machine learning* merupakan pendekatan yang menjanjikan (Takeshita et al., 2023). Penelitian sebelumnya menunjukkan bahwa model *machine learning* untuk klasifikasi kanker payudara masih memiliki keterbatasan, terutama terkait potensi bias pada model (Thalor et al., 2022). Hal ini disebabkan oleh kompleksitas data ekspresi gen yang sulit dianalisis, sehingga dapat menurunkan akurasi model.

Analisis data ekspresi gen memerlukan pendekatan yang melibatkan struktur kombinatorial, sistem dinamis, algoritma, serta teknik estimasi dan optimasi. Meskipun demikian, penggunaan algoritma *machine learning* berpotensi mengurangi bias apabila dataset dianalisis secara cermat. Oleh karena itu, diperlukan penelitian lebih lanjut untuk meningkatkan akurasi identifikasi kanker payudara berbasis ekspresi gen melalui analisis yang lebih komprehensif dan pemilihan parameter yang tepat dalam menangani kompleksitas data (Dagadu et al., 2025).

Berdasarkan penelitian terdahulu, berbagai metode *Machine learning* telah banyak diterapkan untuk klasifikasi kanker payudara menggunakan data *Gene Expression* dan data klinis pasien. Penelitian sebelumnya menunjukkan bahwa optimasi hyperparameter mampu meningkatkan performa model klasifikasi kanker payudara berbasis mRNA, dengan algoritma *Decision Tree* memperoleh akurasi tertinggi sebesar 99,73%, sedangkan *K-Nearest Neighbors* (KNN) memperoleh akurasi 79,63% (Arifin et al., 2024). Selain itu, algoritma K-Nearest Neighbors (KNN) memiliki performa klasifikasi yang baik pada dataset kanker payudara dengan tingkat akurasi di atas 94%, meskipun masih berada di bawah beberapa metode *ensemble* dan Naïve Bayes (Bhat et al., 2026). Hasil tersebut menunjukkan bahwa KNN tetap menjadi algoritma yang kompetitif dalam proses klasifikasi data medis karena memiliki kemampuan mengenali pola berdasarkan kedekatan antar data. Penelitian lain juga menunjukkan bahwa KNN mampu digunakan untuk klasifikasi tipe kanker payudara berbasis Gene Expression dengan akurasi mencapai 87% (Rasheda et al., 2022).

Beberapa penelitian memanfaatkan metode optimasi dan feature selection untuk meningkatkan performa klasifikasi kanker payudara. Kombinasi metode *Support Vector Machine* (SVM) dengan Backward Elimination berhasil meningkatkan akurasi dari 65,22% menjadi 95,65% (Resmiati & Arifin,

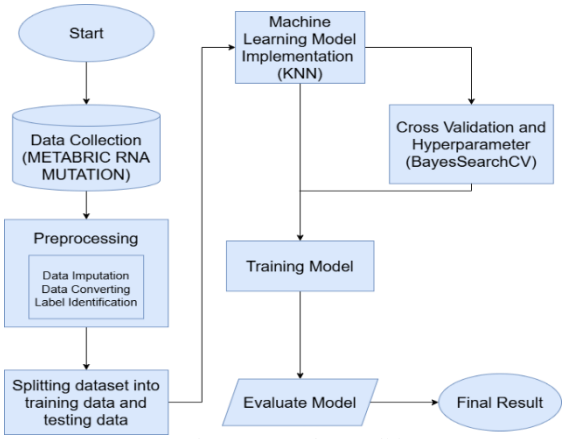
2021). Hasil tersebut menunjukkan bahwa proses seleksi fitur mampu membantu model dalam memilih atribut yang paling relevan sehingga performa klasifikasi menjadi lebih optimal. Penelitian lain juga menemukan bahwa penggunaan teknik *feature selection seperti Information Gain with Grey Wolf Optimization* (IG-GWO) mampu meningkatkan akurasi SVM hingga 94,87%, sedangkan KNN mencapai 91,96% (Elnaby et al., 2021). Selain itu, penerapan *Bayesian Optimization* pada model klasifikasi berbasis *Gene Expression* memperoleh Balanced *Accuracy* sebesar 97,31%, sementara model KNN juga menunjukkan performa tinggi dengan akurasi mendekati 97,4% (Phan et al., 2021).

Penelitian terbaru mulai banyak menerapkan pendekatan *Bayesian Optimization* atau *Bayes Search* CV untuk meningkatkan performa model pada data genetik berdimensi tinggi. Metode optimasi ini dinilai efektif karena mampu menemukan kombinasi hyperparameter terbaik secara lebih efisien dibandingkan metode pencarian konvensional. Penelitian menunjukkan bahwa *Bayesian Optimization* mampu meningkatkan akurasi model klasifikasi RNA Expression hingga mencapai lebih dari 99% (Alanazi et al., 2024). Hasil tersebut menunjukkan bahwa proses optimasi hyperparameter memiliki peran penting dalam meningkatkan kemampuan model dalam mengelola data genetik yang kompleks dan berdimensi tinggi. Penelitian lain juga menjelaskan bahwa *Bayesian Optimization* efektif digunakan pada algoritma seperti SVM, KNN, dan *Deep Learning* untuk mengatasi kompleksitas data genetik dengan jumlah atribut yang besar (Alharbi & Vakanski, 2023). Berdasarkan penelitian-penelitian tersebut, dapat disimpulkan bahwa optimasi hyperparameter menggunakan *Bayes Search* CV memiliki potensi besar dalam meningkatkan performa algoritma KNN pada klasifikasi kanker payudara berbasis data ekspresi gen.

2. Metode

Tahapan metodologi penelitian secara komprehensif dipaparkan dalam kerangka penelitian pada Gambar 1. Gambar tersebut mengilustrasikan langkah-langkah sistematis yang dilakukan untuk mengimplementasikan model *K-Nearest Neighbors* (KNN) dan optimasi *Bayes Search Cross Validation* (CV) pada data ekspresi mRNA dari dataset METABRIC. Alur penelitian dimulai dengan pengumpulan data yang dilanjutkan ke tahap pra-pemrosesan. Selanjutnya, data dipartisi menjadi set pelatihan (*training*) dan pengujian (*testing*), diikuti oleh proses pelatihan KNN serta konfigurasi hyperparameter melalui metode *Bayes Search* CV.

Tahap akhir melibatkan evaluasi model untuk memperoleh hasil performa yang definitif. Sementara itu, Tabel 1 merinci atribut klinis dari dataset mRNA kanker payudara METABRIC yang memberikan konteks fundamental dalam menganalisis data ekspresi gen serta mendukung pengembangan model *machine learning* yang lebih presisi (Arifin et al., 2024a).



Gambar 1. Kerangka penelitian

Berdasarkan studi literatur terdahulu, implementasi algoritma *K-Nearest Neighbors* (KNN) mencatatkan tingkat akurasi sebesar 79,63%. Penelitian ini bertujuan untuk mengoptimalkan akurasi identifikasi kanker payudara melalui integrasi algoritma KNN dengan metode *Bayes Search Cross Validation* (CV). Pengembangan model ini diharapkan dapat memberikan kontribusi signifikan bagi para klinisi dalam melakukan diagnosis dini yang lebih presisi.

Tahapan awal penelitian difokuskan pada pengumpulan data menggunakan dataset *Breast Cancer Messenger* RNA yang diperoleh dari *Molecular Taxonomy of Breast Cancer International Consortium* (https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric/data). Basis data ini mencakup skor-z tingkat mRNA untuk 331 gen pada 1.904 pasien kanker payudara, yang dilengkapi dengan 31 karakteristik klinis, pengobatan, serta data demografi. Secara komprehensif, dataset ini memiliki 692 atribut dan 1.904 baris data (Arifin et al., 2025).

Tabel 1. Daftar Atribut Klinis Dataset Breast Cancer *Gene Expression* Profiles (METABRIC)

Nama	Keterangan
patient_id	Identitas pasien.
age_at_diagnosis	Usia pasien saat dilakukan diagnosa.
type_of_breast_surgery	Jenis operasi pada kanker.
cancer_type	Jenis kanker payudara.
cancer_type_detailed	Jenis kanker payudara terperinci.
Cellularity	Jumlah sel kanker sesudah kemoterapi.
Chemotherapy	Riwayat pasien menjalani kemoterapi.
pam50_+_claudin-low_subtype cohort	Sub tipe kanker payudara berdasarkan tes pam 50 dan subjek

	kelompok claudin-low dengan karakteristik tertentu.
er_status_measured_by_ihc	Sub tipe kanker payudara berdasarkan tes pam 50 dan subjek kelompok claudin-low dengan karakteristik tertentu.
er_status	Status reseptor estrogen diukur dengan imunohistokimia.
neoplasm_histologic_grade	Status reseptor estrogen.
her2_status_measured_by_snp6	Tingkat agresivitas sel kanker.
her2_status	Status HER2 diukur dengan teknik molekuler lanjut. Status HER2.
tumor_other_histologic_subtype	Sub tipe tumor histologis.
hormone_therapy	Riwayat terapi hormon.
inferred_menopausal_state	Status pasien monofos.
integrative_cluster	Sub tipe molekuler kanker berdasarkan ekspresi gen.
primary_tumor_laterality	Lokasi tumor primer.
lymph_nodes_examined_positive	Jumlah kelenjar getah bening.
mutation_count	Jumlah gen yang bermutasi.
nottingham_prognostic_index	Indek prognostik nottingham.
oncotree_code	Kode oncotree untuk diagnosis kanker.
overall_survival_months	Durasi pasien yang bertahan keseluruhan.
pr_status	Status reseptor progesteron.
radio_therapy	Riwayat radioterapi.
3-gene_classifier_subtype	Sub tipe kanker berdasarkan klasifikasi 3 gen.
tumor_size	Ukuran tumor.
tumor_stage	Tahapan tumor.
death_from_cancer	Riwayat kematian sebab kanker.

Tabel 1 menjelaskan atribut klinis yang digunakan pada dataset *Breast Cancer Gene Expression Profiles* (METABRIC). Dataset ini memuat informasi penting terkait karakteristik pasien kanker payudara yang digunakan dalam proses analisis maupun pemodelan *machine learning*. Informasi klinis pada dataset tersebut dapat membantu dalam memahami kondisi pasien serta mendukung proses analisis hubungan antara faktor klinis dan hasil prediksi penelitian.

Tahap kedua adalah pra-pemrosesan (*preprocessing*) yang alurnya disajikan pada Gambar 2 hingga Gambar 6. Pada tahap ini, dilakukan imputasi data yang bertujuan untuk menangani nilai yang hilang (*missing values*) guna menjaga integritas serta kelengkapan dataset yang digunakan dalam penelitian(Alromema et al., 2023). Proses berikutnya adalah konversi data menjadi format numerik (*data converting*), sebagaimana diilustrasikan pada Gambar 7 hingga Gambar 9, yang bertujuan agar data dapat diproses secara optimal oleh algoritma *machine learning*. Tahap akhir dari pra-pemrosesan adalah identifikasi label (*label identification*) untuk menentukan target klasifikasi, sehingga model mampu mengidentifikasi status diagnosis kanker pada pasien.

```
[ ]  # Mengubah Data NaN menjadi 0
     dg = dg.fillna(0)
```

Gambar 2. Koding Imputasi fillna

```
[ ] dg.head()
```

	patient_id	age_at_diagnosis	type_of_breast_surgery	cancer_type	cancer_type_detailed	cellularity	c
0	0	75.65	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	NaN	
1	2	43.19	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High	
2	5	48.87	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	High	
3	6	47.68	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Moderate	
4	8	76.97	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	High	

5 rows × 693 columns

Gambar 3. Data Sebelum Imputasi

```
dg.head()
```

	patient_id	age_at_diagnosis	type_of_breast_surgery	cancer_type	cancer_type_detailed	cellularity
0	0	75.65	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	0
1	2	43.19	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High
2	5	48.87	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	High
3	6	47.68	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Moderate
4	8	76.97	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	High

5 rows × 693 columns

Gambar 4. Data Sesudah Imputasi

```
[ ]  # Memilih kolom-kolom dengan tipe data 'object' (kategori)
     kolom_kategorikal = dg.select_dtypes(include=['object']).columns

     print("Kolom-kolom kategorikal:")
     print(kolom_kategorikal)
```

Gambar 5. Koding Identifikasi Fitur Kategori

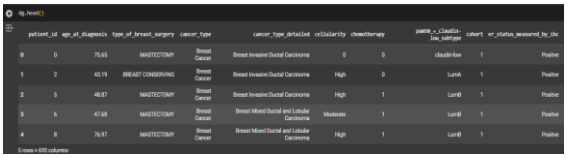
```
Kolom-kolom kategorikal:
Index(['type_of_breast_surgery', 'cancer_type', 'cancer_type_detailed',
       'cellularity', 'pam50_+_claudin-low_subtype',
       'er_status_measured_by_ihc', 'er_status',
       'her2_status_measured_by_snp6', 'her2_status',
       'tumor_other_histologic_subtype', 'inferred_menopausal_state',
       'integrative_cluster', 'primary_tumor_laterality', 'oncotree_code',
       'pr_status', '3-gene_classifier_subtype', 'death_from_cancer',
       'ryr2_mut'],
      dtype='object')
```

Gambar 6. Hasil Identifikasi Fitur Kategori

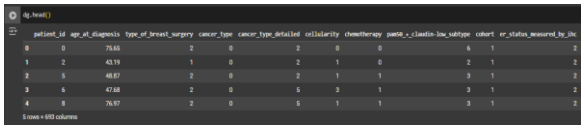
```
# Konversi semua kolom kategorikal ke kode numerik dalam satu loop
categorical_columns = [
    'type_of_breast_surgery', 'cancer_type', 'cancer_type_detailed',
    'cellularity', 'her2_status', 'er_status_measured_by_ihc',
    'er_status', 'inferred_menopausal_state', 'primary_tumor_laterality',
    'pr_status', 'death_from_cancer', 'pam50_+_claudin-low_subtype',
    'her2_status_measured_by_snp6', 'tumor_other_histologic_subtype',
    'integrative_cluster', 'oncotree_code', '3-gene_classifier_subtype',
    'ryr2_mut'
]

for col in categorical_columns:
    if dg[col].dtype == 'object':
        dg[col] = dg[col].astype('category').cat.codes
```

Gambar 7. Transformasi Fitur Kategori ke Numerik



Gambar 8. Hasil Sebelum Transformasi



Gambar 9. Hasil Sesudah Transformasi

Tahap berikutnya adalah pembagian data (data splitting) sebagaimana diilustrasikan pada Gambar 10 hingga Gambar 12. Proses ini melibatkan pemisahan dataset menjadi set pelatihan (*training*) dan pengujian (*testing*). Data pelatihan digunakan untuk mengestimasi parameter model yang tidak diketahui atau melakukan penyesuaian (*fitting*) model, sementara tingkat keakuratan model dievaluasi menggunakan data pengujian (Arifin et al., 2024a). Dalam penelitian ini, dataset yang berjumlah 1.904 *records* dipartisi secara seimbang dengan rasio 50:50, sehingga diperoleh 952 sampel untuk data latih dan 952 sampel untuk data uji. Pembagian ini dilakukan untuk memastikan bahwa model dapat divalidasi secara objektif sebelum diimplementasikan lebih lanjut.

```
x = dg.iloc[:, :-1] #semua data kecuali kolom terakhir
y = dg['cancer_type_detailed']
```

Gambar 10. Konfigurasi Pemilihan Variabel x dan y

```
[12] from sklearn.model_selection import train_test_split
     X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.5, random_state=42)
```

Gambar 11. Kode Pembagian Data

```
[13] from sklearn.preprocessing import StandardScaler
     scaler = StandardScaler()
     scaler.fit(X_train)
     X_train_scaled = scaler.transform(X_train)
     X_test_scaled = scaler.transform(X_test)
```

Gambar 12. Kode Menstandarisasi Data

Tahapan selanjutnya adalah implementasi model *machine learning* menggunakan algoritma *K-Nearest Neighbors* (KNN) dan optimasi hyperparameter melalui metode *Bayes Search Cross Validation* (CV), yang prosedurnya diilustrasikan pada Gambar 13, 14, dan 15. Metode *Bayes Search* CV bekerja secara iteratif dengan menentukan konfigurasi titik evaluasi berikutnya berdasarkan hasil observasi sebelumnya, proses ini terus berlanjut hingga diperoleh kombinasi parameter yang paling optimal (Ubaidillah et al., 2022). Setelah parameter terbaik teridentifikasi, dilakukan tahap pelatihan ulang (re*training*) pada model KNN menggunakan konfigurasi tersebut untuk mencapai tingkat akurasi yang maksimal dalam mengklasifikasikan data.

```
# Define parameter space
param= {'n_neighbors': [1, 2, 4, 5, 6, 7, 8, 9],
        'p': [1, 2],
        'metric': ['minkowski', 'manhattan', 'euclidean'],
        'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
        'leaf_size': [10, 20, 30, 40, 50]
       }
```

Gambar 13. Mengaplikasikan Parameter *K-Nearest Neighbors*

```
# Initialize BayesSearchCV
bayes = BayesSearchCV(knn_bayes,
                      param,
                      n_iter=10,
                      cv=kf,
                      random_state=42)
```

Gambar 14. Mengaplikasikan *Bayes Search* CV

```
[10] from sklearn.model_selection import KFold, StratifiedKFold, cross_val_score, cross_val_predict
     kf =StratifiedKFold(n_splits=10, shuffle=True, random_state=42)

     cnt = 1
     # split()  method generate indices to split data into training and test set.
     for train_index, test_index in kf.split(x, y):
         print(f'Fold:{cnt}, Train set: {len(train_index)}, Test set:{len(test_index)}')
         cnt += 1

Fold:1, Train set: 1713, Test set:191
Fold:2, Train set: 1713, Test set:191
Fold:3, Train set: 1713, Test set:191
Fold:4, Train set: 1713, Test set:191
Fold:5, Train set: 1714, Test set:190
Fold:6, Train set: 1714, Test set:190
Fold:7, Train set: 1714, Test set:190
Fold:8, Train set: 1714, Test set:190
Fold:9, Train set: 1714, Test set:190
Fold:10, Train set: 1714, Test set:190
```

Gambar 15. Kode Cross Validation dengan StratifiedKFold

K-Nearest Neighbors (KNN) merupakan algoritma *supervised learning* yang mengklasifikasikan query instance baru berdasarkan kategori mayoritas dari tetangga terdekatnya. Tujuan utama algoritma ini adalah menentukan label kelas objek baru dengan menganalisis kemiripan atribut terhadap data pelatihan, di mana setiap sampel pengujian diklasifikasikan berdasarkan konsensus kategori dalam ruang fitur (Ozturk Kiyak et al., 2023). Untuk mengoptimalkan kinerja model, digunakan *Bayes Search* CV, sebuah metode optimasi dalam kerangka *Bayesian Optimization* yang menerapkan algoritma Sequential *Model-Based Optimization* (SMBO) (Qiu & Liu, 2025).

Struktur SMBO melibatkan mekanisme inner loop yang mencakup optimasi numerik melalui model pengganti (surrogate model). Dalam proses ini, titik x^* yang memaksimalkan fungsi pengganti diusulkan sebagai lokasi evaluasi fungsi objektif f yang sebenarnya. Algoritma SMBO secara dinamis mengoptimalkan kriteria tertentu untuk menentukan nilai x berdasarkan estimasi model f dan akumulasi riwayat observasi H (Qiu & Liu, 2025).

Tahap evaluasi dilakukan untuk mengukur efektivitas model melalui *Confusion Matrix*, yang menghasilkan nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score* sebagai indikator performa akhir (Mudzakir & Arifin, 2022). Komponen fundamental dalam *Confusion Matrix* didefinisikan sebagai berikut:

- *True Positive* (TP): Jumlah data positif yang diprediksi secara benar oleh model.
- *True Negative* (TN): Jumlah data negatif yang diprediksi secara benar oleh model.

- *False Positive* (FP): Jumlah data negatif yang salah diprediksi sebagai positif oleh model.
- *False Negative* (FN): Jumlah data positif yang salah diprediksi sebagai negatif oleh model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{FP+TP} \quad (2)$$

$$Recall = \frac{TP}{FN+TP} \quad (3)$$

$$F1\,Score = 2 * \frac{Precision * Recall}{Precision+Recall} \quad (4)$$

3. Hasil dan Pembahasan

Hasil penelitian ini menyajikan perbandingan kinerja algoritma *K-Nearest Neighbors* (KNN) sebelum dan sesudah penerapan hyperparameter tuning, dengan menampilkan nilai evaluasi berupa *Accuracy, Recall, Precision*, dan *F1-Score*. Parameter yang digunakan dalam proses tuning disajikan pada Tabel 2, sedangkan Tabel 3 dan Tabel 4 menunjukkan perbandingan hasil evaluasi sebelum dan sesudah penerapan hyperparameter tuning pada KNN. Penerapan metode *Bayes Search* CV pada KNN bertujuan untuk memperoleh kombinasi parameter yang optimal sehingga dapat meningkatkan kinerja model, khususnya dalam hal akurasi.

Tabel 2. Parameter *K-Nearest Neighbors* dengan *Bayes Search CV*

Parameter	Nilai
n_neighbors	(1, 30)
p	[1, 6]
metric	[‘manhattan’]
weights	[‘uniform’]
algorithm	[‘auto’]
leaf_size	[30]

Tabel 3. Hasil evaluasi sebelum *Hyperparameter Tuning*

Nama	Nilai
Accuracy	79.42%
Precision	0.63
Recall	0.78
F1-Score	0.68

Tabel 4. Hasil evaluasi sesudah *Hyperparameter Tuning*

Nama	Nilai
Accuracy	80.15%
Precision	0.64
Recall	0.78
F1-Score	0.69

Tabel 2 menunjukkan parameter yang digunakan dalam proses hyperparameter tuning algoritma *K-Nearest Neighbors* (KNN) menggunakan metode *Bayes Search* CV. Parameter yang dioptimasi meliputi jumlah tetangga terdekat (n_neighbors) dengan rentang nilai 1–30 serta parameter p pada perhitungan jarak *Minkowsk*i dengan rentang 1–6, sementara parameter lainnya menggunakan nilai tetap seperti *metric = manhattan, weights = uniform, algorithm = auto, dan leaf_size =* 30.

Hasil evaluasi akhir menunjukkan adanya perubahan pada metrik *Accuracy, Precision, Recall,* dan *F1-Score* setelah penerapan hyperparameter optimal yang diperoleh melalui *metode Bayes Search* CV, sebagaimana ditunjukkan pada Tabel 3 dan Tabel 4. Sebelum dilakukan optimasi, model memperoleh nilai *Accuracy* sebesar 79,42%, *precision* 0,63, *recall* 0,78, dan *F1-Score* 0,68. Setelah proses optimasi dilakukan, performa model meningkat dengan nilai *Accuracy* mencapai 80,15%, precision 0,64, recall tetap 0,78, serta F1-Score meningkat menjadi 0,69.

Peningkatan performa ini dipengaruhi oleh penggunaan parameter *weights = uniform*, yang memberikan bobot setara pada seluruh tetangga terdekat dalam proses klasifikasi. Pendekatan tersebut menghasilkan model yang lebih stabil dan mampu mengurangi pengaruh *outlier* dibandingkan penggunaan parameter *distance* yang lebih menitikberatkan pada jarak terdekat. Stabilnya nilai recall menunjukkan bahwa sensitivitas model dalam mendeteksi kelas positif tetap terjaga, sedangkan peningkatan nilai *Accuracy*, *precision*, dan *F1-Score* menunjukkan bahwa model KNN hasil optimasi menjadi lebih efektif dan seimbang dalam melakukan proses klasifikasi.

4. Kesimpulan

Hasil Penelitian ini berhasil mengoptimalkan kinerja algoritma *K-Nearest Neighbors* (KNN) melalui implementasi *Bayes Search Cross Validation* (CV) untuk identifikasi kanker payudara. Hasil optimasi menunjukkan peningkatan akurasi yang signifikan, yaitu dari 79,63% menjadi 80,15%. Temuan ini membuktikan bahwa penggunaan *Bayes Search* CV sangat efektif dalam mengidentifikasi kombinasi hyperparameter terbaik guna meningkatkan daya prediksi model terhadap data klinis dan genetik kanker payudara. Sebagai tindak lanjut, penelitian selanjutnya disarankan untuk melakukan studi komparatif dengan algoritma optimasi lain, seperti *Random Search* CV, guna mengevaluasi efektivitas relatif antar metode dalam penanganan dataset medis yang kompleks.

Daftar Pustaka:

Abd-Elnaby, M., Alfonse, M., & Roushdy, M. (2021). Classification of breast cancer using microarray *Gene Expression* data: A survey. In *Journal of Biomedical Informatics* (Vol. 117). Academic Press Inc. https://doi.org/10.1016/j.jbi.2021.103764

Abunasser, B. S., AL-Hiealy, M. R. J., Zaqout, I. S., & Abu-Naser, S. S. (2023). Literature review of breast cancer detection using *machine learning* algorithms. *AIP Conference Proceedings*, *2808*(1), 040006. https://doi.org/10.1063/5.0133688

Alanazi, S. A., Alshammari, N., Alruwaili, M., Junaid, K., Abid, M. R., & Ahmad, F. (2024). Integrative analysis of RNA expression data unveils distinct cancer types through *machine learning* techniques. *Saudi Journal of Biological Sciences*, *31*(3). https://doi.org/10.1016/j.sjbs.2023.103918

Alharbi, F., & Vakanski, A. (2023). *Machine learning* Methods for Cancer Classification Using *Gene Expression* Data: A Review. *Bioengineering*, *10*(2). https://doi.org/10.3390/bioengineering10020173

Alromema, N., Syed, A. H., & Khan, T. (2023). A Hybrid *Machine Learning* Approach to Screen Optimal Predictors for the Classification of Primary Breast Tumors from *Gene Expression* Microarray Data. *Diagnostics*, *13*(4). https://doi.org/10.3390/diagnostics13040708

Arifin, T., Agung, I. W. P., Junianto, E., Agustin, D. D., Wibowo, I. R., & Rachman, R. (2025). Breast cancer identification using a hybrid *machine learning* system. *International Journal of Electrical and Computer Engineering (IJECE)*, *15*(4), 3928. https://doi.org/10.11591/ijece.v15i4.pp3928-3937

Arifin, T., Agung, I. W. P., Junianto, E., Rachman, R., Wibowo, I. R., & Agustin, D. D. (2024). Breast cancer identification using *machine learning* and hyperparameter optimization. *Indonesian Journal of Electrical Engineering and Computer Science*, *36*(3), 1620–1630. https://doi.org/10.11591/ijeecs.v36.i3.pp1620-1630

Assegie, T. A. (2021). An optimized K-Nearest neighbor-based breast cancer detection. *Journal of Robotics and Control (JRC)*, *2*(3), 115–118. https://doi.org/10.18196/jrc.2363

Babichev, S., Liakh, I., & Škvor, J. (2024). *Integrating Data Mining, Deep Learning, and GeneOntology Analysis for Gene Expression-BasedDisease Diagnosis Systems*. https://doi.org/10.21203/rs.3.rs-3978499/v1

Bhat, M. A., Mir, M. A., Lakshmi, R. V., Pradhan, T., Rao, G. V. V. J., Tejani, G. G., & Hussain, S. A. (2026). *Machine learning* approaches for predicting breast cancer recurrence using clinical and histopathological data. *Clinical and Experimental Medicine*, *26*(1). https://doi.org/10.1007/s10238-025-02018-x

Boeri, C., Chiappa, C., Galli, F., De Berardinis, V., Bardelli, L., Carcano, G., & Rovera, F. (2020). *Machine learning* techniques in breast cancer prognosis prediction: A primary evaluation. *Cancer Medicine*, *9*(9), 3234–3243. https://doi.org/10.1002/cam4.2811

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for *machine learning* algorithms used for arabic sentiment analysis. *Informatics*, *8*(4). https://doi.org/10.3390/informatics8040079

Farhad Khorshid, S., & Mohsin Abdulazeez, A. (2021). Breast Cancer Diagnosis Based On *K-Nearest Neighbors*: A Review Pjaee, 18 (4) (2021) Breast Cancer Diagnosis Based On *K-Nearest Neighbors*: A REVIEW. In *Journal of Archaeology of Egypt/Egyptology* (Vol. 18, Number 4).

Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of *Machine learning* Techniques, and Their Analysis. In *IEEE Access* (Vol. 8, pp. 150360–150376). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2020.3016715

Kallah-Dagadu, G., Mohammed, M., Nasejje, J. B., Mchunu, N. N., Twabi, H. S., Batidzirai, J. M., Singini, G. C., Nevhungoni, P., & Maposa, I. (2025). Breast cancer prediction based on *Gene Expression* data using interpretable *machine learning* techniques. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-85323-5

Kim, D. H., & Lee, K. E. (2022). Discovering Breast Cancer Biomarkers Candidates through mRNA Expression Analysis Based on The Cancer Genome Atlas Database. *Journal of Personalized Medicine*, *12*(10). https://doi.org/10.3390/jpm12101753

Mudzakir, I., & Arifin, T. (2022). Klasifikasi Penggunaan Masker dengan Convolutional Neural Network Menggunakan Arsitektur MobileNetv2. *EXPERT: Jurnal Manajemen Sistem Informasi Dan Teknologi*, *12*(1), 76. https://doi.org/10.36448/expert.v12i1.2466

Nisanova, A., Yavary, A., Deaner, J., Ali, F. S., Gogte, P., Kaplan, R., Chen, K. C., Nudleman, E., Grewal, D., Gupta, M., Wolfe, J., Klufas, M., Yiu, G., Soltani, I., & Emami-Naeini, P. (2024). Performance of Automated *Machine Learning* in Predicting Outcomes of Pneumatic Retinopexy. *Ophthalmology Science*, *4*(5). https://doi.org/10.1016/j.xops.2024.100470

Ozturk Kiyak, E., Ghasemkhani, B., & Birant, D. (2023). High-Level *K-Nearest Neighbors* (HLKNN): A Supervised *Machine Learning* Model for Classification Analysis. *Electronics (Switzerland)*, *12*(18). https://doi.org/10.3390/electronics12183828

Phan, N. N., Huang, C. C., Tseng, L. M., & Chuang, E. Y. (2021). Predicting Breast Cancer *Gene Expression* Signature by Applying Deep Convolutional Neural Networks from

Unannotated Pathological Images. *Frontiers in Oncology*, *11*. https://doi.org/10.3389/fonc.2021.769447

Qiu, Y., & Liu, P. (2025). Investigation of ML algorithms for prediction of CFD data of fluid flow inside a packed-bed reactor. *Case Studies in Thermal Engineering*, *70*. https://doi.org/10.1016/j.csite.2025.106093

Rahman, Md. M., Rahman, A., Akter, S., & Pinky, S. A. (2023). Hyperparameter Tuning Based *Machine Learning* Classifier for Breast Cancer Prediction. *Journal of Computer and Communications*, *11*(04), 149–165. https://doi.org/10.4236/jcc.2023.114007

Rajaguru, H., & Sannasi Chakravarthy, S. R. (2019). Analysis of *Decision Tree* and k-nearest neighbor algorithm in the classification of breast cancer. *Asian Pacific Journal of Cancer Prevention*, *20*(12), 3777–3781. https://doi.org/10.31557/APJCP.2019.20.12.3777

Ranti, N., 1∗, M., & Hanif, K. H. (2022). *Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine learning*. *3*(1), 1–6. http://creativecommons.org/licences/by/4.0/

Rasheda, A., Arifin, T., Studi, P., Informatika, T., Adhirajasa, U., Sanjaya, R., Neighbor, K., & Belakang, P. T. (2022). *Penerapan K-Nearest Neighbor Untuk Sistem Pakar Diagnosa Penyakit Tulang Belakang*. *3*(2).

Resmiati, R., & Arifin, T. (2021). Klasifikasi Pasien Kanker Payudara Menggunakan Metode *Support Vector Machine* dengan Backward Elimination. *Sistemasi*, *10*(2), 381. https://doi.org/10.32520/stmsi.v10i2.1238

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. In *International Journal of Biological Sciences* (Vol. 13, Number 11, pp. 1387–1397). Ivyspring International Publisher. https://doi.org/10.7150/ijbs.21635

Takeshita, T., Iwase, H., Wu, R., Ziazadeh, D., Yan, L., & Takabe, K. (2023). Development of a *Machine Learning*-Based Prognostic Model for Hormone Receptor-Positive Breast Cancer Using Nine-*Gene Expression* Signature. *World Journal of Oncology*, *14*(5), 406–422. https://doi.org/10.14740/wjon1700

Thalor, A., Kumar Joon, H., Singh, G., Roy, S., & Gupta, D. (2022). *Machine learning* assisted analysis of breast cancer *Gene Expression* profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer. *Computational and Structural Biotechnology Journal*, *20*, 1618–1631. https://doi.org/10.1016/j.csbj.2022.03.019

Ubaidillah, R., Muliadi, M., Nugrahadi, D. T., Faisal, M. R., & Herteno, R. (2022). Implementasi XGBoost Pada Keseimbangan Liver Patient Dataset dengan SMOTE dan Hyperparameter Tuning Bayesian Search. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *6*(3), 1723. https://doi.org/10.30865/mib.v6i3.4146

Zhao, Y., Zhang, W., & Liu, X. (2024). Grid search with a weighted error function: Hyper-parameter optimization for financial time series forecasting. *Applied Soft Computing*, *154*. https://doi.org/10.1016/j.asoc.2024.111362