Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.503**
Page 159-172

Analisis Perbandingan Metode Support Vector Machine dan Random Forest dalam Prediksi Kanker Payudara

Nathasya Taruli Aprillia Purba*✉, Afninda Yovini Lumban Toruan✉, Elvis Sastra Ompusunggu✉
Sistem Informasi, Universitas Prima Indonesia, Indonesia

Abstrak: Salah satu jenis kanker yang memiliki prevalensi dan angka kematian tertinggi di seluruh dunia adalah kanker payudara. Metode diagnosis yang akurat dan efisien diperlukan karena deteksi dini berperan penting dalam meningkatkan kemungkinan pasien sembuh. Penelitian ini bertujuan untuk menguji dan membandingkan tingkat kinerja algoritma Support Vector Machine (SVM) serta Random Forest dalam memprediksi kasus kanker payudara. Kajian ini memanfaatkan dataset Breast Cancer Wisconsin (Diagnostic) yang mencakup 30 variabel numerik guna menggambarkan karakteristik sel dari 569 pasien. Status diagnosis pada data tersebut terbagi menjadi dua kategori, yaitu tumor jinak (benign) dan tumor ganas (malignant). Pada 114 data uji, SVM dan Random Forest menghasilkan matriks konfusi yang sama (TP = 39, TN = 72, FP = 0, FN = 3), sehingga akurasi 97,37%, presisi 100%, sensitivitas 92,86%, spesifisitas 100%, dan F1-score 96,30% bernilai identik. Perbedaan hanya tampak pada AUC, yaitu 1,00 untuk SVM dan 0,99 untuk Random Forest pada pembulatan dua desimal. Seluruh temuan ini diperoleh melalui prosedur sistematis mencakup pengumpulan data, tahap pra-pemrosesan, pemisahan proporsi data sebesar 80% untuk pelatihan dan 20% untuk pengujian, normalisasi nilai fitur menggunakan StandardScaler yang hanya dipelajari dari data latih, proses pelatihan model, sampai tahap pengujian performa. Evaluasi kinerja dilakukan berdasarkan metrik akurasi, presisi, sensitivitas (recall), spesifisitas, F1-score, serta Area Under Curve (AUC). Kedua algoritma menunjukkan kinerja yang setara dalam klasifikasi akhir, dan keunggulan SVM terbatas pada AUC dengan selisih yang kecil.
Kata Kunci: Kanker Payudara; Support Vector Machine; Random Forest; Machine Learning; Prediksi Klasifikasi

Abstract: Breast cancer is one of the types of cancer with the highest prevalence and mortality rates worldwide. Accurate and efficient diagnostic methods are essential, as early detection plays a crucial role in improving patients' chances of recovery. This study aims to test and compare the performance of the Support Vector Machine (SVM) and Random Forest algorithms in predicting breast cancer cases. This study utilizes the Wisconsin Breast Cancer (Diagnostic) dataset, which includes 30 numerical variables to describe the cellular characteristics of 569 patients. The diagnostic status in the dataset is divided into two categories: benign and malignant tumors. On the 114-sample test set, SVM and Random Forest produced identical confusion matrices (TP = 39, TN = 72, FP = 0, FN = 3), resulting in identical accuracy of 97.37%, precision of 100%, sensitivity of 92.86%, specificity of 100%, and an F1-score of 96.30%. The only difference was observed in the AUC, which was 1.00 for SVM and 0.99 for Random Forest when rounded to two decimal places. All of these findings were obtained through a systematic

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.503**
Page 159-172

procedure that included data collection, preprocessing, splitting the dataset into 80% for training and 20% for testing, normalizing feature values using StandardScaler—which was trained solely on the training data—model training, and performance testing. Performance evaluation was conducted based on the metrics of accuracy, precision, sensitivity (recall), specificity, F1-score, and Area Under the Curve (AUC). Both algorithms demonstrated comparable performance in the final classification, and SVM's advantage was limited to the AUC, with only a slight difference.
Keywords: Breast Cancer; Support Vector Machine; Random Forest; Machine Learning; Classification Prediction

1. PENDAHULUAN

Kanker payudara merupakan salah satu tipe keganasan yang sangat banyak ditemukan sekaligus jadi pemicu utama kematian akibat kanker secara luas, dengan populasi perempuan sebagai kelompok yang paling rentan. Informasi epidemiologi terbaru mengindikasikan bahwa angka kejadian penyakit ini terus menampilkan tren kenaikan di berbagai belahan dunia, sehingga menimbulkan beban dan tantangan besar untuuk layanan kesehatan modern [1], [2]. Merujuk pada data World Health Organization (WHO), kanker payudara menduduki peringkat teratas sebagai tipe keganasan yang sangat kerap di nyatakan pada wanita di segala dunia, dengan catatan lebih dari 2, 3 juta permasalahan baru dan dekat 670. 000 jiwa kematian pada tahun 2022. Lonjakan angka tersebut menjadikan patologi ini selaku fokus utama dalam pengembangan sistem penaksiran berbasis kecerdasan buatan (Artificial Intelligence), yang diarahkan untuk membantu klinisi dalam mengenali penyakit secara lebih cepat dan akurat [1]. Dalam upaya penanganannya, identifikasi penyakit pada stadium awal terbukti efektif dalam memperbesar peluang kesembuhan dan tingkatkan angka harapan hidup penderita, khususnya apabila diagnosis dilakukan pada tahap awal perkembangan sel kanker [14]. Berdasarkan hal tersebut, pengembangan metode deteksi yang tepat dan efisien merupakan kebutuhan yang mendesak.

Mammografi dan biopsi jaringan merupakan pemeriksaan medis secara tradisional dalam diagnosis kanker payudara. Meskipun metode tersebut masih menjadi standar klinis, upaya ini memiliki sejumlah keterbatasan, antara lain biaya yang relatif tinggi, sifatnya yang invasif, serta ketergantungan pada interpretasi team medis yang dapat memunculkan kesalahan diagnosis [1], [11]. Di bidang kesehatan, pemilihan algoritma machine learning tidak cukup berpacu pada tingkat akurasi, akan tetapi harus menilai seberapa tepat model mengenali kasus positif. Pada kanker payudara, False Negative dapat menyebabkan diagnosis dan penanganan tertunda, sehingga berpotensi memperbesar risiko komplikasi maupun kematian [1], [18]. Situasi demikian memicu adopsi teknologi kecerdasan buatan, terpenting pada pendekatan machine learning, sebagai solusi potensial untuk memaksimalkan akurasi dan efisiensi dalam mendeteksi kanker payudara melalui analisis data klinis. Untuk itu, evaluasi dilakukan dengan memperhitungkan sensitivitas, spesifisitas, dan Area Under Curve (AUC) pada samping akurasi, sehingga bisa dinilai secara lebih komprehensip pada kualitas prediksi tiap algoritma.

Sejumlah literatur terdahulu mencatat Support Vector Machine menampilkan performa yang lebih bagus tatkala mengolah data berdimensi tinggi, sementara riset lainnya menggarisbawahi keunggulan Random Forest dalam menangani keragaman data serta mekanisme seleksi fitur [11], [12], [13], [15]. Konsekuensinya, hingga saat ini belum tercapai konsensus ilmiah yang seragam mengenai algoritma terbaik untuk klasifikasi kasus kanker payudara. Berdasarkan kesenjangan tersebut, studi komparatif mendalam mutlak diperlukan guna mengidentifikasi metode yang mampu menghadirkan tingkat performa tertinggi secara spesifik pada dataset Breast Cancer Wisconsin (Diagnostic)

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.503**
Page 159-172

Sejumlah literatur terdahulu mencatat Support Vector Machine menampilkan performa yang lebih bagus tatkala mengolah data berdimensi tinggi, sementara riset lainnya menggarisbawahi keunggulan Random Forest dalam menangani keragaman data serta mekanisme seleksi fitur [11], [12], [13], [15]. Konsekuensinya, hingga saat ini belum tercapai konsensus ilmiah yang seragam mengenai algoritma terbaik untuk klasifikasi kasus kanker payudara. Berdasarkan kesenjangan tersebut, studi komparatif mendalam mutlak diperlukan guna mengidentifikasi metode yang mampu menghadirkan tingkat performa tertinggi secara spesifik pada dataset Breast Cancer Wisconsin (Diagnostic). Walaupun SVM dan Random Forest sama-sama berkinerja sangat baik, studi-studi sebelumnya mencatat adanya perbedaan hasil pada dataset yang sama. Sebagian studi melaporkan keunggulan SVM dalam hal akurasi. Pada sudut pandang lain, studi terpisah mengindikasikan kalau model Random Forest menawarkan tingkatan stabilitas yang lebih besar dan ketahanan (robustness) yang baik terhadap fluktuasi ataupun anomali data (noise) [3], [7], [8]. Perbedaan temuan tersebut mengindikasikan bahwa algoritma yang tepat sangat bergantung pada karakteristik data, teknik pra-pemrosesan, serta parameter model yang digunakan [13], [14], [15].

Penerapan machine learning dapat mendorong atas kemajuan teknologi yang semakin luas di sektor kesehatan untuk mendukung diagnosis penyakit. Dengan mempelajari pola pada data medis, algoritma machine learning menciptakan model prediksi yang bisa berperan sebagai sistem pendukung sebuah keputusan. Berbagai literatur riset memperlihatkan bahwa penerapan metode pada diagnosis kanker payudara terbukti mampu mendongkrak tingkat ketepatan analisis dan juga menekan potensi kesalahan interpretasi yang diakibatkan oleh faktor subjektivitas manusia [11], [12], [15]. Dataset Breast Cancer Wisconsin (Diagnostic) merupakan salah satu basis data yang diimplementasikan dalam penelitian prediksi kanker payudara, yang memuat 569 rekam medis pasien serta 30 variabel numerik penciri karakteristik inti sel. Segala entri informasi dibagi ke dalam 2 kelas, ialah tumor jinak (benign) serta tumor ganas (malignant) [3], [5]. Dari struktur informasinya yang terstandarisasi dan jelas, dataset ini sangat ideal untuk implementasi algoritma supervised learning serta kerap diadopsi dalam berbagai studi komparatif machine learning di ranah medis [14]. Algoritma Support Vector Machine (SVM) diakui mempunyai kinerja handal dalam mencerna informasi yang besar dan berdimensi tinggi. Tata cara ini bekerja dengan metode membangun hyperplane optimal guna memaksimalkan batas pemisahan antar dua kelas data yang berbeda [6][10]. Dalam konteks klasifikasi kanker payudara, sejumlah kajian empiris melaporkan bahwa SVM kerap mengungguli performa algoritma lain semacam K-Nearest Neighbor (KNN), Logistic Regression, dan Decision Tree [2], [6], [10]. Lebih lanjut, lewat penerapan normalisasi yang benar serta pemilihan guna kernel yang sesuai, metode ini tercatat mampu mencatatkan tingkat akurasi melampaui 97% pada dataset Breast Cancer Wisconsin [11], [13]. Berbeda dari mekanisme SVM, Random Forest bekerja dengan pendekatan ensemble learning yang mengintegrasikan beberapa besar tumbuhan keputusan (decision tree) serta memastikan hasil akhir klasifikasi lewat mekanisme pemungutan suara kebanyakan (majority voting) [7],[10]. Keunggulan utama dari algoritma ini mencakup ketangguhannya dalam menanggulangi struktur informasi yang rumit, potensi overfitting yang relatif kecil, serta stabilitas performa yang terjaga di tengah berbagai keragaman karakteristik data [5], [7], [12]. Penelitian terbaru juga menunjukkan bahwa hasil klasifikasinya kompetitif dan pada sebagian kasus dapat melebihi SVM, khususnya jika didukung feature selection yang tepat [12], [13], [15].

Melalui pemanfaatan dataset Breast Cancer Wisconsin (Diagnostic), studi ini melaksanakan evaluasi komparatif secara empiris terhadap kinerja algoritma Support Vector Machine dan Random Forest untuk keperluan prediksi kanker payudara, di mana seluruh rangkaian eksperimen serta pengolahan datanya dieksekusi melalui lingkungan

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

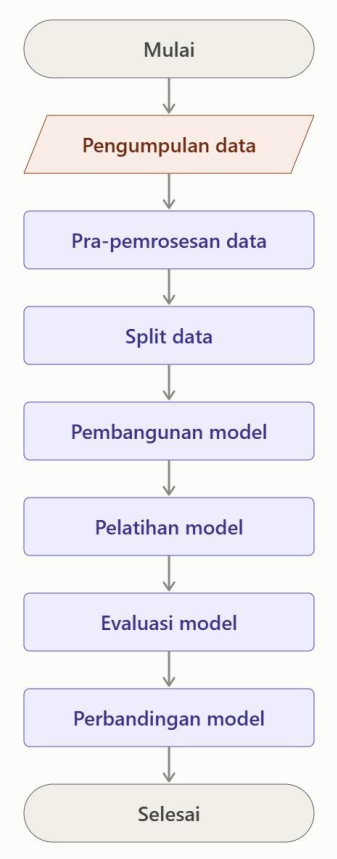
Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.503**
Page 159-172

Google Colab. Penemuan dari studi ini diharapkan sanggup memberikan rekomendasi model klasifikasi yang sangat akurat serta efektif, sekaligus menjadi rujukan berharga dalam pengembangan sistem pendukung keputusan klinis berbasis machine learning guna menunjang optimalisasi deteksi awal mula kanker payudara.

2. METODE PENELITIAN

2.1. Desain Penelitian

Studi ini menerapkan desain komparatif kuantitatif berbasis data sekunder guna mengevaluasi serta membandingkan performa dua algoritma machine learning, yaitu Support Vector Machine (SVM) dan Random Forest (RF), dalam konteks prediksi kanker payudara. Penggunaan pendekatan kuantitatif didasari oleh kebutuhan untuk mengolah variabel numerik secara sistematis serta mengukur performa model melalui metrik statistik yang objektif. Lebih lanjut, karakteristik basis data yang bersifat terbuka (open dataset) memungkinkan rangkaian analisis komparatif ini dieksekusi secara mandiri tanpa memerlukan intervensi langsung terhadap subjek penelitian di dunia nyata.



Gambar 1. Flowchart Penelitian

Gambar diatas menjelaskan tahapan penelitian meliputi (1) Pengumpulan Data: Data hasil pemeriksaan sel payudara pasien untuk diagnosis kanker menjadi isi dataset Breast Cancer Wisconsin (Diagnostic) yang dipakai dalam penelitian ini. Pada fase ini, dataset tersebut dikumpulkan dari Kaggle, platform terbuka yang terpercaya; (2) Pra-pemrosesan Data: meliputi penghapusan atribut yang tidak diperlukan, pengkodean label, serta pemisahan fitur dan target agar data siap diolah; (3) Split Data: Dataset yang telah bersih kemudian dibagi menjadi data latih (training data) dan data uji (testing data) dengan

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.503**
Page 159-172

proporsi 80% untuk melatih model dan 20% untuk mengevaluasi kinerjanya. Normalisasi fitur dilakukan sesudah pemisahan ini dan hanya mempelajari data latih; (4) Pembangunan Model: Model dibangun menggunakan Support Vector Machine (SVM), yang menentukan hyperplane optimal untuk memisahkan dua kelas data [6], dan Random Forest (RF), yang menggabungkan banyak pohon keputusan melalui voting [7]. Konfigurasi parameter kedua model dirinci pada Tabel 2; (5) Pelatihan Model: Pelatihan dilakukan supaya model dapat menangkap pola keterkaitan antara fitur sel payudara dan kelas targetnya, baik jinak maupun ganas. Untuk keperluan itu, model yang sudah disusun dilatih menggunakan data latih. Pelatihan dilakukan menggunakan fungsi fit() pada masing-masing algoritma; (6) Evaluasi Model: Metrik evaluasi yakni akurasi, sensitivitas (recall), spesifisitas, dan AUC (Area Under Curve), dipakai untuk mengevaluasi model yang sudah dilatih dengan data uji. Hasilnya menunjukkan sejauh mana model mampu mengklasifikasikan kanker payudara; (7) Perbandingan Model: Hasil evaluasi menjadi dasar perbandingan antara Support Vector Machine dan Random Forest pada fase penutup. Dari perbandingan itu, ditentukan model algoritma yang paling optimal untuk prediksi kanker payudara.

2.2. Instrumen Penelitian

Bahasa pemrograman utama untuk mengolah data dan menerapkan machine learning. Untuk mengumpulkan, mengolah, dan menganalisis data, penelitian memerlukan instrumen sebagai alat bantu. Adapun instrumen dalam penelitian ini terdiri dari:

a) Google Colab
Berfungsi sebagai lingkungan kerja daring untuk menulis, menjalankan, dan mengelola kode program.

b) Python
Bahasa pemrograman utama bagi tahap pengolahan data dan menerapkan algoritma machine learning.

c) Scikit-learn
Dipakai untuk mengimplemetasikan algoritma Support Vector Machine dan Random Forest, sekaligus mengevaluasi model [10].

d) Matplotlib
Dipakai untuk visualisasi hasil analisis, termasuk grafik perbandingan dan kurva ROC.

e) Metrik Evaluasi
Menjadi indikator untuk menilai cara kerja model, yaitu akurasi, sensitivitas (recall), spesifisitas, dan AUC yang dihitung dari data uji.

2.3. Dataset Penelitian

Populasi dalam riset ini merepresentasikan rekam data pasien yang mencakup berbagai parameter klinis dari hasil pemeriksaan sel payudara. Seluruh entri data tersebut dikelompokkan ke dalam dua kategori kelas kondisi, yakni tumor payudara jinak dan tumor ganas. Melalui platform Kaggle (https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data), dataset publik "Breast Cancer Wisconsin (Diagnostic)" dapat diakses dan dijadikan sampel dalam penelitian ini. Basis data ini memuat 569 rekam medis pasien yang didukung oleh 30 atribut numerik penciri morfologi sel payudara meliputi parameter inti, tekstur, keliling (perimeter), luas area, hingga tingkat kehalusan serta dilengkapi satu kolom label target yang membagi status pasien ke dalam kategori tumor jinak benign (B) atau ganas malignant (M) [3], [5]. Berkas CSV dari Kaggle berisi 33 kolom, yaitu kolom id, kolom diagnosis, 30 fitur numerik, dan satu kolom kosong bernama 'Unnamed: 32'. Dua kolom yang tidak membawa informasi (id dan Unnamed: 32) dibuang pada tahap pra-pemrosesan, sehingga data yang dianalisis terdiri atas 569 sampel, 30 fitur numerik, dan 1 label target.

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.503**
Page 159-172

3. HASIL DAN PEMBAHASAN

3.1. Hasil Koleksi Data

Bersumber dari UCI Machine Learning Repository, dataset publik "Breast Cancer Wisconsin (Diagnostic) Dataset" diperoleh untuk studi ini melalui situs Kaggle. Isinya adalah data inti sel kanker payudara yang bersumber dari digitasi citra aspirasi jarum halus (FNA) pada massa payudara.

```
df = pd.read_csv('data.csv')

print(f"Shape dataset : {df.shape}")
print(f"Kolom         : {list(df.columns[:5])} ...")
print("\nDistribusi kelas:")
print(df['diagnosis'].value_counts())
df.head()
```

```
Shape dataset : (569, 33)
Kolom         : ['id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean'] ...

Distribusi kelas:
diagnosis
B    357
M    212
Name: count, dtype: int64
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280

5 rows × 33 columns

Gambar 2. Hasil Koleksi Data

Sebanyak 569 sampel tercatat dalam dataset ini, dengan 33 kolom pada berkas aslinya (id, diagnosis, 30 fitur numerik, dan kolom kosong 'Unnamed: 32'). Setelah kolom id dan 'Unnamed: 32' dibuang, setiap sampel dideskripsikan oleh 30 fitur numerik dan 1 kolom target berupa diagnosis. Label pada kolom target ada dua, yaitu M (Malignant/Ganas) dan B (Benign/Jinak). Dari total 569 sampel, terdapat 212 sampel dengan diagnosis ganas (M) dan 357 sampel dengan diagnosis jinak (B). Sepuluh karakteristik inti sel, yaitu radius, tekstur, perimeter, area, kehalusan, kekompakan, konkavitas, titik cekung, simetri, dan dimensi fraktal, masing-masing diukur dalam tiga bentuk (mean, standard error, dan nilai terburuk), sehingga setiap sampel memiliki 30 fitur numerik.

3.2. Analisis Karakteristik Dataset

Sebelum pemodelan, dilakukan pemeriksaan awal terhadap sebaran kelas dan karakteristik fitur. Dari 569 sampel, 357 sampel (62,74%) berlabel jinak dan 212 sampel (37,26%) berlabel ganas, atau sekitar 1,68 sampel jinak untuk setiap sampel ganas (Gambar 3). Ketidakseimbangan ini tergolong ringan, tetapi cukup untuk membuat akurasi saja kurang memadai. Model yang cenderung memilih kelas jinak tetap dapat memperoleh akurasi di atas 62% tanpa mengenali kasus ganas sama sekali. Karena itu, sensitivitas, spesifisitas, dan AUC dilaporkan bersama akurasi.

Ringkasan statistik pada Tabel 1 memperlihatkan bahwa kedua kelas memiliki karakteristik sel yang berbeda. Sel pada tumor ganas rata-rata berukuran lebih besar, dengan radius rata-rata 17,46 dibanding 12,15 pada tumor jinak, serta luas area rata-rata 978,38 dibanding 462,79. Perbedaan yang lebih tajam terlihat pada bentuk kontur sel. Rata-rata konkavitas pada kelas ganas (0,161) sekitar 3,5 kali kelas jinak (0,046), dan rata-rata titik cekung (concave points) pada kelas ganas (0,088) sekitar 3,4 kali kelas jinak

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : https://doi.org/10.58602/itsecs.v5i1.503
Page 159-172

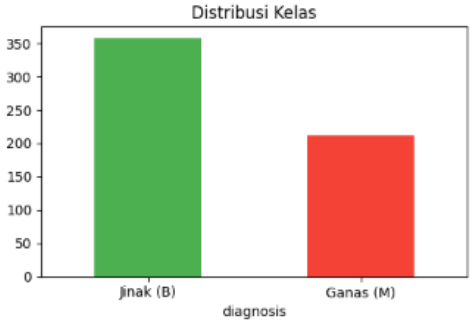
(0,026). Sebaliknya, dimensi fraktal rata-rata hampir sama pada kedua kelas (0,063), sehingga fitur ini nyaris tidak membantu membedakan kelas. Fitur yang paling memisahkan kedua kelas adalah fitur-fitur dengan nilai terburuk (worst), terutama worst concave points dan worst perimeter.

Tabel 1. Statistik Fitur Terpilih menurut Kelas Diagnosis (Rata-rata ± Simpangan Baku)

Fitur	Jinak (B), n = 357	Ganas (M), n = 212
Radius rata-rata	12,15 ± 1,78	17,46 ± 3,20
Tekstur rata-rata	17,92 ± 4,00	21,61 ± 3,78
Perimeter rata-rata	78,08 ± 11,81	115,37 ± 21,86
Area rata-rata	462,79 ± 134,29	978,38 ± 367,94
Kehalusan rata-rata	0,092 ± 0,013	0,103 ± 0,013
Konkavitas rata-rata	0,046 ± 0,043	0,161 ± 0,075
Titik cekung rata-rata	0,026 ± 0,016	0,088 ± 0,034
Dimensi fraktal rata-rata	0,063 ± 0,007	0,063 ± 0,008
Titik cekung terburuk	0,074 ± 0,036	0,182 ± 0,046

Dua temuan lain dari pemeriksaan awal berkaitan dengan tahap pemodelan. Pertama, rentang nilai antarfitur sangat berbeda. Area rata-rata berkisar 143,5 sampai 2.501, sedangkan kehalusan rata-rata hanya 0,053 sampai 0,163. Selisih skala seperti ini dapat membuat fitur berskala besar mendominasi perhitungan jarak pada SVM, sehingga normalisasi diperlukan. Kedua, fitur radius, perimeter, dan area saling berkorelasi sangat kuat (koefisien korelasi radius-perimeter 0,998 dan radius-area 0,987). Simpangan baku yang cukup besar pada kelas ganas juga menunjukkan bahwa nilai kedua kelas masih saling tumpang tindih di sekitar batas keputusan, sehingga sebagian kesalahan klasifikasi pada hasil evaluasi wajar terjadi pada sampel di daerah tersebut.

3.3. Hasil Prapemrosesan Data



Gambar 3. Hasil Prapemrosesan Data

Pada fase ini, data tabular hasil pengukuran sel yang sudah dikumpulkan diolah menjadi data klasifikasi yang siap dipakai, dengan urutan langkah sebagai berikut. Tiga langkah pertama (a sampai c) bekerja pada seluruh data karena hanya membersihkan dan menyusun ulang kolom tanpa menghitung statistik apa pun dari nilai fitur. Pemisahan data dan normalisasi dilakukan sesudahnya dan dijelaskan pada subbab berikutnya. Urutan ini dijaga supaya tidak ada informasi dari data uji yang ikut masuk ke proses pelatihan.

a) Penghapusan Atribut dan Pemeriksaan Nilai Kosong: Kolom 'Unnamed: 32' ditemukan kosong seluruhnya (569 missing values) sehingga dihapus. Kolom 'id' juga dihapus karena hanya berfungsi sebagai pengenal dan tidak membawa informasi diagnostik. Setelah kedua kolom dihapus, tersisa 31 kolom (30 fitur dan 1 target) dan tidak ada nilai kosong pada kolom mana pun.

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : https://doi.org/10.58602/itsecs.v5i1.503
Page 159-172

b) Encoding Label: Kolom target 'diagnosis' yang awalnya berupa kategori teks (M dan B) dikonversi menjadi nilai numerik biner, yaitu label kategorik diganti dengan nilai numerik, yaitu 1 untuk M (Malignant/Ganas) dan 0 untuk B (Benign/Jinak), agar model machine learning dapat memprosesnya.
c) Pemisahan Fitur dan Target: Kolom diagnosis dipisahkan sebagai vektor target y, sedangkan 30 kolom fitur dijadikan matriks fitur X, sebagai persiapan untuk pembagian data latih dan data uji.

3.4. Pemisahan Data dan Normalisasi

Langkah pemisahahan data yaitu matriks fitur X dan vektor target y lebih dulu dibagi menjadi 455 data latih (80%) dan 114 data uji (20%) menggunakan fungsi train_test_split dengan test_size=0.2 dan random_state=42. Proporsi kelas pada kedua subset hampir sama dengan proporsi dataset asli. Data uji berisi 42 sampel ganas dan 72 sampel jinak (36,8% ganas), sedangkan data latih berisi 170 sampel ganas dan 285 sampel jinak (37,4% ganas), dibanding 37,3% ganas pada seluruh dataset. Langkah (e): normalisasi. Setelah data dipisah, StandardScaler dilatih (fit) hanya pada 455 data latih, sehingga rata-rata dan simpangan baku tiap fitur dihitung dari data latih saja. Parameter yang sama kemudian dipakai untuk mentransformasi data latih dan data uji (transform), tanpa menghitung ulang statistik dari data uji. Dengan begitu tidak ada kebocoran informasi (data leakage) dari data uji ke tahap pelatihan, dan hasil evaluasi tetap mencerminkan kinerja pada data yang belum pernah dilihat model. Normalisasi terutama dibutuhkan SVM karena kernel RBF bekerja berdasarkan jarak antarsampel, sedangkan Random Forest berbasis pohon pada dasarnya tidak terpengaruh oleh skala fitur. Meski begitu, kedua model dilatih pada data hasil normalisasi yang sama agar perbandingannya setara.

3.5. Pengembangan Model Support Vector Machine dan Random Forest

Dua model dibangun dengan pustaka scikit-learn dan dilatih pada data latih hasil normalisasi yang sama. Konfigurasi lengkap disajikan pada Tabel 2. Parameter yang tidak tercantum memakai nilai bawaan scikit-learn, dan pada penelitian ini tidak dilakukan penyetelan hyperparameter. Pada SVM dipakai kernel Radial Basis Function (RBF) dengan C=1.0 sebagai parameter regularisasi dan gamma='scale', yang menetapkan nilai gamma secara otomatis dari jumlah fitur dan varians data latih. Parameter probability=True diaktifkan agar probabilitas prediksi dapat dihitung untuk keperluan AUC. Pada Random Forest dipakai 100 pohon keputusan (n_estimators=100). Setiap pohon dilatih pada sampel bootstrap dari data latih dan pada tiap percabangan hanya mempertimbangkan sebagian fitur yang dipilih acak (max_features='sqrt', sekitar 5 dari 30 fitur), sehingga model ini cukup andal menghadapi data berdimensi tinggi. Parameter random_state=42 dipakai pada kedua model agar hasilnya dapat diulang.

Tabel 2. Konfigurasi Parameter Eksperimen

Model	Parameter	Nilai	Keterangan
SVM	kernel	'rbf'	Radial Basis Function
SVM	C	1.0	Parameter regularisasi (bawaan)
SVM	gamma	'scale'	1 / (jumlah fitur × varians X), dengan jumlah fitur = 30
SVM	probability	True	Mengaktifkan estimasi probabilitas untuk menghitung AUC
SVM	random_state	42	Keterulangan hasil
RF	n_estimators	100	Jumlah pohon keputusan
RF	criterion	'gini'	Ukuran kualitas pemisahan (bawaan)
RF	max_depth	None	Kedalaman pohon tidak dibatasi (bawaan)
RF	max_features	'sqrt'	Sekitar 5 dari 30 fitur pada tiap percabangan (bawaan)

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : https://doi.org/10.58602/itsecs.v5i1.503
Page 159-172

RF	bootstrap	True	Tiap pohon dilatih pada sampel bootstrap (bawaan)
RF	random_state	42	Keterulangan hasil
Data	train_test_split	test_size=0.2, random_state=42	455 data latih dan 114 data uji
Data	StandardScaler	fit pada data latih	Data uji hanya ditransformasi

3.6. Pelatihan Model

Model Support Vector Machine dilatih untuk menemukan hyperplane optimal yang memisahkan kelas ganas (M) dan jinak (B) dalam ruang fitur berdimensi tinggi. Dengan menggunakan kernel RBF, model ini mampu membuat batas keputusan non-linear yang lebih sesuai untuk data medis yang kompleks. Metode bootstrap aggregating(bagging) dipakai supaya masing- masing tumbuhan keputusan dalam Random Forest dilatih pada subset informasi latih yang diseleksi secara acak. Sehabis segala tumbuhan tercipta, keputusan akhir model diambil bersumber pada suara paling banyak(voting kebanyakan) dari seluruh tumbuhan tersebut.

```
# Latih kedua model
model_svm.fit(X_train, y_train)
model_rf.fit(X_train, y_train)

print("✅ Pelatihan model selesai!")
print(f"SVM          : {model_svm}")
print(f"Random Forest: {model_rf}")
```

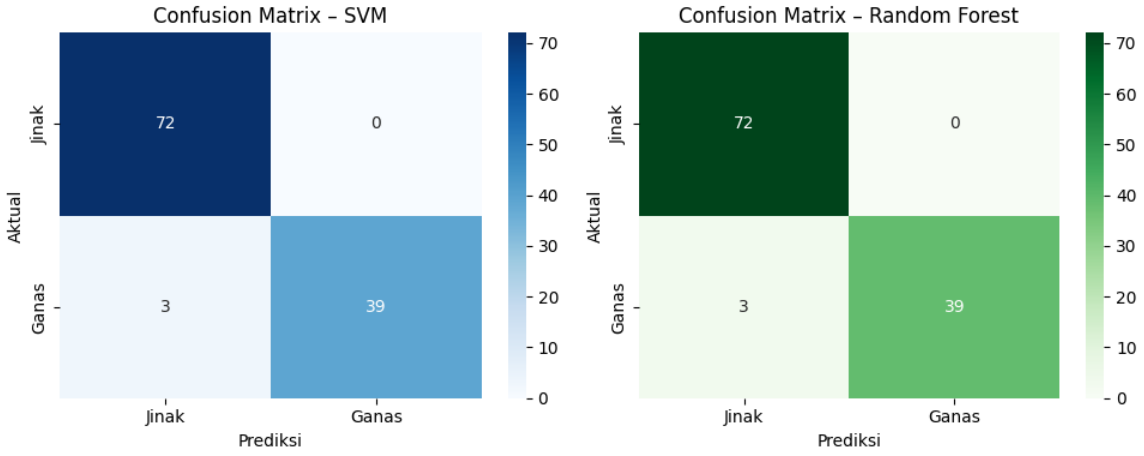
```
✅ Pelatihan model selesai!
SVM          : SVC(probability=True, random_state=42)
Random Forest: RandomForestClassifier(random_state=42)
```

Gambar 4. Pelatihan Model

Lewat metode .fit(), SVM dan Random Forest mengetahui pola dari data pelatihan dan membangun fungsi prediktif. Pelatihan ini menjadi tahap penting yang mendahului evaluasi dan perbandingan kinerja kedua algoritma pembelajaran mesin tersebut dalam mendeteksi kanker payudara, berdasarkan fitur sel yang diekstraksi dari pemeriksaan FNA.

3.7. Evaluasi Model

3.7.1 Support Vector Machine (SVM) & Random Forest



Gambar 5. Confusion Matrix SVM & RF

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.503**
Page 159-172

Gambar 5. Confusion matrix SVM (kiri) dan Random Forest (kanan) pada 114 data uji. Baris menunjukkan kelas aktual, kolom menunjukkan kelas prediksi, dan angka di dalam sel adalah jumlah sampel. Kedua model menghasilkan TN = 72, FP = 0, FN = 3, dan TP = 39. Gambar 5 memperlihatkan bahwa SVM dan Random Forest menghasilkan matriks konfusi yang sama pada 114 data uji. Dengan kelas ganas sebagai kelas positif, seluruh 72 sampel jinak terklasifikasi benar (TN = 72, FP = 0), sedangkan dari 42 sampel ganas, 39 terdeteksi (TP = 39) dan 3 terlewat (FN = 3). Artinya tidak ada pasien dengan tumor jinak yang salah didiagnosis ganas, tetapi masing-masing model melewatkan tiga kasus ganas. Karena kedua matriks identik, semua metrik yang dihitung dari label prediksi (akurasi, presisi, sensitivitas, spesifisitas, dan F1-score) bernilai sama untuk kedua model, sebagaimana perhitungan di bawah dan Tabel 3. Perbedaan antarmodel baru muncul pada AUC, yang dihitung dari probabilitas prediksi dan bukan dari label akhir.

Dalam konteks penaksiran kanker payudara secara klinis, tingkat sensitivitas (kapasitas untuk menemukan kasus kanker) sangat penting, karena kesalahan negatif palsu (pasien dengan kanker yang tidak terdeteksi) dapat menimbulkan akibat yang jauh lebih sungguh-sungguh dibanding dengan hasil positif palsu. Oleh sebab itu, model dengan sensitivitas tinggi lebih diprioritaskan dalam konteks diagnostik. Pada penelitian ini sensitivitas kedua model sebesar 92,86%, sehingga tiga dari 42 kasus ganas tidak terdeteksi dan hal ini perlu dicatat sebagai keterbatasan.

1. **Support Vector Machine dan Random Forest**

a) Akurasi $= \frac{(TP + TN)}{(TP + TN + FP + FN)}$
$= \frac{(39+ 72)}{(39 +72 + 0 + 3)}$
$= \frac{111}{114} = 0.9737$

b) Sensitivitas (Recall) $= \frac{TP}{(TP + FN)}$
$= \frac{39}{(39+3)}$
$= 0.9286$

c) Spesifitas $= \frac{TN}{TN + FP}$
$= \frac{72}{72 + 0}$
$= 1.0000$

d) Presisi $= \frac{TP}{TP + FP}$
$= \frac{39}{39 + 0}$
$= 1.0000$

e) F1-score $= 2 \text{ x } (\frac{(presisi\ x\ recall)}{(presisi+recall)})$
$= 2 \text{ x } \frac{(1.0000\ x\ 0.9286)}{(1.0000+0.9286)}$
$= 0.9630$

f) AUC dibaca dari SVM = 1,00 dan Random Forest = 0,99 (pembulatan dua desimal).

3.8. Tabel Hasil Evaluasi Model

Tabel hasil evaluasi model menyajikan perbandingan nilai dari berbagai metrik evaluasi yang digunakan untuk mengukur kinerja model dalam menghasilkan prediksi secara akurat dan konsisten.

Tabel 3. Hasil Evaluasi SVM dan Random Forest pada 114 Data Uji

Model	Akurasi	Presisi	Sensitivitas (Recall)	Spesifisitas	F1-score	AUC
SVM	0,9737	1,0000	0,9286	1,0000	0,9630	1,00*
Random Forest	0,9737	1,0000	0,9286	1,0000	0,9630	0,99*

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

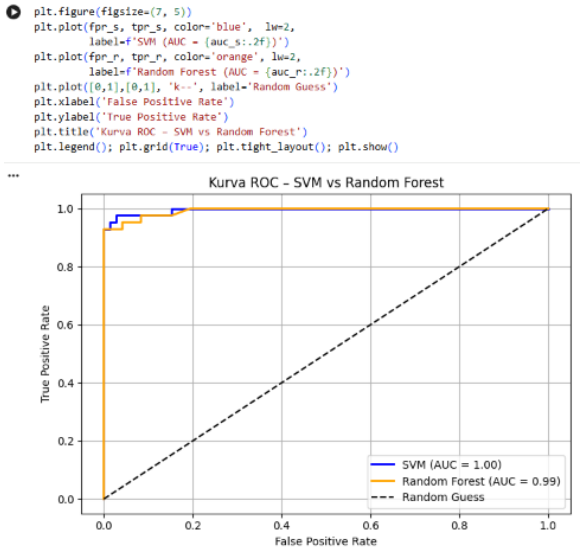
Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.503**
Page 159-172

*AUC dibulatkan dua desimal sesuai legenda Gambar 6.

Tabel 3 menunjukkan bahwa SVM dan Random Forest memiliki nilai yang sama pada akurasi (97,37%), presisi (100%), sensitivitas (92,86%), spesifisitas (100%), dan F1-score (96,30%). Presisi dan spesifisitas sempurna karena tidak ada sampel jinak yang diprediksi ganas, sedangkan F1-score sebesar 96,30% lebih rendah daripada presisi karena tiga kasus ganas terlewat. Selisih satu sampel pada 114 data uji setara dengan 0,88 poin persentase akurasi, sehingga perbedaan kecil antarmodel mudah berubah bila data uji berbeda. Satu-satunya perbedaan terdapat pada AUC, yang dibahas pada bagian berikut.

3.9. AUC (Area Under Curve)

AUC (Area Under the Curve) digunakan untuk mengukur kemampuan model dalam membedakan kelas positif dan negatif berdasarkan luas area di bawah kurva ROC, sehingga dapat menggambarkan tingkat kinerja dan akurasi klasifikasi model.



Gambar 6. AUC (Area Under Curve)

Gambar diatas menampilkan kurva ROC SVM (biru) dan Random Forest (oranye) pada data uji. Sumbu-x adalah false positive rate, sumbu-y adalah true positive rate, dan garis putus-putus menunjukkan klasifikasi acak. AUC SVM = 1,00 dan AUC Random Forest = 0,99 (pembulatan dua desimal).

Kedua kurva pada Gambar 6 berada jauh di atas garis klasifikasi acak dan hampir berimpit, sehingga secara visual perbedaannya sulit terlihat. Perbedaan AUC muncul di sisi kiri atas kurva. Pada false positive rate yang sangat rendah, kurva SVM naik lebih cepat dan mencapai true positive rate mendekati 1 lebih awal, sedangkan kurva Random Forest berbentuk tangga dan baru menyusul kemudian. Bentuk tangga ini wajar karena probabilitas Random Forest merupakan proporsi suara 100 pohon, sehingga hanya bernilai kelipatan 0,01 dan banyak sampel memperoleh skor yang sama [7]. Probabilitas SVM dengan probability=True dikalibrasi dari jarak ke batas keputusan dan bernilai lebih kontinu [10]. Dengan demikian, AUC SVM yang sedikit lebih tinggi (1,00 dibanding 0,99) menandakan urutan skor SVM sedikit lebih rapi dalam memisahkan kasus ganas dari jinak. Selisih ini kecil dan tidak mengubah keputusan klasifikasi pada ambang 0,5, yang terlihat dari matriks konfusi yang identik. Perlu dicatat bahwa AUC 1,00 hasil pembulatan tidak

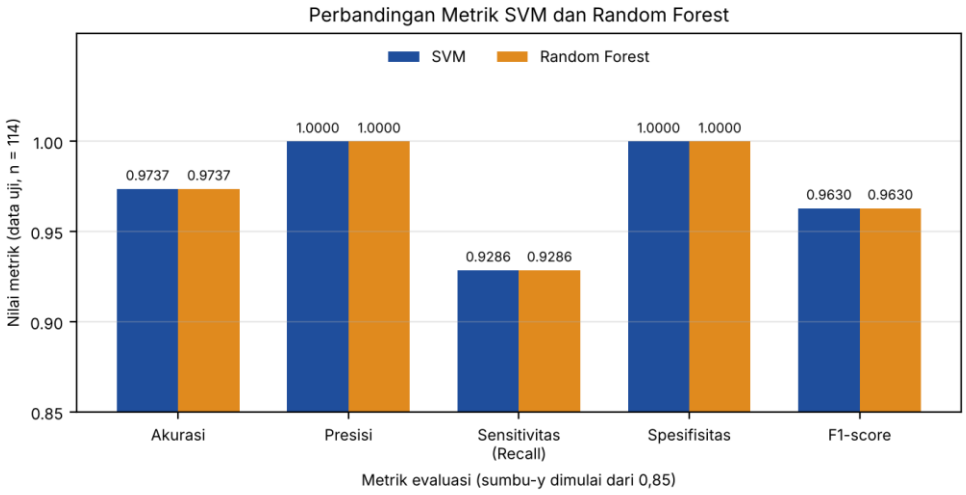
Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.503**
Page 159-172

berarti pemisahan yang sempurna, karena tiga kasus ganas tetap terlewat pada ambang tersebut.

3.10. Perbandingan Model SVM dan Random Forest

Perbandingan lima metrik evaluasi SVM dan Random Forest pada 114 data uji. Sumbu-x menunjukkan metrik evaluasi, sumbu-y menunjukkan nilai metrik (dimulai dari 0,85), dan angka di atas batang adalah nilai masing-masing metrik.



Gambar 7. Perbandingan Model SVM dan Random Forest

Gambar 7 menegaskan bahwa pada kelima metrik berbasis label, batang SVM dan Random Forest setara. Dengan demikian, hasil penelitian ini tidak mendukung pernyataan bahwa SVM lebih unggul dalam klasifikasi akhir. Bukti keunggulan SVM hanya berupa AUC yang sedikit lebih tinggi. Beberapa faktor dapat menjelaskan mengapa kedua model sangat dekat, sementara skor probabilitasnya berbeda.

Ada tiga hal yang bisa menjelaskan mengapa SVM dan Random Forest menghasilkan kinerja yang sama. Hal pertama ada pada datanya sendiri. Berdasarkan Tabel 1, sel ganas rata-rata memiliki radius, perimeter, area, konkavitas, dan titik cekung yang jauh lebih besar dibandingkan sel jinak. Karena kedua kelas relatif mudah dipisahkan, algoritma dengan cara kerja berbeda pun cenderung sampai pada keputusan yang sama untuk sebagian besar sampel. Tiga kasus ganas yang gagal dikenali kedua model diduga berada di wilayah tumpang tindih antarkelas, dugaan yang sejalan dengan besarnya simpangan baku pada kelas ganas, meskipun belum diperiksa pada tingkat sampel. Hal kedua menyangkut cara kerja algoritma. SVM berkernel RBF menghitung jarak antarsampel dari seluruh 30 fitur yang telah dinormalisasi, dengan batas keputusan yang dibentuk oleh sampel di sekitar margin [6]. Akibatnya, radius, perimeter, dan area yang berkorelasi di atas 0,98 sama-sama menguatkan informasi ukuran. Pada Random Forest, fitur yang saling berkorelasi memang dapat saling menggantikan di setiap percabangan, tetapi suara mayoritas dari 100 pohon tetap mengarah ke kelas yang sama [7]. Adapun perbedaan pada AUC muncul karena skor yang dihasilkan tidak sama, yaitu probabilitas kontinu pada SVM dan proporsi suara pohon pada Random Forest. Hal ketiga adalah keterbatasan ukuran data dan konfigurasi. Model dilatih dengan 455 sampel dan diuji dengan 114 sampel, menggunakan parameter bawaan tanpa penyetelan. Dengan data uji sekecil itu, selisih satu sampel saja mengubah akurasi sebesar 0,88 poin persentase, sementara validasi silang dan uji signifikansi tidak dilakukan. Karena itu, selisih AUC antara 1,00 dan 0,99 belum bisa dianggap sebagai perbedaan yang nyata antara kedua algoritma.

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.503**
Page 159-172

4. KESIMPULAN

Penelitian ini membandingkan Support Vector Machine dan Random Forest pada dataset Breast Cancer Wisconsin (Diagnostic) yang terdiri atas 569 sampel dan 30 fitur numerik. Pada 114 data uji, kedua algoritma menghasilkan matriks konfusi yang sama sehingga akurasi (97,37%), presisi (100%), sensitivitas (92,86%), spesifisitas (100%), dan F1-score (96,30%) identik. Perbedaan hanya tampak pada AUC, yaitu 1,00 untuk SVM dan 0,99 untuk Random Forest pada pembulatan dua desimal. Dengan demikian, SVM dan Random Forest sama-sama mampu memprediksi kanker payudara dengan baik pada dataset ini, dan keunggulan SVM terbatas pada kualitas urutan skor probabilitas, bukan pada keputusan klasifikasi akhir. Hasil ini memiliki beberapa keterbatasan, yaitu hanya satu kali pembagian data (80:20), parameter bawaan tanpa penyetelan, tidak ada validasi silang maupun uji signifikansi, serta hanya satu dataset. Kedua model juga masih melewatkan tiga kasus ganas, sehingga belum memadai untuk dipakai sebagai alat diagnosis tunggal. Karena itu, temuan ini hanya berlaku pada konfigurasi dan dataset yang digunakan dan tidak dapat digeneralisasi bahwa salah satu algoritma selalu lebih unggul. Penelitian selanjutnya disarankan menerapkan validasi silang dan penyetelan parameter, membandingkan algoritma lain seperti XGBoost, Gradient Boosting, atau deep learning, serta memakai data yang lebih banyak dan beragam, misalnya citra histopatologi, informasi genetik, atau hasil biopsi jaringan.

PERNYATAAN KONTRIBUSI PENULIS (CREDIT)

Nathasya Taruli Aprillia Purba: Penulisan draft awal, peninjauan dan penyuntingan. **Afninda Yovini Lumban Toruan**: Metodologi dan Konseptualisasi. **Elvis Sastra Ompusunggu**: Validasi dan Perangkat lunak.

PERNYATAAN KONFLIK KEPENTINGAN

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan finansial yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam makalah ini.

PERNYATAAN PENDANAAN

Penelitian ini tidak menerima hibah khusus dari lembaga pendanaan mana pun di sektor publik, komersial, atau nirlaba.

PERNYATAAN PENGGUNAAN AI GENERATIF DAN TEKNOLOGI BERBANTUAN AI DALAM PROSES PENULISAN

Selama penyusunan karya ini, penulis menggunakan claude.ai untuk membantu meyempurnakan kalimat yang akan dibuat ke dalam artikel. Setelah menggunakan alat ini, penulis meninjau dan mengedit konten sesuai kebutuhan dan bertanggung jawab penuh atas isi publikasi ini.

REFERENCES

[1] World Health Organization, "Breast Cancer," WHO, Geneva, Switzerland, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer. [Accessed: Jul. 29, 2026].

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.503**
Page 159-172

[2] J. Ferlay, M. Colombet, I. Soerjomataram, D. M. Parkin, M. Piñeros, A. Znaor, and F. Bray, "Cancer Statistics for the Year 2020: An Overview," International Journal of Cancer, vol. 149, no. 4, pp. 778–789, 2021.
[3] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast Cancer Wisconsin (Diagnostic) Data Set," UCI Machine Learning Repository, University of Wisconsin, Madison, WI, USA, 1995.
[4] D. Dua and C. Graff, "UCI Machine Learning Repository," School of Information and Computer Sciences, University of California Irvine, Irvine, CA, USA, 2019.
[5] Kaggle, "Breast Cancer Wisconsin (Diagnostic) Dataset," 2026. [Online]. Available: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.
[6] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[7] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[8] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2023.
[9] S. Raschka, V. Mirjalili, and M. Khan, Python Machine Learning, 4th ed. Birmingham, UK: Packt Publishing, 2023.
[10] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[11] A. Bilal, A. Imran, T. I. Baig, X. Liu, E. A. Nasr, and H. Long, "Breast Cancer Diagnosis Using Support Vector Machine Optimized by Improved Quantum Inspired Grey Wolf Optimization," Scientific Reports, vol. 14, no. 1, p. 10714, 2024.
[12] M. Alshamrani, "Diagnosis of Breast Cancer Using Random Forests," Procedia Computer Science, vol. 218, pp. 429–437, 2023.
[13] R. Oktafiani and E. I. Sela, "Breast Cancer Classification with Principal Component Analysis and SMOTE Using Random Forest Method and Support Vector Machine," International Journal of Computer Applications, vol. 186, no. 16, pp. 1–8, 2024.
[14] M. M. Mostafa, M. A. H. Al-Masri, and M. A. Al-Sharif, "The Stratified K-Folds Cross-Validation and Class-Balancing Methods with High-Performance Ensemble Classifiers for Breast Cancer Classification," Healthcare Analytics, vol. 4, p. 100247, 2023.
[15] M. Ebrahim, A. A. H. Sedky, and S. Mesbah, "Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer," Data, vol. 8, no. 2, p. 35, 2023.
[16] Y. Liu, H. Yin, and Y. Zhang, "Comparative Study of Machine Learning Algorithms for Medical Image Classification," Journal of Healthcare Engineering, vol. 2023, pp. 1–10, 2023.
[17] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?," Journal of Machine Learning Research, vol. 15, pp. 3133–3181, 2014.
[18] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
[19] N. Cruz-Roa et al., "Automatic Detection of Invasive Ductal Carcinoma in Whole Slide Images with Convolutional Neural Networks," Proceedings of SPIE Medical Imaging, vol. 9041, pp. 1–8, 2020.
[20] M. A. Khan, M. Sharif, M. Raza, and T. Saba, "Machine Learning Framework for Breast Cancer Classification," Microscopy Research and Technique, vol. 84, no. 12, pp. 2917–2928, 2021.

Nathasya Taruli Aprillia Purba: *Penulis Korespondensi
Copyright © 2027, Nathasya Taruli Aprillia Purba, Afninda Yovini Lumban Toruan, Elvis Sastra Ompusunggu.