BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1489-1497
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1087

Perbandingan Grid Search dan Random Search untuk Optimasi Hyperparameter Random Forest pada Klasifikasi Kanker Payudara

Fiona Yenisya Dewi*, Bayu Rizkya Pratama, Sunaryono

Program Studi Teknik Informatika, STMIK Widya Utama, Purwokerto, Indonesia
Email: [1,*]fionayenisyadewi16@gmail.com, [2]bayurizykapratama@swu.ac.id, [3]sunaryono@swu.ac.id
Email Penulis Korespondensi: fionayenisyadewi16@gmail.com

Abstrak−Kanker payudara merupakan jenis kanker dengan jumlah kasus terbanyak di Indonesia, dengan 71% pasien terdiagnosis pada stadium lanjut akibat keterbatasan akses deteksi dini. Kondisi ini menuntut pengembangan sistem skrining berbasis machine learning yang tidak hanya akurat, tetapi juga efisien secara komputasi agar dapat diimplementasikan secara luas pada fasilitas kesehatan dengan sumber daya terbatas, sehingga pemilihan metode optimasi *hyperparameter* yang efisien menjadi krusial. Penelitian ini membandingkan dua metode optimasi *hyperparameter*, yaitu *Grid Search* dan *Random Search*, pada algoritma Random Forest menggunakan UCI Wisconsin Diagnostic Breast Cancer Dataset (569 sampel, 30 fitur numerik) dengan ruang pencarian identik mencakup 288 kombinasi *hyperparameter* dan stratified 5-fold cross-validation. Hasil eksperimen menunjukkan bahwa *Random Search* RF mencapai performa setara dengan Baseline RF pada metrik berbasis threshold (akurasi 0,9737; F1-Score 0,9630) sekaligus menghasilkan AUC-ROC tertinggi sebesar 0,9950 dalam waktu 88,26 detik. Sebaliknya, *Grid Search* RF menghasilkan performa di bawah baseline (akurasi 0,9561; F1-Score 0,9367) dengan waktu komputasi 526,73 detik, akibat fenomena *optimizer's curse* di mana kombinasi terpilih berdasarkan cross-validation tidak menghasilkan generalisasi optimal pada data uji. *Random Search* terbukti 5,97 kali lebih efisien dibandingkan *Grid Search* dengan kualitas solusi yang lebih baik, mengkonfirmasi secara empiris proposisi teoritis bahwa *Random Search* mampu menemukan konfigurasi kompetitif dengan biaya komputasi yang jauh lebih rendah dibandingkan pencarian exhaustive pada ruang *hyperparameter* berdimensi tinggi.

Kata Kunci: Random Forest; Grid Search ; Random Search; Optimasi Hyperparameter; Klasifikasi Kanker Payudara

Abstract−Breast cancer is the most prevalent type of cancer in Indonesia, with 71% of patients diagnosed at advanced stages due to limited access to early detection. This condition necessitates the development of machine learning-based screening systems that are not only accurate but also computationally efficient to enable widespread implementation in healthcare facilities with limited resources, making the selection of an efficient hyperparameter optimization method crucial. This study compares two hyperparameter optimization methods, namely Grid Search and Random Search, applied to the Random Forest algorithm using the UCI Wisconsin Diagnostic Breast Cancer Dataset (569 samples, 30 numerical features) with an identical search space encompassing 288 hyperparameter combinations and stratified 5-fold cross-validation. Experimental results demonstrate that Random Search RF achieves performance equivalent to Baseline RF on threshold-based metrics (accuracy 0.9737; F1-Score 0.9630) while producing the highest AUC-ROC of 0.9950 in 88,26 seconds. In contrast, Grid Search RF yields performance below the baseline (accuracy 0.9561; F1-Score 0.9367) with a computation time of 526,73 seconds, attributable to the optimizer's curse phenomenon in which the selected combination based on cross-validation does not produce optimal generalization on the test data. Random Search is proven to be 5.97 times more efficient than Grid Search with superior solution quality, empirically confirming the theoretical proposition that Random Search is capable of finding competitive configurations at substantially lower computational cost compared to exhaustive search in high-dimensional hyperparameter spaces.

Keywords: Random Forest; Grid Search; Random Search; Hyperparameter Optimization; Breast Cancer Classification

1. PENDAHULUAN

Kanker payudara merupakan salah satu penyakit tidak menular dengan beban kesehatan terbesar di dunia. Berdasarkan data GLOBOCAN 2022, kanker payudara tercatat sebagai jenis kanker paling umum pada perempuan dengan estimasi 2,3 juta kasus baru, mewakili 11,6% dari seluruh kasus kanker global, sekaligus menjadi penyebab kematian akibat kanker tertinggi keempat di dunia dengan 665.684 kematian atau setara 6,9% dari total kematian akibat kanker [1]. Di antara seluruh jenis kanker yang menyerang perempuan, kanker payudara mendominasi sebagai kanker paling banyak didiagnosis di 157 negara dan menjadi penyebab kematian utama akibat kanker pada perempuan di 112 negara di seluruh dunia [1]. Kondisi ini semakin mengkhawatirkan mengingat proyeksi demografis memperkirakan jumlah kasus baru kanker secara global akan meningkat hingga 35 juta pada tahun 2050, bertumbuh sebesar 77% dibandingkan estimasi tahun 2022, dengan beban peningkatan terbesar diprediksi terjadi di negara-negara berpenghasilan rendah dan menengah termasuk Indonesia [1]. Besarnya beban global ini mendorong pemanfaatan teknologi komputasi dan *machine learning* sebagai pendekatan yang menjanjikan untuk mendukung deteksi dini dan klasifikasi kanker payudara secara lebih cepat, akurat, dan skalabel.

Di Indonesia, kanker payudara menempati posisi pertama sebagai jenis kanker dengan jumlah kasus terbanyak sekaligus penyebab kematian akibat kanker tertinggi pada perempuan. Berdasarkan data GLOBOCAN 2022, dari total 408.661 kasus baru kanker yang tercatat di Indonesia, kanker payudara mendominasi dengan 66.271 kasus atau 16,2% dari seluruh kasus kanker nasional, disertai 22.598 kematian setiap tahunnya [2]. Tantangan terbesar yang dihadapi bukan sekadar tingginya angka kejadian, melainkan pola keterlambatan diagnosis yang masih sangat memprihatinkan. Berdasarkan data registri berbasis rumah sakit PERABOI yang mencakup 4.959 pasien dari 16 provinsi di Indonesia, sebesar 71% pasien kanker payudara terdiagnosis pada stadium lanjut (stadium III dan IV), padahal deteksi dini pada stadium awal dapat meningkatkan peluang kesembuhan hingga 90% [3]. Keterlambatan diagnosis ini tidak terjadi secara merata; faktor pendidikan rendah, tinggal di wilayah pedesaan, serta keterbatasan akses layanan kesehatan di luar Pulau

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1489-1497
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1087

Jawa-Bali terbukti secara statistik menjadi prediktor independen yang signifikan terhadap diagnosis stadium lanjut [3]. Keterbatasan kesadaran masyarakat terhadap pentingnya deteksi dini serta belum meratanya akses fasilitas kesehatan di berbagai daerah menjadi faktor utama yang menghambat upaya diagnosis lebih awal [3], sehingga pengembangan pendekatan berbasis teknologi yang mampu mendukung proses skrining dan klasifikasi diagnosis menjadi sangat relevan dan mendesak untuk dikaji.

Perkembangan *machine learning* dalam satu dekade terakhir telah membuka peluang baru untuk mendukung diagnosis medis secara lebih cepat, konsisten, dan skalabel. Berbagai algoritma *machine learning* telah diuji untuk klasifikasi kanker payudara menggunakan dataset yang sama dengan penelitian ini, yakni UCI Wisconsin Diagnostic Breast Cancer Dataset dengan 569 sampel dan 30 fitur numerik. Lemons & Rathnathungalage [4] membandingkan algoritma Naïve Bayes dan Random Forest pada dataset tersebut dan menemukan bahwa Random Forest mencapai akurasi klasifikasi tertinggi sebesar 97,82% dengan menggunakan 17 fitur terpilih, mengungguli Naïve Bayes yang hanya mencapai 94,55% pada pengujian tanpa seleksi fitur [4]. Batool & Byun [5] juga menerapkan Random Forest pada dataset yang sama menggunakan validasi *10-fold cross-validation* dan melaporkan akurasi pelatihan 91,5% serta akurasi validasi 90,3%, lebih tinggi dibandingkan model Tree Ensemble sebagai pembanding [5]. Konsistensi performa Random Forest pada berbagai benchmark klasifikasi medis menjadikannya kandidat algoritma yang kuat untuk tugas ini. Namun demikian, performa optimal Random Forest tidak diperoleh secara otomatis, melainkan sangat bergantung pada konfigurasi *hyperparameter* yang tepat, mencakup parameter seperti jumlah pohon, kedalaman maksimum pohon, dan jumlah fitur yang dipertimbangkan pada setiap pemisahan node [6].

Rahman et al. [7] membandingkan lima algoritma pada dataset WDBC menggunakan *Grid Search* dan menemukan peningkatan akurasi dari 93,56% menjadi 98,83%, namun Random Forest tidak menunjukkan peningkatan dan tetap berada di 97,08% [7]. Jain et al. [8] menemukan hasil berbeda, di mana Random Forest dengan *Grid Search* CV berhasil meningkatkan akurasi dari 96,48% menjadi 98,43%, meskipun penelitian tersebut tidak menyertakan perbandingan dengan metode *Random Search* [8]. Sholeh et al. [9] yang secara spesifik membandingkan *Grid Search* dan *Random Search* pada Decision Tree menggunakan dataset WDBC menemukan akurasi meningkat dari 92,98% menjadi 95,61% dengan *Grid Search* dan 97,37% dengan *Random Search*, namun belum menguji kedua metode tersebut pada algoritma Random Forest [9]. Ketiga penelitian ini secara kolektif menunjukkan bahwa perbandingan langsung *Grid Search* dan *Random Search* khusus pada algoritma Random Forest untuk klasifikasi kanker payudara belum pernah dilakukan. Lebih dari sekadar kelengkapan empiris, perbandingan ini memiliki relevansi ilmiah tersendiri. Random Forest memiliki ruang *hyperparameter* yang lebih tinggi-dimensional dibandingkan Decision Tree, mencakup setidaknya lima parameter utama yang saling berinteraksi, sehingga perilaku kedua metode pencarian tidak dapat diasumsikan identik dengan hasil yang ditemukan pada algoritma lain [9]. Bergstra dan Bengio secara teoritis membuktikan bahwa *Random Search* lebih efisien dibandingkan *Grid Search* ketika dimensi ruang pencarian tinggi dan tidak semua *hyperparameter* sama-sama berpengaruh terhadap performa model [10]. Oleh karena itu, validasi empiris pada Random Forest secara spesifik diperlukan untuk menentukan apakah keunggulan teoritis *Random Search* tersebut terkonfirmasi dalam konteks klasifikasi kanker payudara.

Berdasarkan uraian permasalahan dan kajian penelitian terdahulu di atas, penelitian ini bertujuan untuk membandingkan efektivitas dua metode optimasi *hyperparameter*, yaitu *Grid Search* dan *Random Search*, dalam meningkatkan performa algoritma Random Forest untuk klasifikasi kanker payudara menggunakan UCI Wisconsin Diagnostic Breast Cancer Dataset. Performa model dievaluasi menggunakan enam metrik, yakni akurasi, *precision*, *recall*, F1-*score*, AUC-ROC, dan waktu komputasi, sehingga perbandingan mencakup dua dimensi sekaligus yaitu kualitas klasifikasi dan efisiensi proses optimasi.

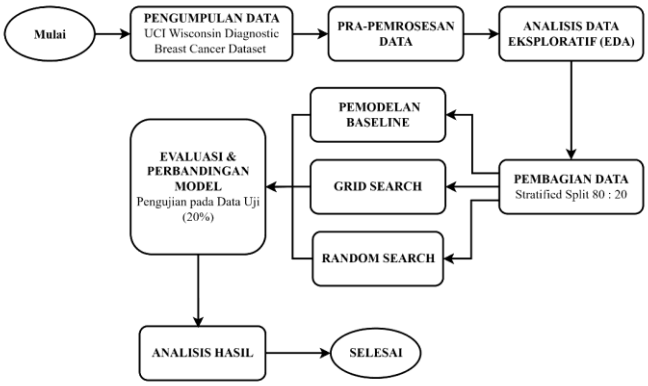
Penelitian ini memberikan tiga kontribusi utama. Pertama, validasi empiris perbandingan langsung *Grid Search* dan *Random Search* pada algoritma Random Forest untuk klasifikasi kanker payudara sejauh yang diketahui penulis berdasarkan tinjauan literatur yang dilakukan melengkapi kesenjangan yang belum ditangani oleh penelitian sebelumnya. Kedua, analisis trade-off antara kualitas klasifikasi dan efisiensi komputasi dari kedua metode optimasi, yang dievaluasi secara komprehensif menggunakan enam metrik sekaligus. Ketiga, analisis sensitivitas pengaruh konfigurasi *hyperparameter* terhadap stabilitas performa model Random Forest, yang memberikan wawasan metodologis tentang perilaku kedua strategi pencarian pada ruang *hyperparameter* berdimensi tinggi.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan eksperimental komparatif untuk membandingkan dua metode optimasi *hyperparameter*, yaitu *Grid Search* dan *Random Search*, dalam meningkatkan performa algoritma Random Forest pada klasifikasi kanker payudara. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 1.

Penelitian diawali dengan pengumpulan data menggunakan UCI Wisconsin Diagnostic Breast Cancer Dataset, dilanjutkan dengan pra-pemrosesan data dan analisis data eksploratif (EDA). Setelah itu, data dibagi menggunakan stratified split dengan rasio 80:20 untuk kemudian digunakan dalam tiga proses pemodelan secara paralel, yaitu Baseline RF, *Grid Search* RF, dan *Random Search* RF. Hasil ketiga model selanjutnya dievaluasi dan dibandingkan pada data uji (20%), sebelum dilakukan analisis hasil secara menyeluruh.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1489-1497
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1087



Gambar 1. Diagram Alur Penelitian

2.1 Dataset

Dataset yang digunakan dalam penelitian ini adalah *Wisconsin Diagnostic Breast Cancer* (WDBC) yang diperoleh dari UCI Machine Learning Repository [11]. Dataset ini pertama kali diperkenalkan oleh Street et al. melalui sistem analisis citra digital untuk mengekstraksi fitur nukleus sel dari hasil *fine needle aspirate* (FNA) jaringan payudara [12].

Dataset terdiri dari 569 sampel dengan 30 fitur numerik kontinu yang merepresentasikan karakteristik geometris dan tekstur nukleus sel, mencakup nilai rata-rata, standar error, dan nilai terburuk dari sepuluh pengukuran dasar yaitu radius, tekstur, perimeter, luas, kehalusan, kekompakan, kecekungan, titik cekung, simetri, dan dimensi fraktal [12]. Setiap sampel diklasifikasikan ke dalam dua kelas, yaitu *Malignant* (M) atau ganas sebanyak 212 sampel (37,3%) dan *Benign* (B) atau jinak sebanyak 357 sampel (62,7%).

Dataset ini dipilih karena telah menjadi benchmark standar yang digunakan secara luas dalam penelitian klasifikasi kanker payudara berbasis machine learning [4], [5], [7] - [9], sehingga hasil penelitian ini dapat dibandingkan secara langsung dengan temuan dari studi-studi sebelumnya.

2.2 Pra-Pemrosesan Data

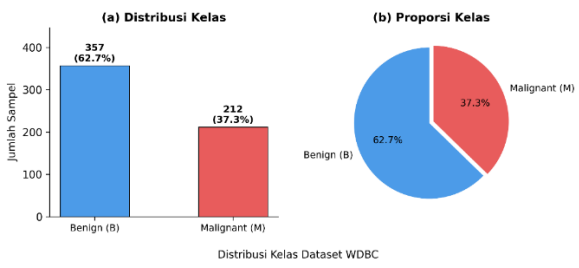
Tahap pra-pemrosesan data dilakukan untuk memastikan kualitas dan konsistensi data sebelum digunakan dalam proses pemodelan. Terdapat tiga langkah utama yang dilakukan pada tahap ini. Langkah pertama adalah pemeriksaan missing value. Seluruh kolom pada dataset diperiksa untuk mendeteksi keberadaan nilai yang hilang. Hasil pemeriksaan menunjukkan bahwa dataset WDBC tidak mengandung missing value pada seluruh 30 fitur numeriknya, sehingga tidak diperlukan penanganan imputasi data [12].

Langkah kedua adalah penghapusan kolom yang tidak relevan. Dataset WDBC memiliki kolom ID yang berfungsi sebagai pengenal unik setiap sampel dan tidak memiliki kontribusi terhadap proses klasifikasi. Kolom ini dihapus sebelum pemodelan untuk menghindari pengaruh yang tidak relevan terhadap performa model.

Langkah ketiga adalah encoding variabel target. Kolom diagnosis yang semula berisi nilai kategorikal M (*Malignant*) dan B (*Benign*) dikonversi menjadi nilai numerik biner, yaitu M=1 dan B=0, agar dapat diproses oleh algoritma machine learning. Normalisasi atau standarisasi fitur tidak diterapkan dalam penelitian ini karena algoritma Random Forest berbasis decision tree tidak sensitif terhadap skala fitur pemisahan node ditentukan oleh urutan nilai, bukan magnitude absolut dari setiap fitur [6][13].

2.3 Analisis Data Eksploratif

Analisis data eksploratif dilakukan untuk memahami karakteristik dataset sebelum proses pemodelan, mencakup pemeriksaan distribusi kelas, pola korelasi antar fitur, dan separabilitas fitur berdasarkan kelas diagnosis.
Hasil pemeriksaan distribusi kelas menunjukkan bahwa dataset WDBC terdiri dari 357 sampel kelas *Benign* (62,7%) dan 212 sampel kelas *Malignant* (37,3%) sebagaimana ditunjukkan pada Gambar 2. Rasio ketidakseimbangan kelas sebesar 1,68:1 tergolong dalam kategori imbalance ringan yang masih dapat ditangani oleh algoritma Random Forest tanpa memerlukan teknik resampling tambahan [14].

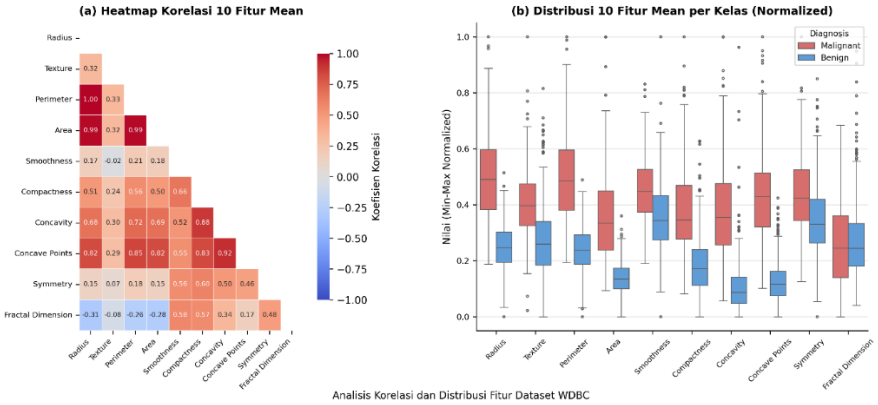


Gambar 2. Distribusi Kelas Dataset WDBC

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1489-1497
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1087

Analisis korelasi dan distribusi fitur dilakukan menggunakan 10 fitur mean sebagai representasi utama karakteristik nukleus sel, mengingat 30 fitur dataset merupakan turunan dari 10 konsep dasar yang sama yang dihitung menggunakan tiga ukuran statistik berbeda, yaitu nilai rata-rata, standar error, dan nilai terburuk [12]. Hasil heatmap korelasi pada Gambar 3(a) menunjukkan bahwa terdapat korelasi sangat tinggi antara fitur radius, perimeter, dan area dengan koefisien korelasi mendekati 1,00, mengindikasikan adanya multikolinearitas pada ketiga fitur tersebut. Fitur concavity dan concave points juga menunjukkan korelasi tinggi sebesar 0,92. Meskipun demikian, penghapusan fitur berkorelasi tinggi tidak dilakukan dalam penelitian ini karena Random Forest secara inheren robust terhadap multikolinearitas melalui mekanisme pemilihan fitur acak pada setiap pemisahan node [6], [15].

Hasil boxplot pada Gambar 3(b) menunjukkan bahwa sebagian besar fitur mean memiliki separabilitas yang jelas antara kelas *Benign* dan *Malignant*, terutama pada fitur *radius, perimeter, area, concavity,* dan *concave points* di mana median kelas *Malignant* secara konsisten lebih tinggi dibandingkan kelas *Benign*. Hal ini mengindikasikan bahwa fitur-fitur tersebut memiliki daya diskriminatif yang kuat untuk membedakan kedua kelas diagnosis. Karakteristik separabilitas yang baik ini menjadi landasan mengapa Random Forest mampu mencapai performa klasifikasi yang tinggi bahkan pada kondisi baseline tanpa optimasi hyperparameter.



Gambar 3. Analisis Korelasi dan Distribusi Fitur Dataset WDBC

2.4 Pembagian Data

Dataset dibagi menjadi data latih dan data uji menggunakan metode stratified split dengan rasio 80:20, menghasilkan 455 sampel data latih dan 114 sampel data uji [16]. Stratified split dipilih untuk memastikan proporsi kelas *Benign* dan *Malignant* pada data latih dan data uji identik dengan proporsi pada dataset asli, sehingga estimasi performa model tidak bias akibat ketidakseimbangan distribusi kelas pada salah satu subset data. Pembagian data dilakukan sebelum tahap preprocessing apapun untuk mencegah terjadinya data leakage dari data uji ke data latih [17].

2.5 Pemodelan Baseline

Model baseline dibangun menggunakan algoritma Random Forest dengan seluruh parameter dalam kondisi *default* tanpa optimasi apapun. Model baseline berfungsi sebagai acuan perbandingan untuk mengukur sejauh mana proses optimasi *hyperparameter* melalui *Grid Search* dan *Random Search* memberikan peningkatan performa yang signifikan [14]. Parameter default yang digunakan mencakup jumlah pohon sebesar 100, kedalaman pohon tidak dibatasi, minimum sampel untuk pemisahan node sebesar 2, minimum sampel pada daun pohon sebesar 1, dan jumlah fitur yang dipertimbangkan pada setiap pemisahan menggunakan akar kuadrat dari total fitur [6].

2.6 Optimasi Hyperparameter

2.6.1 Grid Search

Grid Search melakukan pencarian exhaustive terhadap seluruh kombinasi nilai *hyperparameter* yang telah didefinisikan dalam parameter grid [7], [18] . Pada penelitian ini, *Grid Search* diimplementasikan menggunakan fungsi GridSearchCV yang tersedia pada library Scikit-learn [16], [19], parameter grid yang digunakan mencakup lima *hyperparameter* utama Random Forest sebagaimana ditunjukkan pada Tabel 1.

Tabel 1. Parameter Grid untuk *Grid Search* dan *Random Search*

Hyperparameter	Nilai yang Diuji	Jumlah Opsi
n_estimators	50, 100, 200, 300	4
max_depth	None, 5, 10, 20	4
min_samples_split	2, 5, 10	3
min_samples_leaf	1, 2, 4	3
max_features	sqrt, log2	2
Total kombinasi		288

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1489-1497
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1087

Kombinasi parameter grid menghasilkan 288 kombinasi *hyperparameter* yang seluruhnya dievaluasi oleh *Grid Search* menggunakan stratified 5-fold cross-validation pada data latih, sehingga total *fitting* yang dilakukan adalah 1.440 kali. Metrik *scoring* yang digunakan selama proses tuning adalah *F1-score* dengan pertimbangan metodologis sebagai berikut. Pertama, F1-score merupakan rata-rata harmonik dari *precision* dan *recall* yang secara bersamaan mempertimbangkan kedua aspek kesalahan klasifikasi *false positive* dan *false negative* sehingga lebih informatif dibandingkan akurasi pada kondisi distribusi kelas yang tidak sepenuhnya seimbang seperti dataset WDBC dengan rasio 1,68:1 [20].Kedua, meskipun *recall* secara klinis lebih kritis untuk meminimalkan kasus kanker yang tidak terdeteksi, penggunaan *recall* tunggal sebagai *scoring metric* berisiko menghasilkan model yang terlalu agresif dalam memprediksi kelas *Malignant* sehingga menurunkan *precision* secara tidak terkendali [18]. Ketiga, penggunaan AUC-ROC sebagai *scoring metric* tuning tidak dipilih karena AUC-ROC mengoptimalkan kemampuan diskriminasi pada seluruh rentang *threshold*, sementara evaluasi akhir model pada praktik klinis dilakukan pada *threshold* tunggal (0,5), sehingga F1-score yang berbasis *threshold* lebih relevan sebagai panduan seleksi *hyperparameter* [20]. Dengan demikian, F1-score dipilih sebagai kompromi yang seimbang antara sensitivitas klinis dan kontrol terhadap *false positive* dalam konteks klasifikasi biner kanker payudara.

2.6.2 Random Search

Random Search melakukan pencarian dengan mengambil sampel kombinasi *hyperparameter* secara acak dari ruang pencarian yang sama dengan *Grid Search* [10], [18]. diimplementasikan menggunakan fungsi RandomizedSearchCV pada library Scikit-learn [16], [21]. Jumlah iterasi ditetapkan sebesar 50 iterasi berdasarkan dua pertimbangan. Pertama, secara teoritis Bergstra dan Bengio membuktikan bahwa untuk menemukan konfigurasi *hyperparameter* yang berada dalam persentil terbaik-5% dari seluruh ruang pencarian dengan probabilitas 95%, dibutuhkan minimal 60 iterasi namun pada ruang pencarian yang tidak seragam di mana hanya sebagian kecil *hyperparameter* yang benar-benar berpengaruh, jumlah iterasi yang lebih kecil sudah cukup [10]. Dengan 50 iterasi yang mencakup 17,4% dari total 288 kombinasi *Grid Search* , probabilitas menemukan konfigurasi kompetitif tetap tinggi karena Random Forest memiliki karakteristik *low effective dimensionality* tidak semua lima *hyperparameter* yang dioptimasi memberikan pengaruh yang setara terhadap performa model [6], [10].

Kedua, secara empiris pemilihan 50 iterasi konsisten dengan praktik yang direkomendasikan dalam literatur optimasi *hyperparameter* untuk dataset berukuran sedang, di mana jumlah iterasi antara 25–60 umumnya menghasilkan solusi yang konvergen [14]. Untuk memastikan reprodusibilitas eksperimen, parameter random_state ditetapkan pada nilai tetap yang identik untuk seluruh proses pembagian data (*stratified split*), *cross-validation*, dan proses pencarian *Random Search* sehingga seluruh hasil yang dilaporkan dalam penelitian ini dapat direplikasi secara eksak pada eksekusi berikutnya. Parameter *space* yang digunakan identik dengan parameter grid pada Tabel 1 untuk memastikan perbandingan yang adil [18]. Dengan 50 iterasi dan 5-*fold cross-validation*, total *fitting* yang dilakukan *Random Search* adalah 250 kali atau hanya 17,4% dari total *fitting Grid Search* sehingga perbandingan efisiensi komputasi antara kedua metode menjadi sangat informatif.

2.7 Metrik Evaluasi

Performa ketiga model dievaluasi menggunakan enam metrik yang mencakup dua dimensi sekaligus, yaitu kualitas klasifikasi dan efisiensi komputasi. Akurasi mengukur proporsi prediksi yang benar dari keseluruhan sampel uji. *Precision* mengukur proporsi prediksi *Malignant* yang benar-benar *Malignant*, yang relevan untuk meminimalkan *false positive* dalam konteks diagnosis medis. *Recall* mengukur proporsi sampel *Malignant* yang berhasil dideteksi oleh model, yang kritis untuk meminimalkan *false negative* mengingat risiko klinis dari kanker yang tidak terdeteksi. F1-*score* merupakan rata-rata harmonik dari *precision* dan *recall* yang memberikan evaluasi seimbang pada kondisi distribusi kelas yang tidak sepenuhnya seimbang [18]. AUC-ROC mengukur kemampuan model dalam membedakan kedua kelas secara keseluruhan pada berbagai *threshold* klasifikasi, di mana nilai mendekati 1,0 menunjukkan diskriminasi yang sempurna [20], [22]. Waktu komputasi diukur dalam satuan detik untuk membandingkan efisiensi proses optimasi antara *Grid Search* dan *Random Search* pada kondisi *hardware* yang identik.

3. HASIL DAN PEMBAHASAN

Secara keseluruhan, hasil evaluasi menunjukkan pola yang menarik secara ilmiah: *Grid Search* RF yang menggunakan pencarian exhaustive terhadap 288 kombinasi *hyperparameter* justru menghasilkan performa di bawah Baseline RF, sementara *Random Search* RF dengan hanya 50 iterasi mampu menghasilkan performa setara Baseline RF sekaligus mencapai AUC-ROC tertinggi. Perbedaan antar model terutama terkonsentrasi pada tiga dimensi: kemampuan deteksi kelas *Malignant* (Recall), kemampuan diskriminasi keseluruhan (AUC-ROC), dan efisiensi waktu komputasi. Analisis mendalam terhadap masing-masing dimensi tersebut diuraikan pada sub-bagian berikut.

3.1 Hasil Evaluasi Model

Ketiga model Random Forest Baseline, *Grid Search* , dan *Random Search* dievaluasi pada data uji sebanyak 114 sampel menggunakan enam metrik evaluasi. Hasil perbandingan lengkap disajikan pada Tabel 2.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1489-1497
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1087

Tabel 2. Perbandingan Performa Ketiga Model Random Forest

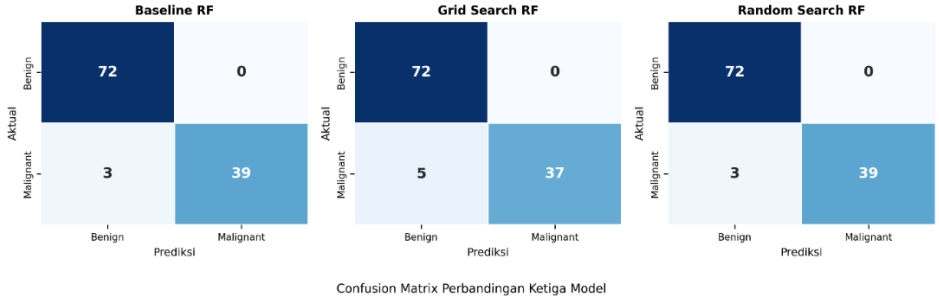
Model	Akurasi	Precision	Recall	F1-Score	AUC-ROC	Waktu (detik)
Baseline RF	0,9737	1,0000	0,9286	0,9630	0,9929	0,30
Grid Search RF	0,9561	1,0000	0,8810	0,9367	0,9942	526,73
Random Search RF	0,9737	1,0000	0,9286	0,9630	0,9950	88,26

Ketiga model mencapai nilai Precision sempurna sebesar 1,0000, artinya tidak satu pun sampel *Benign* diprediksi sebagai *Malignant* (false positive = 0). Perbedaan performa antar model terutama terletak pada dimensi Recall, F1-Score, dan efisiensi waktu komputasi.

3.1 Analisis Confusion Matrix

Hasil confusion matrix ketiga model ditunjukkan pada Gambar 4. Ketiga model mencapai nilai True Negative (TN) yang identik sebesar 72, artinya seluruh 72 sampel *Benign* pada data uji berhasil diklasifikasikan dengan benar tanpa satu pun false positive, yang tercermin dari nilai Precision sempurna sebesar 1,0000 pada ketiga model.

Perbedaan signifikan antar model terletak pada nilai False Negative (FN), yaitu kasus *Malignant* yang gagal terdeteksi. Baseline RF dan *Random Search* RF menghasilkan FN sebesar 3 dari 42 sampel *Malignant*, sementara *Grid Search* RF menghasilkan FN lebih tinggi sebesar 5. Dalam konteks diagnosis kanker, false negative memiliki implikasi klinis yang lebih serius dibandingkan false positive karena kegagalan mendeteksi kanker dapat mengakibatkan keterlambatan penanganan yang berdampak fatal bagi pasien [16]. Oleh karena itu, nilai Recall Baseline RF dan *Random Search* RF sebesar 0,9286 lebih unggul secara klinis dibandingkan *Grid Search* RF sebesar 0,8810. Secara keseluruhan, hasil confusion matrix mengkonfirmasi bahwa Random Search RF merupakan pilihan yang lebih optimal dibandingkan Grid Search RF, karena menghasilkan performa klasifikasi yang setara dengan Baseline RF sekaligus lebih efisien secara komputasi.



Gambar 4. Confusion Matrix untuk Perbandingan Ketiga Model

3.2 Analisis Performa Klasifikasi

Baseline RF dan *Random Search* RF mencapai akurasi tertinggi sebesar 0,9737 dan F1-Score sebesar 0,9630, melampaui *Grid Search* RF yang hanya mencapai akurasi 0,9561 dan F1-Score 0,9367. Hasil ini menunjukkan fenomena yang menarik secara ilmiah: optimasi *hyperparameter* menggunakan *Grid Search* tidak memberikan peningkatan performa dibandingkan model *default*, bahkan menghasilkan penurunan akurasi sebesar 1,76 poin persentase dan penurunan F1-Score sebesar 2,63 poin persentase.

Fenomena ini dapat dijelaskan melalui analisis konfigurasi *hyperparameter* terpilih oleh masing-masing metode. *Grid Search* memilih kombinasi *max_depth* = 10 dan *max_features* = log2 berdasarkan CV F1-score tertinggi pada data latih sebesar 0,9557, sementara *Random Search* memilih *max_depth* = 5 dan *max_features* = sqrt dengan CV F1-score 0,9553. Konfigurasi *hyperparameter* lengkap yang dipilih oleh masing-masing metode optimasi disajikan pada Tabel 3.

Tabel 3. *Hyperparameter* Optimal yang Dipilih oleh Masing-Masing Metode

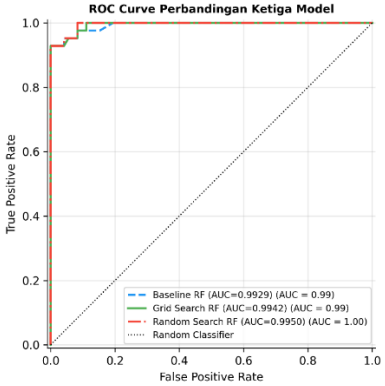
Hyperparameter	Baseline RF	*Grid Search* RF	*Random Search* RF
n_estimators	100	200	100
max_depth	None	10	5
min_samples_split	2	2	2
min_samples_leaf	1	1	1
max_features	sqrt	log2	sqrt
CV F1-Score	-	0,9557	0,9553

Tabel 3 mengungkap temuan yang secara ilmiah sangat informatif: *Random Search* RF hanya memodifikasi satu *hyperparameter* dari konfigurasi *default*, yaitu *max_depth* dari None menjadi 5, sementara empat *hyperparameter* lainnya (*n_estimators* = 100, *min_samples_split* = 2, *min_samples_leaf* = 1, *max_features* = sqrt) tetap identik dengan Baseline RF. Hal ini menjelaskan secara langsung mengapa performa *Random Search* RF pada seluruh metrik berbasis threshold identik dengan Baseline RF. Pada dasarnya, *Random Search* menemukan bahwa konfigurasi *default* Random Forest sudah mendekati optimal untuk dataset WDBC, dengan satu-satunya modifikasi pada *max_depth* yang tidak berdampak

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1489-1497
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1087

signifikan terhadap metrik berbasis threshold namun sedikit meningkatkan AUC-ROC dari 0,9929 menjadi 0,9950. Temuan ini memiliki implikasi metodologis penting: pada dataset dengan separabilitas kelas yang tinggi dan ukuran yang relatif terbatas seperti WDBC, parameter *default* scikit-learn untuk Random Forest sudah dirancang secara konservatif untuk menghasilkan generalisasi yang baik, sehingga ruang peningkatan melalui tuning menjadi sempit [6].

Sebaliknya, *Grid Search* RF memodifikasi dua *hyperparameter* secara bersamaan dari konfigurasi *default*: *max_depth* dari None menjadi 10 dan *max_features* dari sqrt menjadi log2, disertai peningkatan *n_estimators* dari 100 menjadi 200. Perubahan *max_features* dari sqrt menjadi log2 merupakan faktor kunci penurunan performa log2 menghasilkan jumlah fitur yang lebih sedikit pada setiap *split* ($\log_2 30 \approx 4{,}9$ fitur vs $\sqrt{30} \approx 5{,}5$ fitur), yang pada dataset berdimensi 30 fitur dengan multikolinearitas tinggi justru membatasi ekspresivitas model [6]. Meskipun selisih matematis antara log2 dan sqrt sangat marginal dalam nilai absolut, dampaknya terhadap generalisasi pada data uji terbukti signifikan. Kombinasi pembatasan *max_depth* = 10 dan *max_features* = log2 yang dipilih *Grid Search* mencerminkan fenomena *optimizer's curse*: pencarian *exhaustive* yang mengevaluasi seluruh 288 kombinasi menemukan kombinasi yang unggul pada cross-validation namun tidak menghasilkan generalisasi optimal pada data uji, terutama karena selisih CV F1-score yang sangat tipis antara kombinasi terpilih (0,9557) dan kombinasi lain dalam ruang pencarian rentan terhadap variance tinggi pada dataset berukuran 455 sampel data latih [18], [23]. Temuan ini konsisten dengan observasi Rahman et al. [7] yang juga menemukan bahwa Random Forest tidak menunjukkan peningkatan signifikan setelah optimasi *Grid Search* pada dataset WDBC, dengan akurasi tetap pada 97,08%.

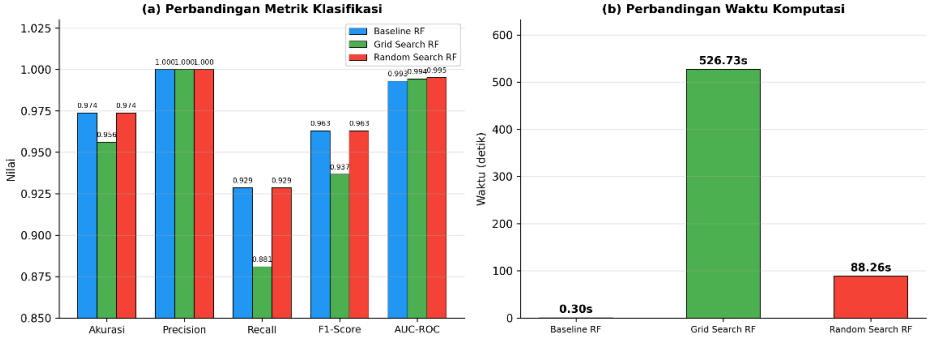


Gambar 5. ROC *Curve* Perbandingan Ketiga Model

Meskipun metrik berbasis threshold seperti akurasi dan F1-score tidak menunjukkan peningkatan setelah tuning, nilai AUC-ROC justru meningkat secara konsisten dari 0,9929 pada Baseline RF menjadi 0,9942 pada *Grid Search* RF dan 0,9950 pada *Random Search* RF sebagaimana ditunjukkan pada Gambar 5. Peningkatan AUC-ROC mengindikasikan bahwa kedua metode tuning berhasil meningkatkan kemampuan diskriminasi model secara keseluruhan pada berbagai threshold klasifikasi, meskipun pada threshold default 0,5 peningkatan tersebut tidak selalu terwujud dalam metrik berbasis threshold [18]. *Random Search* RF mencapai AUC-ROC tertinggi sebesar 0,9950, menjadikannya model dengan kemampuan diskriminasi terbaik di antara ketiga model yang diuji.

3.3 Analisis Efisiensi Komputasi

Analisis waktu komputasi mengungkap perbedaan efisiensi yang sangat signifikan antara *Grid Search* dan *Random Search* sebagaimana ditunjukkan pada Gambar 6(b). *Grid Search* membutuhkan waktu 526,73 detik untuk mengevaluasi seluruh 1.440 *fitting* (288 kombinasi × 5-fold CV), sementara *Random Search* hanya membutuhkan 88,26 detik untuk 250 *fitting* (50 iterasi × 5-fold CV). *Random Search* terbukti 5,97 kali lebih cepat dibandingkan *Grid Search*, sekaligus hanya menggunakan 17,4% dari total ruang pencarian *Grid Search*. Sementara itu, Gambar 6(a) menunjukkan bahwa *Random Search* RF tetap mampu mempertahankan performa klasifikasi yang kompetitif dibandingkan Baseline RF meskipun dengan waktu komputasi yang jauh lebih singkat.



Gambar 6. Perbandingan Performa dan Efisiensi Komputasi Ketiga Model Random Forest

Temuan ini secara empiris mengkonfirmasi proposisi teoritis Bergstra dan Bengio [10] bahwa *Random Search* mampu menemukan konfigurasi yang setara atau lebih baik dengan biaya komputasi yang jauh lebih rendah. Dalam konteks praktis, penghematan waktu komputasi sebesar 438,47 detik per eksperimen menjadi sangat relevan ketika optimasi perlu dilakukan berulang kali pada pipeline *machine learning* skala besar [14], [20]. Seluruh eksperimen dijalankan menggunakan Google Colaboratory versi gratis dengan spesifikasi default, yakni prosesor Intel Xeon CPU @ 2,20GHz dengan 2 virtual CPU, RAM sebesar 12,7 GB, sistem operasi Linux 6.6.113, dan Python versi 3.12.13 tanpa akselerasi GPU. Spesifikasi ini merepresentasikan kondisi komputasi yang realistis dan dapat direproduksi oleh peneliti lain tanpa memerlukan infrastruktur komputasi khusus, sehingga perbandingan waktu komputasi yang dilaporkan mencerminkan kondisi penggunaan umum di lingkungan riset dengan sumber daya terbatas. Karena Colaboratory gratis menggunakan *shared resources*, waktu eksperimen dapat bervariasi antar sesi tergantung beban server pada saat eksperimen dijalankan.

3.4 Analisis Trade-off Performa vs Efisiensi Komputasi

Hasil penelitian ini menunjukkan bahwa pada klasifikasi kanker payudara menggunakan dataset WDBC, *Random Search* memberikan keunggulan ganda dibandingkan *Grid Search*: performa klasifikasi yang setara dengan Baseline RF sekaligus efisiensi komputasi yang 5,05 kali lebih tinggi dibandingkan *Grid Search* *Grid Search* yang secara teoritis lebih exhaustive justru menghasilkan performa di bawah baseline, mengindikasikan bahwa kelengkapan pencarian tidak selalu berkorelasi positif dengan kualitas solusi yang dihasilkan, terutama pada dataset dengan karakteristik yang relatif bersih dan berdimensi terbatas [9], [20].

Perbandingan konfigurasi *hyperparameter* terpilih antara *Grid Search* dan *Random Search* menunjukkan bahwa kedua metode menemukan solusi yang berbeda meskipun mengeksplorasi ruang pencarian yang identik. *Grid Search* dengan pencarian exhaustive justru terjebak pada kombinasi yang overfit terhadap proses cross-validation, sementara *Random Search* dengan eksplorasi acak berhasil menemukan kombinasi yang lebih generalizable pada data uji. Perlu dicatat bahwa perbedaan Recall antara *Grid Search* RF (0,8810) dan kedua model lainnya (0,9286) setara dengan 2 sampel dari 42 sampel *Malignant* pada test set berukuran terbatas, sehingga interpretasinya perlu dilakukan dengan hati-hati tanpa dukungan uji signifikansi statistik. Temuan ini memperkuat argumen bahwa karakteristik dataset terutama ukuran dan tingkat kebersihan data memiliki pengaruh besar terhadap efektivitas strategi optimasi *hyperparameter* yang dipilih [14].

Dari perspektif klinis, model dengan Recall tertinggi adalah prioritas utama dalam konteks skrining kanker karena meminimalkan kasus kanker yang tidak terdeteksi [18]. Baseline RF dan *Random Search* RF yang mencapai Recall 0,9286 lebih unggul secara klinis dibandingkan *Grid Search* RF dengan Recall 0,8810. Dalam konteks skrining kanker payudara, perbedaan ini tetap memiliki implikasi klinis yang perlu diperhatikan meskipun validasi pada dataset yang lebih besar diperlukan untuk mengkonfirmasi konsistensi temuan ini.

Mempertimbangkan seluruh dimensi evaluasi akurasi, F1-score, AUC-ROC tertinggi (0,9950), Recall klinis, dan efisiensi komputasi, *Random Search* RF merupakan pendekatan optimasi yang paling direkomendasikan untuk klasifikasi kanker payudara berbasis Random Forest pada dataset WDBC. *Random Search* terbukti mampu menghasilkan performa setara atau lebih baik dari baseline dengan biaya komputasi yang jauh lebih efisien dibandingkan *Grid Search*, menjadikannya pilihan yang lebih pragmatis untuk pipeline machine learning pada dataset medis berukuran sedang dengan separabilitas kelas tinggi.

4. KESIMPULAN

Penelitian ini membandingkan *Grid Search* dan *Random Search* untuk optimasi *hyperparameter* Random Forest pada klasifikasi kanker payudara menggunakan dataset WDBC. Hasil eksperimen menunjukkan bahwa *Random Search* RF mencapai performa setara dengan Baseline RF (akurasi 0,9737; F1-Score 0,9630; AUC-ROC 0,9950) sekaligus mengungguli *Grid Search* RF (akurasi 0,9561; F1-Score 0,9367) yang mengalami fenomena *optimizer's curse*. Dari sisi efisiensi, *Random Search* terbukti 5,97 kali lebih cepat dengan hanya mengeksplorasi 17,4% ruang pencarian, mengkonfirmasi proposisi teoritis Bergstra dan Bengio [10]. Secara klinis, *Recall Random Search* RF (0,9286) juga lebih unggul dibanding *Grid Search* RF (0,8810), setara dengan 2 kasus *Malignant* tambahan yang tidak terdeteksi. Meski demikian, generalisasi hasil ini perlu divalidasi lebih lanjut mengingat eksperimen hanya dilakukan pada satu dataset benchmark berukuran sedang. Penelitian lanjutan disarankan mencakup validasi pada dataset klinis lebih besar, perbandingan dengan Bayesian Optimization, serta eksplorasi kombinasi *Random Search* dengan seleksi fitur .

REFERENCES

[1] F. Bray *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834.

[2] J. Ferlay *et al.*, "Global Cancer Observatory: Cancer Today — Indonesia Fact Sheet (GLOBOCAN 2022)," International Agency for Research on Cancer (IARC). Accessed: Mar. 09, 2026. [Online]. Available: https://gco.iarc.who.int/media/globocan/factsheets/populations/360-indonesia-fact-sheet.pdf

[3] I. B. Setyawan *et al.*, "Sociodemographic disparities associated with advanced stages and distant metastatic breast cancers at diagnosis in Indonesia: a cross-sectional study," *Annals of Medicine & Surgery*, vol. 85, no. 9, pp. 4211–4217, 2023, doi: 10.1097/ms9.0000000000001030.

[4] K. Lemons and I. W. Rathnathungalage, "A Comparison Between Naïve Bayes and Random Forest to Predict Breast Cancer," *International Journal of Undergraduate Research and Creative Activities*, vol. 12, no. 1, p. 1, 2023, doi: 10.7710/2168-0620.0287.
[5] A. Batool and Y. C. Byun, "Breast Cancer Classification using Random Forest Algorithm," in *Journal of Physics: Conference Series*, 2023. doi: 10.1088/1742-6596/2559/1/012002.
[6] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[7] Md. M. Rahman, A. Rahman, S. Akter, and S. A. Pinky, "Hyperparameter Tuning Based Machine Learning Classifier for Breast Cancer Prediction," *Journal of Computer and Communications*, vol. 11, no. 04, pp. 149–165, 2023, doi: 10.4236/jcc.2023.114007.
[8] P. Jain, S. Aggarwal, S. Adam, and M. Imam, "Parametric optimization and comparative study of machine learning and deep learning algorithms for breast cancer diagnosis," *Breast Dis.*, vol. 43, no. 1, pp. 257–270, 2024, doi: 10.3233/BD-240018.
[9] M. Sholeh, U. Lestari, and D. Andayati, "Hyperparameter Optimization Using Grid Search and Random Search to Improve the Performance of Prediction Models with Decision Trees," *Jurnal Riset Multidisiplin dan Inovasi Teknologi*, vol. 3, no. 03, pp. 453–464, 2025, doi: 10.59653/jimat.v3i03.2025.
[10] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
[11] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository. Accessed: Apr. 08, 2026. [Online]. Available: https://doi.org/10.24432/C5DW2B
[12] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, vol. 1905, no. 870, pp. 861–870, 1993, doi: 10.1117/12.148698.
[13] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems*, 3rd ed. United States of America: O'Reilly Media, 2022.
[14] Q. A. Hidayaturrohman and E. Hanada, "A Comparative Analysis of Hyper-Parameter Optimization Methods for Predicting Heart Failure Outcomes," *Applied Sciences (Switzerland)*, vol. 15, no. 6, p. 3393, 2025, doi: 10.3390/app15063393.
[15] M. Quintana *et al.*, "It is not always greener on the other side: Greenery perception across demographics and personalities in multiple cities," *Landsc. Urban Plan.*, vol. 271, p. 105618, 2026, doi: 10.1016/j.landurbplan.2026.105618.
[16] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, [Online]. Available: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
[17] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
[18] Y. A. Ali, E. M. Awwad, M. Al-Razgan, and A. Maarouf, "Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity," *Processes*, vol. 11, no. 2, p. 349, 2023, doi: 10.3390/pr11020349.
[19] scikit-learn developers, "GridSearchCV — scikit-learn 1.9.0 documentation." 2026. [Online] Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[20] J. Li, "Area under the ROC Curve has the most consistent evaluation for binary classification," *PLoS One*, vol. 19, no. 12 December, 2024, doi: 10.1371/journal.pone.0316019.
[21] scikit-learn developers, "RandomizedSearchCV — scikit-learn 1.9.0 documentation," 2026. [Online] Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
[22] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-56706-x.
[23] L. Schneider, B. Bischl, and M. Feurer, "Overtuning in Hyperparameter Optimization," *Proc. Mach. Learn. Res.*, vol. 293, pp. 1–43, 2025, doi: https://doi.org/10.48550/arXiv.2506.19540.