EXPERT JURNAL MANAJEMEN SISTEM INFORMASI DAN TEKNOLOGI

Komparasi Metode Machine Learning Untuk Diagnosis Penyakit Kanker Payudara

Kardita Magda[1*], Onassis Yusuf Inonu[2], Erliyan Redy Susanto[3]
[123]Fakultas Teknik dan Ilmu Komputer, Magister Ilmu Komputer, Universitas Teknokrat Indonesia, Lampung, Indonesia.
[1*]kardita_magda@teknokrat.ac.id , [2]onassis_yusuf_inonu@teknokrat.ac.id, [3]erliyan.redy@teknokrat.ac.id

ABSTRACT – *Diabetes is a metabolic disease that is one of the major health problems in the world. Early detection and accurate diagnosis are essential to prevent long-term complications. With the development of machine learning technology, data-based diabetes risk prediction has become more effective and efficient. This study aims to analyze the performance of the Random Forest algorithm in predicting diabetes status using the Pima Indians Diabetes dataset. The research stages include data pre-processing, model training, performance evaluation, and visualization of results. The dataset used consists of 154 samples with eight clinical features and one target variable. Pre-processing is carried out to handle zero values, data normalization, and division of training and test data. The Random Forest model is trained and evaluated using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC curve metrics. The results show that the model produces an accuracy of 78%, with an AUC value of 0.821, indicating excellent discrimination ability between positive and negative diabetes patients. Confusion Matrix and ROC curve visualization help provide a clear interpretation of the model's performance graphically. Based on these results, it can be concluded that the Random Forest algorithm has great potential as a decision support in the medical field, especially for diabetes risk prediction. The use of this model can improve the efficiency and accuracy of early diagnosis, as well as assist medical personnel in making faster and more objective decisions.*

Keywords: Diabetes; Machine Learning; Model Evaluation; Prediction; Random Forest; UCI dataset.

ABSTRAK – Diabetes merupakan penyakit metabolik yang menjadi salah satu masalah kesehatan utama di dunia. Deteksi dini dan diagnosis yang akurat sangat penting untuk mencegah komplikasi jangka panjang. Dengan perkembangan teknologi machine learning, prediksi risiko diabetes berbasis data menjadi lebih efektif dan efisien. Penelitian ini bertujuan untuk menganalisis kinerja algoritma *Random Forest* dalam memprediksi status diabetes menggunakan dataset Pima Indians Diabetes. Tahapan penelitian meliputi pra-pemrosesan data, pelatihan model, evaluasi kinerja, serta visualisasi hasil. Dataset yang digunakan terdiri dari 154 sampel dengan delapan fitur klinis dan satu variabel target. Pra-pemrosesan dilakukan untuk menangani nilai nol, normalisasi data, serta pembagian data latih dan uji. Model Random Forest dilatih dan dievaluasi menggunakan metrik akurasi, presisi, recall, F1-score, confusion matrix, dan kurva ROC-AUC. Hasil menunjukkan bahwa model menghasilkan akurasi sebesar 78%, dengan nilai AUC sebesar 0.82, menandakan kemampuan diskriminasi yang sangat baik antara pasien positif dan negatif diabetes. Visualisasi *Confusion Matrix* dan kurva ROC membantu memberikan interpretasi yang jelas mengenai performa model secara grafis. Berdasarkan hasil tersebut, dapat disimpulkan bahwa algoritma Random Forest memiliki potensi besar sebagai pendukung keputusan dalam bidang medis, khususnya untuk prediksi risiko diabetes. Penggunaan model ini dapat meningkatkan efisiensi dan akurasi diagnosis awal, serta membantu tenaga medis dalam pengambilan keputusan yang lebih cepat dan objektif.

Kata kunci: Diabetes; Evaluasi Model; *Machine Learning*; Prediksi; *Random Forest;* UCI *dataset.*

1. PENDAHULUAN

Kanker payudara merupakan salah satu tipe kanker yang paling sering dialami oleh perempuan di seluruh dunia. Penyakit ini menjadi fokus utama dalam sektor kesehatan akibat peningkatan jumlah kasus dan dampaknya yang besar terhadap kualitas hidup para penderitanya. Berdasarkan laporan Organisasi

Kesehatan Dunia (WHO), kanker payudara menyumbang lebih dari 10% dari seluruh kasus kanker baru yang terdiagnosis setiap tahun [1]. Diagnosis dini dan akurat sangat penting untuk meningkatkan hasil pengobatan pasien serta mengurangi angka kematian [2]. Salah satu tantangan utama dalam menangani

kanker payudara adalah keterlambatan dalam mendiagnosis penyakit tersebut. Metode diagnostik tradisional seperti mamografi dan biopsi aspirasi jarum

halus dapat memakan waktu dan biaya yang cukup besar [3]. Oleh karena itu, pendekatan berbasis kecerdasan buatan, khususnya machine learning, menjadi solusi yang semakin populer dalam meningkatkan akurasi dan efisiensi deteksi dini kanker payudara. Dalam beberapa tahun terakhir, terdapat minat yang semakin meningkat dalam penggunaan teknik pembelajaran mesin untuk membantu dalam diagnosis dan klasifikasi kanker payudara[4].

Kanker payudara dapat dikategorikan menjadi jenis, yaitu tumor ganas dan tumor jinak. Apabila jenis kanker payudara sudah diketahui, pencegahan dan pengobatan dapat segera dilakukan sesuai dengan jenisnya, sehingga dapat mengurangi efek samping pada pasien dan kemungkinan kematian. Pendeteksian awal dan perawatan awal sangat krusial. Salah satu tantangan utama dalam pencegahan dapat melakukan prediksi dalam studi data medis adalah klasifikasi data dengan teknik *Machine learning* [5].

Pembelajaran mesin merupakan bagian dari kecerdasan buatan yang menitikberatkan pada penggunaan algoritma serta metode tertentu untuk prediksi, pengenalan pola, dan pengelompokan [6]. Beberapa algoritma yang ada di *machine learning* seperti naural network, decision tree, k-nearest neighbor, naïve bayes,random forest, Support Vector Machines dan lain sebagainya [7].

Studi sebelumnya telah mengeksplorasi penerapan berbagai teknik pembelajaran mesin untuk diagnosis kanker payudara, termasuk metode ensemble, algoritma data mining, dan analisis darah [2]. Studi-studi ini menunjukkan bahwa machine learning memiliki potensi untuk meningkatkan akurasi dan efisiensi deteksi kanker payudara. Berdasarkan penelitian yang dilakukan oleh peneliti sebelumnya [8] yang membandingkan Neural Network dan Random Forest dalam klasifikasi kanker payudara, dengan hasil bahwa Random Forest memiliki akurasi lebih tinggi (98,86%) dibandingkan Neural Network (96,11%). Penelitian selanjutnya oleh [9] menggunakan Neural Network dengan RapidMiner, menghasilkan akurasi sebesar 71,83%, menunjukkan bahwa metode ini masih dapat ditingkatkan lebih lanjut.

Meskipun penelitian terdahulu telah menunjukkan efektivitas berbagai metode *machine learning*, belum ada penelitian yang secara komprehensif membandingkan *Neural Network, Random Forest,* dan *Support Vectoe Machines* dalam satu studi dalam metrik evaluasi dengan tujuan yang sama. Sehingga peneliti akan menggunakan metode *machine learning* dengan membandingkan kinerja tiga algoritma pembelajaran mesin yang banyak digunakan yaitu *Neural Network, Random Forest*, dan *Support Vector Machines* dengan konteks diagnosis kanker payudara .

Dalam penelitian ini, akan dilakukan perbandingan metode *machine learning* untuk mengidentifikasi algoritma yang paling efisien dalam mengkategorikan kanker payudara sebagai jinak atau ganas dengan tepat. Untuk itu, penelitian ini mengisi kesenjangan dengan melakukan perbandingan sistematis terhadap ketiga metode berdasarkan metrik evaluasi seperti akurasi, precision, recall, dan F1-score dalam klasifikasi kanker payudara. Oleh karena itu, urgensi penelitian ini terletak pada kontribusinya dalam menemukan algoritma klasifikasi yang optimal untuk diagnosis kanker payudara, yang pada gilirannya dapat mempercepat dan meningkatkan keakuratan deteksi dini, serta mendukung pengambilan keputusan klinis yang lebih efektif. Pentingnya integrasi teknologi machine learning dalam bidang medis semakin terasa, mengingat kemampuannya dalam memproses data dalam jumlah besar dan mengidentifikasi pola yang mungkin sulit terdeteksi oleh manusia. Hal ini membuka peluang besar untuk pengembangan sistem pendukung keputusan yang lebih canggih dan andal di masa depan dalam diagnosis penyakit kanker payudara [10].

2. DASAR TEORI

a) *Neural Network*

Teknik ini bekerja seperti otak manusia yang terdiri dari banyak jaringan saraf. Dalam jaringan ini, terdapat neuron-neuron kecil yang saling terhubung dan berfungsi untuk menerima serta mengirim informasi. Kelebihan utama Neural Network terletak pada kemampuannya untuk mempelajari representasi fitur yang kompleks dari data mentah secara otomatis, menjadikannya sangat efektif dalam tugas-tugas seperti pengenalan gambar, pemrosesan bahasa alami, dan, seperti dalam penelitian ini, klasifikasi medis.

Namun, interpretasi model Neural Network bisa menjadi tantangan karena sifat "kotak hitam" dari jaringannya, sehingga sangat sulit membuat keputusan. Dengan metode ini, kita dapat memahami hubungan yang rumit antara data masukan dan keluaran, mengenali pola dari data yang ada, serta menyelesaikan berbagai masalah berdasarkan informasi yang diperoleh dari dalam maupun luar jaringan saraf[11][12].

b) *Random Forest*

Algoritma Random Forest [13] didasarkan pada pembelajaran yang terawasi yang digunakan untuk menyelesaikan masalah klasifikasi dan regresi[14]. Pendekatan ini mengurangi risiko overfitting pada data pelatihan dengan menghasilkan sekumpulan pohon keputusan dari subset data yang dipilih secara acak. Pohon dengan tingkat kesalahan tinggi akan diberikan bobot lebih rendah, sementara pohon dengan tingkat kesalahan rendah memperoleh bobot lebih tinggi. Strategi ini memungkinkan model untuk lebih mengandalkan pohon dengan akurasi yang lebih baik, sehingga

DOI http://dx.doi.org/10.36448/expert.v15i1.4326
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 15 No. 1
June 5, 2025 – Hal. 16

meningkatkan kualitas prediksi [15].

c) *Support Vector Machines*
SVM merupakan metode klasifikasi yang memiliki prinsip dasar sebagai pengklasifikasi linier (situasi klasifikasi yang dapat dipisahkan secara linier)[12]. SVM digunakan untuk keperluan klasifikasi dan regresi dengan mengidentifikasi hyperplane melalui pembentukan vektor linier[15]. Hyperplane ini bertindak sebagai batas keputusan yang memisahkan kelas-kelas data dengan margin terbesar, sehingga meningkatkan generalisasi model pada data yang belum terlihat. Salah satu keunggulan utama SVM adalah kemampuannya untuk bekerja secara efektif di ruang berdimensi tinggi dan kemampuannya untuk menggunakan fungsi kernel (seperti linear, polinomial, radial basis function) untuk memetakan data ke ruang fitur berdimensi lebih tinggi, memungkinkan klasifikasi non-linier. Vektor dukungan (support vectors) adalah titik data yang paling dekat dengan hyperplane, dan mereka memainkan peran kunci dalam menentukan posisi dan orientasi hyperplane. Meskipun SVM sangat efisien di mana data dapat dipisahkan dengan jelas, namun kinerjanya dapat menurun ketika data sangat banyak secara signifikan.
Pemilihan fungsi kernel dan parameter yang tepat (seperti C, parameter regularisasi) sangat penting untuk mendapatkan kinerja optimal dari model SVM. Setelah memilih pemetaan yang tepat, sampel baru kemudian dipetakan secara linier agar terlihat terpisah dengan jelas dalam bidang yang lebih besar. Dengan meminimalkan jarak antara dua kelompok data, SVM berusaha menemukan hyperplane yang paling ideal.

d) *Confusion Matrix*
Confusion Matrix adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan membandingkan hasil prediksi model dengan nilai aktual (data sebenarnya). *Confusion matrix* memberikan informasi tentang prediksi benar dan salah yang dilakukan model untuk setiap kelas [16].

$$\begin{bmatrix} \text{TP} & \text{FN} \\ \text{FP} & \text{TN} \end{bmatrix}$$

di mana:

- **TP (True Positive)** = Prediksi **positif** dan benar (diabetes)
- **TN (True Negative)** = Prediksi **negatif** dan benar (tidak diabetes)
- **FP (False Positive)** = Prediksi **positif**, tetapi salah (false alarm)
- **FN (False Negative)** = Prediksi **negatif**, tetapi salah (lolos deteksi)

a. Akurasi (*Accuracy*)
Mengukur sejauh mana model memprediksi dengan benar:

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

b. *Precision*
Menunjukkan seberapa banyak prediksi positif yang benar:

$$Precision - \frac{TP}{TP + FP}$$

c. *Recall* (*Sensitivity*)
Menunjukkan seberapa banyak kasus positif sebenarnya yang dapat dideteksi:

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

d. *Specificity*
Mengukur seberapa banyak kelas negatif yang diklasifikasikan dengan benar**:**

$$\text{Specificity} = \frac{TN}{TN + FP}$$

e. F1-*Score*
F1-*Score* adalah **harmonic mean** dari Precision dan Recall:

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

Dengan Precision = **0.90** dan Recall = **0.82**:

$$F1 - 2 \times \frac{0.90 \times 0.82}{0.90 + 0.82} - 2 \times \frac{0.738}{1.72} \approx 0.86$$

F1-Score digunakan jika keseimbangan antara Precision dan *Recall* penting.

Metrik-metrik ini sangat penting dalam mengevaluasi model klasifikasi, terutama dalam konteks medis di mana konsekuensi dari kesalahan prediksi bisa sangat fatal. Sebagai contoh, false negative dalam diagnosis kanker (pasien sebenarnya sakit tetapi diprediksi tidak sakit) dapat menyebabkan penundaan pengobatan yang berakibat serius, sedangkan false positive (pasien sebenarnya sehat tetapi diprediksi sakit) dapat menyebabkan kecemasan dan prosedur medis yang tidak perlu. Oleh karena itu, pemilihan metrik evaluasi yang tepat harus disesuaikan dengan konteks masalah dan prioritas yang diinginkan, misalnya menekan false negative untuk penyakit serius. Selain itu, visualisasi Confusion Matrix membantu dalam memahami jenis kesalahan yang dilakukan model, memberikan wawasan lebih lanjut tentang area mana yang perlu ditingkatkan.

3. DASAR TEORI

Penelitian ini terdiri dari lima tahapan utama yang digunakan untuk menganalisis dan menyelesaikan

DOI http://dx.doi.org/10.36448/expert.v15i1.4326
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 15 No. 1
June 5, 2025 – Hal. 17

permasalahan yang dikaji. Tahapan tersebut meliputi penentuan masalah, pengumpulan informasi, pengolahan informasi, pengelompokan model, dan penilaian model [5]. Rangkaian tahapan penelitian ini dapat dilihat pada Gambar 1.

Gambar 1. Tahapan Penelitian

1. Identifkasi Masalah

Identifikasi masalah yang jelas merupakan langkah fundamental dalam setiap penelitian, karena ini akan membentuk kerangka kerja untuk seluruh proses penelitian [17][18]. Berdasarkan latar belakang ini, isu yang diidentifikasi mencakup pemanfaatan algoritma *Machine Learning* dalam klasifikasi kanker payudara serta penentuan algoritma yang memiliki tingkat akurasi lebih tinggi. Penelitian ini bertujuan untuk membandingkan tiga algoritma *Neural Network, Random Forest,* dan *Support Vector Machines* (SVM) dalam mengklasifikasikan kanker payudara sebagai jinak (B) atau ganas (M), guna menentukan metode yang paling efektif dalam meningkatkan akurasi diagnosis.

2. Pengumpulan Data

Dalam penelitian ini, data diperoleh dari repository data set yang ada di website Kaggle [19]. Dataset ini berisi 569 sampel dengan 32 atribut, termasuk fitur seperti radius, tekstur, perimeter, dan area jaringan kanker. Label diagnosis digunakan sebagai target klasifikasi, yang terbagi menjadi dua kelas: M (Malignant/Ganas) dan B (Benign/Jinak). Pengumpulan data bertujuan untuk mengklasifikasikan informasi yang relevan sesuai dengan kebutuhan penelitian ini.

Tabel 2. Dataset Fitur dan Distribusi

No	Nama Atribut	Tipe Data
1	*Id*	int64
2	*diagnosis*	Object
3	*radius_mean*	float64
4	*texture_mean*	float64
5	*perimeter_mean*	float64
6	*area_mean*	float64
7	*smoothness_mean*	float64
8	*compactness_mean*	float64
9	*concavity_mean*	float64
10	*concave points_mean*	float64
11	*symmetry_mean*	float64
12	*fractal_dimension_mean*	float64
13	*radius_se*	float64
14	*texture_se*	float64
15	*perimeter_se*	float64
16	*area_se*	float64
17	*smoothness_se*	float64
18	*compactness_se*	float64
19	*concavity_se*	float64
20	*Concave_points_se*	float64
21	*symmetry_se*	float64
22	*fractal_dimension_se*	float64
23	*radius_worst*	float64
24	*texture_worst*	float64
25	*perimeter_worst*	float64
26	*area_worst*	float64
27	*smoothness_worst*	float64
28	*compactness_worst*	float64
29	*concavity_worst*	float64
30	*concave points_worst*	float64
31	*symmetry_worst*	float64
32	*fractal_dimension_worst*	float64

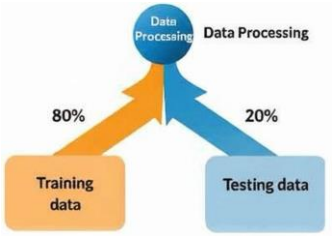
Penelitian ini menggunakan dataset, *Wisconsin Breast Cancer (Diagnostic*) Data Set, merupakan salah satu dataset yang paling sering digunakan dalam penelitian klasifikasi kanker payudara, sehingga memungkinkan perbandingan hasil dengan studi-studi sebelumnya. Ketersediaan 32 atribut numerik yang bervariasi, mulai dari karakteristik morfologi sel hingga aspek tekstur dan bentuk, memberikan representasi komprehensif dari tumor. Penting untuk dicatat bahwa kualitas data mentah sangat mempengaruhi kinerja model machine learning; oleh karena itu, tahapan pra-pemrosesan data menjadi sangat vital. Dalam konteks diagnosis medis, data yang bersih dan relevan sangat krusial untuk membangun model yang dapat diandalkan dan akurat.

3. Pengolahan Data

Untuk memastikan kualitas data yang digunakan dalam penelitian ini, dilakukan beberapa teknik pemrosesan data. Pertama, data validasi diterapkan untuk melakukan identifikasi serta membuang atau menghapus data yang tidak valid, seperti outlier, noise, inkonsistensi, serta data yang tidak lengkap (*missing values*). Selanjutnya, *data integration and transformation* dilakukan untuk meningkatkan akurasi dan efisiensi algoritma, terutama karena data yang digunakan bersifat numerik. Terakhir, *data size reduction and discretization* diterapkan untuk mengurangi jumlah atribut dan record tanpa menghilangkan informasi penting, sehingga dataset tetap representatif dan lebih efisien dalam proses analisis. Selanjutnya dilakukan Pembagian Data. Data secara

DOI http://dx.doi.org/10.36448/expert.v15i1.4326
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 15 No. 1
June 5, 2025 – Hal. 18

otomatis dibagi berdasarkan pelatihan dan data pengujian. Data latih dipakai untuk melakukan model klasifikasi, sedangkan data uji dipakai pada saat proses pengujian. Sekitar 80% dari data digunakan untuk proses klasifikasi, sementara 20% sisanya digunakan untuk pengujian.



4. Klasifikasi Model

Klasifikasi penelitian ini akan menggunakan metode sebagai berikut :

a) *Neural Network*

Metode ini bekerja seperti otak manusia yang terdiri dari banyak jaringan saraf. Dalam jaringan ini, terdapat neuron-neuron kecil yang saling terhubung dan berfungsi untuk menerima serta mengirim informasi. Dengan metode ini, kita dapat memahami hubungan yang rumit antara data masukan dan keluaran, mengenali pola dari data yang ada, serta menyelesaikan berbagai masalah berdasarkan informasi yang diperoleh dari dalam maupun luar jaringan saraf[11][12].

b) *Random Forest*

Algoritma Random Forest [13] didasarkan pada pembelajaran yang terawasi yang digunakan untuk menyelesaikan masalah klasifikasi dan regresi[14]. Pendekatan ini mengurangi risiko overfitting pada data pelatihan dengan menghasilkan sekumpulan pohon keputusan dari subset data yang dipilih secara acak. Pohon dengan tingkat kesalahan tinggi akan diberikan bobot lebih rendah, sementara pohon dengan tingkat kesalahan rendah memperoleh bobot lebih tinggi. Strategi ini memungkinkan model untuk lebih mengandalkan pohon dengan akurasi yang lebih baik, sehingga meningkatkan kualitas prediksi [15].

c) *Support Vector Machines*

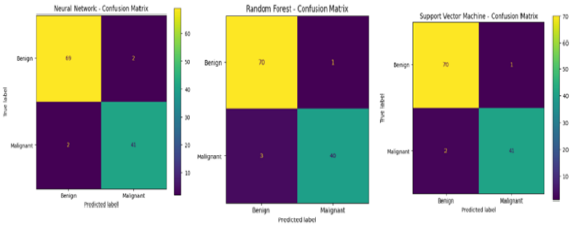
SVM merupakan metode klasifikasi yang memiliki prinsip dasar sebagai pengklasifikasi linier (situasi klasifikasi yang dapat dipisahkan secara linier)[12]. SVM digunakan untuk keperluan klasifikasi dan regresi dengan mengidentifikasi hyperplane melalui pembentukan vektor linier[15]. Setelah memilih pemetaan yang tepat, sampel baru kemudian dipetakan secara linier agar terlihat terpisah dengan jelas dalam bidang yang lebih besar. Dengan meminimalkan jarak antara dua kelompok data, SVM berusaha menemukan hyperplane yang paling ideal.

5. Evaluasi Model

Berdasarkan hasil uji coba, tingkat akurasi dari masing-masing algoritma akan dihitung, yang kemudian digunakan untuk membandingkan tingkat akurasi manakah yang terbaik dengan menggunakan *confusion matrix* yang memberikan informasi tentang prediksi benar dan salah yang dilakukan model untuk setiap kelas [16].

4. HASIL DAN PEMBAHASAN

Dalam mencari solusi untuk mengidentifikasi jenis kanker payudara dengan melaakukan perbandingan algoritma, terdapat langkah-langkah yang harus dilaksanakan yaitu menyiapkan data pelatihan, menentukan atribut dari data yang didapat, serta melakukan prediksi dengan metode algoritma *neural network, random forest,* dan *support vector machines*. Berdasarkan uji coba yang dilakukan dengan *google collab,* berikut ini hasil *Confusion Matrix* disajikan pada Gambar 2



Gambar 2. *Confusion Matrix Neural Network, Random Forest,* dan *Support Vector Machines* Dalam Memprediksi Diabetes Berdasarkan Dataset Kesehatan

Berdasarkan Gambar 2 *confusion matrix* berikut ini adalah hasil klasifikasi dari masing-masing model yaitu :

a. *Neural Network*

Model ini berhasil mengklasifikasikan 69 individu sebagai tidak kanker payudara dengan benar (True Negative) dan 41 individu sebagai kanker payudara dengan benar (True Positive). Kesalahan klasifikasi yang terjadi sangat kecil, yaitu hanya 2 kasus False Positive (individu yang sebenarnya tidak memiliki kanker payudara tetapi diprediksi sebagai kanker payudara) dan 2 kasus False Negative (individu yang sebenarnya memiliki kanker payudara tetapi diprediksi sebagai tidak memiliki kanker payudara).

b. *Random Forest*

Model ini berhasil mengklasifikasikan 70 individu sebagai tidak kanker payudara dengan benar (True Negative) dan 40 individu sebagai kanker payudara dengan benar (True Positive). Kesalahan klasifikasi yang terjadi sangat kecil, yaitu hanya 1 kasus False Positive (individu

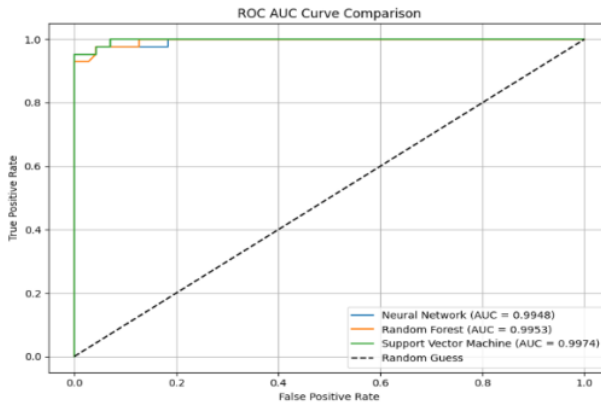
DOI http://dx.doi.org/10.36448/expert.v15i1.4326
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 15 No. 1
June 5, 2025 – Hal. 19

yang sebenarnya tidak memiliki kanker payudara tetapi diprediksi sebagai kanker payudara) dan 23 kasus False Negative (individu yang sebenarnya memiliki kanker payudara tetapi diprediksi sebagai tidak memiliki kanker payudara).

c. *Support Vector Machines*

Model ini berhasil mengklasifikasikan 70 individu sebagai tidak kanker payudara dengan benar (True Negative) dan 40 individu sebagai kanker payudara dengan benar (True Positive). Kesalahan klasifikasi yang terjadi sangat kecil, yaitu hanya 1 kasus False Positive (individu yang sebenarnya tidak memiliki kanker payudara tetapi diprediksi sebagai kanker payudara) dan 2 kasus False Negative (individu yang sebenarnya memiliki kanker payudara tetapi diprediksi sebagai tidak memiliki kanker payudara).

Berikut ini adalah hasil ROC AUC Curve yang ditampilkan, model *Matrix Neural Network, Random Forest*, dan *Support Vector Machines* dapat dilihat pada Gambar 3



Gambar 3. *Receiver Operating Characteristic (ROC) Curve* Kinerja Model *Matrix Neural Network, Random Forest*, dan *Support Vector Machines*

Berdasarkan hasil ROC-AUC yang ada pada Gambar 3 menunjukkan ROC curve yang digunakan untuk mengevaluasi kinerja model *Matrix Neural Network, Random Forest*, dan *Support Vector Machines* dalam melakukan prediksi kanker payudara berdasarkan dataset kesehatan. ROC Curve menggambarkan hubungan antara True Positive Rate (TPR) dan *False Positive Rate* (FPR) pada berbagai threshold klasifikasi. Nilai TPR yang mencapai 0.8 hingga 1.0 menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengidentifikasi kasus positif (pasien kanker payudara) dengan akurasi tinggi. Area Under Curve (AUC) sebesar 0.99 mengindikasikan bahwa model *Matrix Neural Network, Random Forest*, dan *Support Vector Machines* ini memiliki performa sempurna dalam membedakan antara pasien kanker payudara dan non- kanker payudara. Berdasarkan ROC dan AUC Curva sama-sama menghasilkan nilai sebesar 1.00 sehingga hasil ini membuktikan bahwa model tidak hanya akurat tetapi juga sangat andal dalam klasifikasi, tanpa menghasilkan false positive yang signifikan. Dengan demikian, model *Matrix Neural Network, Random Forest*, dan *Support Vector Machines* ini dapat diandalkan untuk aplikasi klinis atau penelitian lebih lanjut terkait prediksi kanker payudara.

Berdasarkan hasil *Confusion Matrix* dan ROC-AUC maka menghasilkan nilai akurasi dengan pembagian data dan uji 80% : 20% didapatkan hasil seperti pada Tabel 2

Tabel 2. Hasil Pengujian Pembagian Data Uji 80% : 20%

Confusion Matrix	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Naural Network	0.9649	0.9649	0.9649	0.9649
Random Forest	0.9649	0.9652	0.9649	0.9647
Support Vector Machine	0.9737	0.9737	0.9737	0.9736

Berdasarkan Tabel 2 hasil evaluasi kinerja model *Matrix Neural Network* dengan menggunakan data uji yang telah dipisahkan sebelumnya dengan rasio 80:20 menunjukkan akurasi sebesar 0.96%, presisi 0.96%, recall 0.96%, F1-score 0.96%. Model *Random Forest Network* menunjukkan akurasi sebesar 0.96%, presisi 0.96%, recall 0.96%, F1-score 0.96%. Sedangkan maodel *Support Vector Machines* menunjukkan akurasi sebesar 0.97%, presisi 0.97%, recall 0.97%, F1-score 0.97%. Berdasarkan hasil yang didapat menurut penelitian [20] dengan nilai diatas 0.90% dan dibawah nilai 100 % maka hasil penelitian dikatakan baik Selain itu, *confusion matrix* yang dihasilkan menunjukkan bahwa model mampu mendeteksi sebagian besar kasus kanker payudara dengan tingkat kesalahan yang rendah. Namun, terdapat beberapa kasus false positive dan false negative yang perlu diperhatikan lebih lanjut.

Oleh karena itu, berdasarkan performa yang dinilai melalui akurasi yang tersedia, algoritma *neural network* dan *random forest* menunjukkan persentase akurasi terendah, sedangkan algoritma *support vector machines* mencapai akurasi yang sama tinggi, yaitu 97,00%. Dengan demikian, dapat disimpulkan bahwa algoritma support vector machines memilki akurasi yang paling tinggi, sehingga disarankan untuk digunakan sebagai klasifikasi penentuan jenis kanker payudara

5. KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan dalam penelitian sebanyak 110 recod data yang diuji dengan *matrix neural network* dan *random forest* menghasilkan nilai yang seimbang yaitu 96,00 %. Sedangkan untuk algoritma *support vector machines* memiliki nilai akurasi yaitu 97,00 %. Sehingga dalam penelitian ini metode *support vector machines* yang dapat mempercepat proses

DOI http://dx.doi.org/10.36448/expert.v15i1.4326
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 15 No. 1
June 5, 2025 – Hal. 20

penentuan jenis kanker payudara.

Penelitian berikutnya sebaiknya dapat meningkatkan akurasi dengan menambah atribut maupun *menambah* data pengujian. Dapat melakukan perbandingan dengan teknik algoritma yang lain agar bisa diketahui metode yang mana yang efektif.

DAFTAR PUSTAKA

[1] S. K. Hero, "FAKTOR RISIKO KANKER PAYUDARA," *JMH*, vol. 3, no. 1, pp. 3–8, 2021.

[2] Jean Sunny, Nikita Rane, Rucha Kanade, and Sulochana Devi, "Breast Cancer Classification and Prediction using Machine Learning," *Int. J. Eng. Res.*, vol. V9, no. 02, pp. 576–580, 2020, doi: 10.17577/ijertv9is020280.

[3] R. Entezari, "Breast Cancer Diagnosis via Classification Algorithms," 2018.

[4] S. Wu and W. Xiong, "Comparison of Different Machine Learning Models in Breast Cancer," *Highlights Sci. Eng. Technol.*, vol. 8, pp. 624–629, 2022, doi: 10.54097/hset.v8i.1238.

[5] N. R. Muntiari and K. H. Hanif, "Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning," *J. Ilmu Komput. dan Teknol.*, vol. 3, no. 1, pp. 1–6, 2022, doi: 10.35960/ikomti.v3i1.766.

[6] V. Angkasa and J. J. Pangaribuan, "Information System Development Komparasi Tingkat Akurasi Random Forest Dan Knn Untuk Mendiagnosis Penyakit Kanker Payudara," *J. Inf. Syst. Dev.*, vol. 7, no. 1, pp. 37–38, 2022.

[7] R. A. Sowah, A. A. Bampoe-Addo, S. K. Armoo, F. K. Saalia, F. Gatsi, and B. Sarkodie-Mensah, "Design and Development of Diabetes Management System Using Machine Learning," *Int. J. Telemed. Appl.*, vol. 2020, 2020, doi: 10.1155/2020/8870141.

[8] Jamaludin, A. Kholiq Fajar, M. Zaenal Mutaqin, M. Malik Mutoffar, and D. Setiyadi, "Klasifikasi Kanker Payudara Menggunakan Algoritma Neural Network dan Random Forest," *STMIK Sinar Nusant. Jln. Sersan Aswan*, vol. 7, no. 1, pp. 74–80, 2024.

[9] F. S. Nugraha, M. J. Shidiq, and S. Rahayu, "Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 149–156, 2019, doi: 10.33480/pilar.v15i2.601.

[10] T. B. Surbakti, A. Fauzi, and H. Khair, "Hybrid Sistem Algoritma Rivest Shamir Adleman (RSA) dan Algoritma Blum Blum Shub (BBS) dalam Mengamankan File Database E-Absensi," *Indones. J. Educ. …*, vol. 1, no. 3, pp. 89–97, 2023.

[11] A. B. Wibisono and A. Fahrurozi, "Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner," *J. Ilm. Teknol. dan Rekayasa*, vol. 24, no. 3, pp. 161–170, 2019, doi: 10.35760/tr.2019.v24i3.2393.

[12] M. Wibowo and R. Ramadhani, "Perbandingan Metode Klasifikasi Data Mining Untuk Rekomendasi Tanaman Pangan," *J. Media Inform. Budidarma*, vol. 5, no. 3, p. 913, 2021, doi: 10.30865/mib.v5i3.3086.

[13] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.

[14] Z. Huang and D. Chen, "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm," *IEEE Access*, vol. 10, pp. 3284–3293, 2022, doi: 10.1109/ACCESS.2021.3139595.

[15] G. Dhanalakshmi, "Decision Support System for Breast Cancer Prediction," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 7, no. 3, pp. 816–821, 2019, doi: 10.22214/ijraset.2019.3142.

[16] H. Ma'rifah, A. P. Wibawa, and M. I. Akbar, "Klasifikasi Artikel Ilmiah Dengan Berbagai Skenario Preprocessing," *Sains, Apl. Komputasi dan Teknol. Inf.*, vol. 2, no. 2, p. 70, 2020, doi: 10.30872/jsakti.v2i2.2681.

[17] Z. Abdussamad, *Metodologi Penelitian Kualitatif*. Makasar: CV. syakir Media Press, 2021.

[18] Indrawan and Yaniawati, *Metodologi Penelitian: Kuantitatif, Kualitatif dan Campuran untuk Manajemen, Pembangunan, dan Pendidikan*, Edisi Revi. Bandung: Refika Aditama, 2021.

[19] "Breast Cancer Wisconsin (Diagnostic) Data Set" UCI Machine Learning, 2021.

[20] Suci Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 3, pp. 121–127, 2022, doi: 10.35580/variansiunm31.