 ISSN:2302-8149
e-ISSN:2540-9719

Komparasi Model pada *Machine Learning* untuk Prediksi Tingkat Kanker Paru

Comparison of Machine Learning Models for Predicting Lung Cancer Severity

[1]Ninik Lestari, [2]Erliyan Redy Susanto*
[1,2]Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia
[1,2]Jl. Z.A. Pagar Alam No. 9-11 Labuhan Ratu, Kota Bandar Lampung, Provinsi Lampung, Indonesia
*e-mail: erliyan.redy@teknokrat.ac.id

(***received*:** 30 April 2025, ***revised*:** 30 July 2025?, ***accepted*:** 31 July 2025)

Abstrak

Studi ini bertujuan untuk membandingkan kinerja empat algoritma machine learning, yaitu *Random Forest, Support Vector Machine* (SVM), *Logistic Regression, dan K-Nearest Neighbors* (KNN), dalam memprediksi tingkat keparahan kanker paru berdasarkan data medis pasien. Dataset yang digunakan mencakup informasi medis dari pasien dengan variabel target berupa tingkat keparahan kanker (rendah, sedang, tinggi). Eksperimen dilakukan dengan *train-test split* 80:20 tanpa *feature scaling*. Hasil menunjukkan RF mencapai akurasi 100 %, LR 99%, KNN 82%, dan SVM 43%. Keunggulan *Random Forest* berasal dari ensemble pohon keputusan yang menekan *overfitting* pada fitur *numerik* berdimensi menengah, sementara SVM (kernel = RBF, C = 1.0, gamma = '*scale'*) gagal menyesuaikan diri karena tidak adanya penskalaan dan tuning hiper-parameter. Recall, precision, dan F1-score mengonfirmasi dominasi RF dan LR. Studi ini memberikan wawasan tentang efektivitas algoritma machine learning dalam diagnosis kanker paru dan kontribusi penggunaan pendekatan multi-algoritma. Hasil studi menyarankan penggunaan RF sebagai model utama dan LR sebagai pengontrol dalam sistem pendukung diagnosis klinis, memungkinkan dokter menentukan terapi dini yang lebih personalisasi dan meningkatkan prognosis pasien kanker paru.
Kata Kunci: prediksi kanker paru, machine learning, random forest, diagnosa dini

Abstract

This study aims to compare the performance of four machine learning algorithms—Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and K-Nearest Neighbors (KNN)—in predicting lung cancer severity based on patient medical data. The dataset includes clinical information with the target variable categorized into three severity levels: low, medium, and high. Experiments were conducted using an 80:20 train-test split without feature scaling. The results show that RF achieved 100% accuracy, LR 99%, KNN 82%, and SVM 43%. The superior performance of Random Forest can be attributed to its ensemble of decision trees, which mitigates overfitting in medium-dimensional numerical features, whereas SVM (kernel = RBF, C = 1.0, gamma = "scale") failed to adapt due to the absence of scaling and hyperparameter tuning. Recall, precision, and F1-score further confirm the dominance of RF and LR. This study provides insights into the effectiveness of machine learning algorithms in lung cancer diagnosis and highlights the contribution of a multi-algorithm approach. The findings recommend using RF as the primary model and LR as a complementary control within clinical decision support systems, enabling physicians to make earlier, more personalized treatment decisions and ultimately improve lung cancer patient prognosis.
Keywords: *lung cancer prediction, machine learning, random forest, early diagnosis*

1 Pendahuluan

Kanker merupakan salah satu penyakit yang paling mematikan di dunia saat ini[1]. Menurut *World Health Organization* WHO (2018), penyebab utama kedua kematian di dunia dengan jumlah kematian 9,6 juta kematian adalah kanker[2]. Menurut perkiraan American Cancer Society untuk kanker paru di Amerika Serikat pada tahun 2022 adalah sekitar 236.740 kasus baru kanker paru dimana 117.910 pada pria dan 118.830 pada perempuan dan sekitar 130.180 kematian akibat kanker

 ISSN:2302-8149
e-ISSN:2540-9719

paru dimana 68.820 pada pria dan 61.360 pada perempuan. Beberapa faktor resiko yang dapat menjadi penyebab kanker paru pada orang yang tidak merokok diantaranya asbestos, radon dan polusi udara[3]. Diagnosis dini kanker sangat penting untuk meningkatkan tingkat kesembuhan dan mengurangi tingkat kematian. Namun, diagnosis kanker masih menghadapi tantangan karena kompleksitas penyakit ini dan variasi gejala yang dialami oleh pasien. Oleh karena itu, pengembangan metode diagnosis yang lebih akurat dan efisien menjadi sangat penting[4].

Dalam beberapa tahun terakhir, *machine learning* telah menunjukkan potensinya dalam membantu diagnosis kanker. Algoritma *machine learning* dapat menganalisis data medis pasien dan mengidentifikasi pola yang sulit terdeteksi oleh manusia[5]. Beberapa algoritma yang telah digunakan dalam penelitian sebelumnya meliputi *Random Forest*, *Support Vector Machine* (SVM), *Logistic Regression*, dan *K-Nearest Neighbors* (KNN). Namun, penelitian sebelumnya masih terbatas dalam menggabungkan beberapa algoritma dan mengevaluasi kinerjanya secara komprehensif pada dataset kanker yang besar dan beragam. Studi-studi lalu memakai algoritma seperti *Random Forest*, SVM, *Logistic Regression*, dan KNN untuk memprediksi kanker paru, tetapi umumnya hanya mengetes satu-dua model di data kecil (<500 pasien), menilai hanya akurasi, dan mengabaikan *recall*, *precision*, serta *F1-score* untuk tiga tingkat keparahan. Mereka pun jarang membahas pengaruh cara validasi atau penskalaan fitur terhadap hasil.

Tujuan studi ini adalah untuk mengembangkan dan mengevaluasi model machine learning untuk diagnosis kanker menggunakan empat algoritma yang berbeda, yaitu *Random Forest*, SVM, *Logistic Regression*, dan KNN[6]. Studi ini bertujuan untuk membandingkan kinerja algoritma-algoritma ini dalam mengklasifikasikan tingkat kanker (rendah, sedang, tinggi) berdasarkan 24 fitur numerik dari dataset 1.000 pasien. Membandingkan kinerja keempat algoritma berdasarkan empat metrik evaluasi, yaitu akurasi, recall, precision, dan F1-score. Serta mengidentifikasi algoritma terbaik untuk integrasi dalam sistem pendukung keputusan klinis guna membantu dokter menentukan terapi dini secara personalisasi. Hasil studi ini diharapkan dapat memberikan wawasan yang lebih dalam tentang algoritma mana yang paling efektif dalam diagnosis kanker.

Manfaat dari studi ini adalah pengembangan model machine learning yang dapat membantu dokter dalam mengambil keputusan diagnosis dini pada kanker paru yang lebih akurat. Model ini dapat digunakan sebagai alat bantu dalam menganalisis data medis pasien dan memberikan prediksi tingkat kanker[7]. Selain itu, studi ini juga memberikan kontribusi dalam membandingkan kinerja beberapa algoritma machine learning dalam konteks diagnosis kanker. Novelty dari studi ini terletak pada penggabungan dan evaluasi komprehensif dari empat algoritma machine learning yang berbeda pada dataset kanker yang besar dan beragam.

2 Tinjauan Literatur

Kanker paru merupakan salah satu penyakit kanker yang paling mematikan di dunia. Penelitian terbaru menunjukkan bahwa kanker paru merupakan penyebab utama kematian akibat kanker secara global[8]. Penelitian terbaru menunjukkan bahwa kanker paru merupakan penyebab utama kematian akibat kanker secara global. Dalam beberapa tahun terakhir, machine learning telah menunjukkan potensinya dalam membantu diagnosis kanker paru. Beberapa algoritma yang telah digunakan dalam penelitian sebelumnya dan studi ini bertujuan untuk mengembangkan dan mengevaluasi model *machine learning* untuk diagnosis kanker paru menggunakan empat algoritma yang berbeda, yaitu *Random Forest*, SVM, *Logistic Regression*, dan KNN[9].

Kanker paru merupakan penyakit yang sangat penting untuk diteliti, dan banyak penelitian telah dilakukan mengenai penggunaan machine learning dalam menangani masalah terkait kanker paru[10]. Penggunaan machine leaening telah membantu mengatasi berbagai tantangan yang terkait dengan diagnosis dan prediksi kanker paru, seperti yang disajikan dalam Tabel 1.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 1 Berbagai metode machine learning dan akurasi

Tahun	Metode	Hasil	Penulis
2024	Random Forest, Support Vector Machine (SVM), Logistic Regression, dan K-Nearest Neighbors (KNN)	Akurasi Random Forest 93%, SVM 87% hingga 91%, Logistic Regression 84% hingga 88%, KNN 82% hingga 86%	Hadrien T. Gayap, dan Moulay A. Akhloufi. [11]
2023	Random Forest	Akurasi 90,61%	Permana, Arifin Yusuf, et al. [9]
2023	Decion Tree	Akurasi 89%	Putra et al. [12]
2023	Support Vector Machine (SVM)	Akurasi 62,3%	Septhya et al. [13]
2022	SVM	Akurasi SVM 85%	A. A. Nagra, I. Mubarik, et al. [14]
2021	SVM, CNN, KNN	Akurasi SVM 95,5%, CNN 92%, dan KNN 88.4%	D. Mustafa Abdullah, et al. [15]
2021	Support Vector Machine (SVM)	Akurasi 56,69%	Hamid, T. M. T., Sallehuddin, R., Yunos, Z. M., & Ali, A. [16]

Tabel 1 merangkum berbagai metode machine learning yang digunakan dalam penelitian prediksi dan klasifikasi dari tahun 2021 hingga 2024. Beragam metode seperti Random Forest, SVM, Logistic Regression, KNN, dan Decision Tree telah diterapkan dalam berbagai konteks untuk mengatasi masalah prediksi atau klasifikasi. Pada tahun 2024, penelitian oleh Hadrien T. Gayap dan Moulay A. Akhloufi [11] menggunakan kombinasi Random Forest, SVM, Logistic Regression, dan KNN. Hasil evaluasi menunjukkan akurasi tinggi: Random Forest mencapai 93%, SVM 87–91%, Logistic Regression 84–88%, dan KNN 82–86%. Pendekatan multi-algoritma ini membuktikan efektivitasnya dalam menyelesaikan masalah yang kompleks. Pada tahun 2023, Permana, Arifin Yusuf, et al. [9] memfokuskan penelitian pada Random Forest dan mencapai akurasi sebesar 90,61%. Penelitian lain oleh Putra et al. [12] menggunakan Decision Tree dan mencatatkan akurasi 89%. Sementara itu, Septhya et al. [13] menggunakan SVM dengan hasil akurasi lebih rendah, yaitu 62,3%, menunjukkan bahwa kinerja SVM sangat bergantung pada konteks aplikasinya. Pada tahun 2022, A. A. Nagra, I. Mubarik, et al. [14] mengevaluasi kombinasi SVM sebesar 85%.

Pada tahun 2021, D. Mustafa Abdullah et al. [15] juga mengevaluasi kombinasi metode SVM, CNN, dan KNN. Hasilnya menunjukkan bahwa Akurasi SVM 95,5%, diikuti oleh CNN 92%, dan KNN 88.4%. Selain itu, penelitian oleh Hamid, T. M. T., Sallehuddin, R., Yunos, Z. M., & Ali, A. [11] pada tahun 2021 menggunakan SVM dan mencatatkan akurasi sebesar 56,69%. Hasil ini menunjukkan bahwa SVM mungkin tidak selalu optimal dalam semua kasus, tergantung pada dataset dan parameter yang digunakan[17]. Secara keseluruhan, tabel ini menunjukkan bahwa metode *Random Forest* sering kali memberikan hasil akurasi yang lebih tinggi dibandingkan metode lainnya, meskipun performa setiap metode dapat bervariasi tergantung pada dataset dan parameter yang digunakan[18]. Pendekatan kombinasi beberapa metode juga menunjukkan potensi dalam meningkatkan akurasi prediksi. Hasil-hasil ini memberikan wawasan penting tentang efektivitas berbagai algoritma machine learning dalam menyelesaikan masalah spesifik, serta membantu dalam pemilihan metode yang paling sesuai untuk aplikasi tertentu. Studi ini bertujuan untuk mengembangkan dan mengevaluasi model machine learning untuk diagnosis kanker paru menggunakan empat algoritma yang berbeda, yaitu *Random Forest*, SVM, *Logistic Regression*, dan KNN[19]. Hasil studi ini diharapkan dapat memberikan wawasan yang lebih dalam tentang algoritma mana yang paling efektif dalam diagnosis kanker paru.

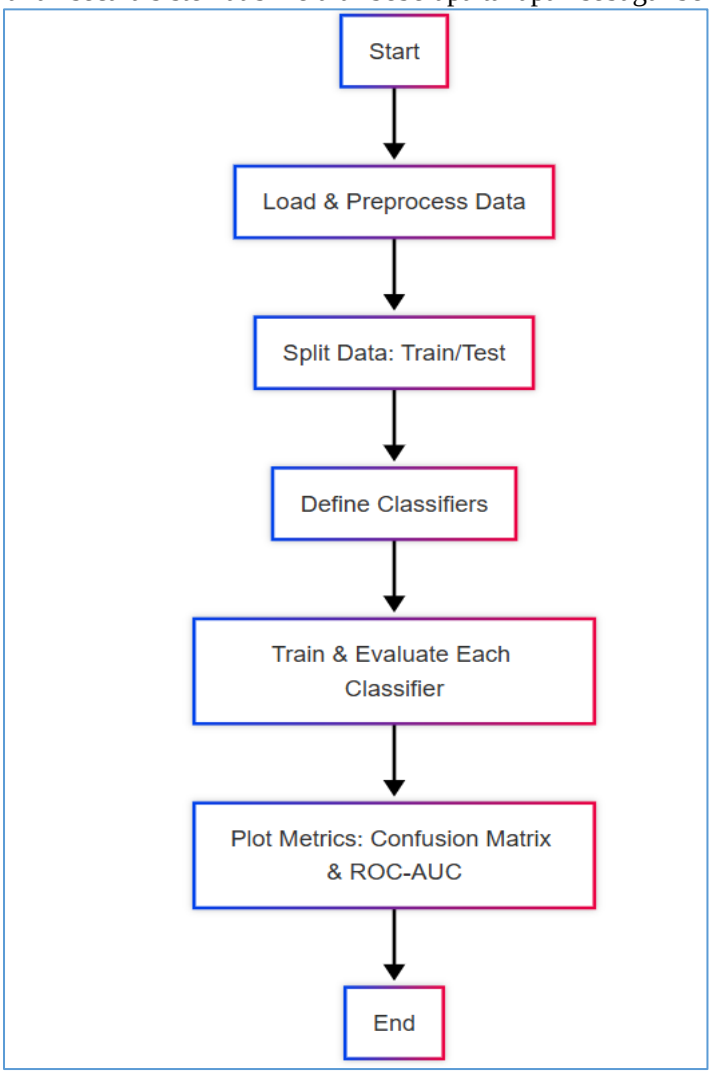
 ISSN:2302-8149
e-ISSN:2540-9719

3 Metode Penelitian

Studi ini menggunakan empat metode prediksi pembelajaran mesin dalam prediksi penyakit jantung, yaitu Random Forest, SVM, Logistic Regression, dan KNN. Penelitian dilakukan dengan tahapan yang sesuai dengan pendekatan masing-masing metode. Dataset yang digunakan dari kaggle "Cancer Patient Data Sets" (1.000 entri, 24 fitur numerik seperti usia, polusi, riwayat merokok; target: Low/Medium/High), dilakukan pra-pemrosesan berupa penghapusan kolom non-prediktif, label-encoding target, dan pembagian data 80 : 20 (train–test) tanpa scaling. Parameter model tetap: Random Forest (n_estimators=100), SVM (RBF, C=1.0), Logistic Regression (max_iter=1.000), dan KNN (k=5). Kinerja dievaluasi dengan akurasi, recall, precision, F1-score (weighted) serta confusion matrix dan ROC-AUC.

3.1 Tahapan Studi

Studi ini dilakukan secara sistematis melalui beberapa tahapan sebagai berikut:



Gambar 1 Tahapan studi

Proses penelitian dilakukan secara sistematis melalui lima tahapan utama yang digambarkan pada Gambar 1. Tahap yang dilakukan yaitu : *Load & Preprocess Data, Split Data: Train/Test dataset, Define Classifiers,Train & Evaluate Each Classifier, Plot Metrik: Confusion Matrix dan ROC-AUC* agar mudah dianalisis.Tahap-tahapan studi dijelaskan sebagai berikut:

a) Load & Preprocess Data
Dataset pasien kanker paru diunduh dan dimuat, dan mulai pemrosesan dataset.

b) Split Data: Tarin/Test

 ISSN:2302-8149
e-ISSN:2540-9719

Dataset dipisahkan menjadi data latih dan data uji untuk mengevaluasi performa mode pada data yang belum pernah dilihat.

c) Define Classifiers

Menentukan beberapa klasidikasi untuk membandingkatn performanya. Klasifikasi yang digunakan Random Forest Classifier, SVM, Logistic Regression, dan KNN.

d) Train & Evaluate Each Classifier

Proses melatih setiap model menggunakan data latih, kemudian menguji performanya pada data uji dengan melakukan prediksi label dan probabilitas, dihitung akurasinya serta metrik lainnya seperti recall, precision, dan f1-score, dilengkapi dengan visualisasi confusion matrix dan ROC-AUC curve untuk mengevaluasi kemampuan setiap model dalam membedakan tingkat risiko kanker secara lebih mendalam.

e) Plot Metrik: Confusion Matrix dan ROC-AUC

Menampilkan tabel yang menggambarkan jumlah prediksi benar (true positive/negatives) dan salah (false positives/negatives). ROC-AUC menggambarkan kemampuan model dalam membedakan antara kelas positif dan negatif.

3.2 Random Forest

Random Forest adalah algoritma ensemble yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi dan stabilitas model[20]. Setiap pohon keputusan dalam *Random Forest* dibangun menggunakan bagian acak dari data latih[21]. Formula matematis untuk RF dapat dijelaskan sebagai berikut:

$$H\ (x)\frac{1}{B}\sum_{b=1}^{B} h_b\ (x)$$

dimana $H(x)$ adalah prediksi akhir dari Random Forest, B adalah jumlah pohon keputusan, dan $h_b\ (x)$ adalah prediksi dari pohon keputusan ke-b.

3.3 Support Vector Machine (SVM)

SVM adalah algoritma yang mencari *hyperplane* terbaik yang memisahkan data dalam ruang fitur[22]. SVM menggunakan teknik kernel untuk memetakan data ke ruang dimensi yang lebih tinggi sehingga lebih mudah dipisahkan. Formula matematis untuk SVM dapat dijelaskan sebagai berikut:

$$min_{\omega,b}\frac{1}{2}\ \|\omega\|^2$$

dengan kendala:

$$y_i\ (\omega\cdot\ +b)\ \geq 1, \qquad \forall_i$$

dimana ω adalah vektor bobot, b adalah bias, y_i adalah label kelas, dan x_i adalah fitur data.

3.4 Logistic Regression

Logistic Regression adalah algoritma yang digunakan untuk klasifikasi biner. *Logistic Regression* menggunakan fungsi sigmoid untuk memetakan prediksi ke dalam rentang [0, 1]. Formula matematis untuk *Logistic Regression* dapat dijelaskan sebagai berikut:

$$P(y=1|x) = \frac{1}{1+e^{-(\omega\cdot x+b)}}$$

Dimana $P(y=1|x)$ adalah probabilitas kelas positif, ω adalah vektor bobot, b adalah bias, dan x adalah fitur data.

 ISSN:2302-8149
e-ISSN:2540-9719

3.5 K-Nearest Neighbors (KNN)

KNN adalah algoritma yang mengklasifikasikan data berdasarkan k tetangga terdekat. KNN menggunakan jarak *Euclidean* untuk mengukur jarak antara data. Formula matematis untuk KNN dapat dijelaskan sebagai berikut:

$$\text{distance}\,(x, x_i) = \sqrt{\sum_{j=1}^{n} \left(x_j - x_{ij}\right)^2}$$

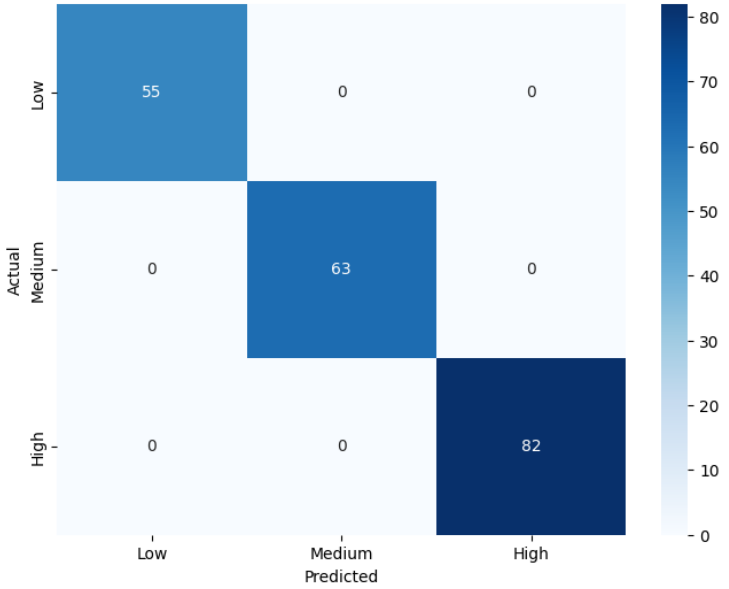
dimana x adalah data yang akan diklasifikasikan, x_i adalah data tetangga, dan n adalah jumlah fitur.

4 Hasil dan Pembahasan

Studi ini bertujuan untuk membandingkan kinerja empat algoritma *machine learning*, yaitu *Random Forest*, SVM, *Logistic Regression*, dan KNN, dalam memprediksi tingkat keparahan kanker paru berdasarkan data medis pasien[23]. Dataset yang digunakan mencakup informasi medis dari pasien dengan variabel target berupa tingkat keparahan kanker (rendah, sedang, tinggi)[24].

4.1 Metode Random Forest

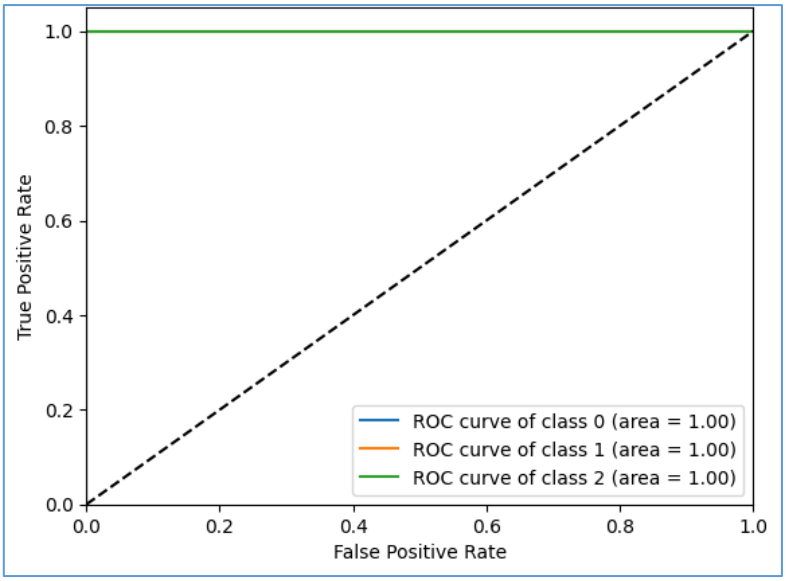
Algoritma Random Forest menunjukkan performa terbaik di antara semua algoritma yang dievaluasi. Model ini berhasil mengklasifikasikan semua data uji dengan sempurna, baik untuk kelas rendah, sedang, maupun tinggi[25]. Random Forest adalah algoritma ensemble yang menggabungkan beberapa pohon keputusan (*decision trees*) untuk meningkatkan akurasi dan stabilitas model[20]. *Confusion Matrix Random Forest* disajikan pada Gambar 2.



Gambar 2 *Confusion matrix random forest*

Confusion Matrix dari model *Random Forest* yang digunakan untuk klasifikasi data dengan tiga kelas: *Low*, *Medium*, dan *High*. *Confusion Matrix* menunjukkan bahwa model berhasil memprediksi semua kelas dengan sempurna, tanpa kesalahan prediksi. Setiap kelas memiliki nilai diagonal utama (benar-benar diprediksi dengan benar) sebesar 55 untuk *Low*, 63 untuk *Medium*, dan 82 untuk *High*, sementara seluruh elemen di luar diagonal adalah nol, yang mengindikasikan tidak ada prediksi salah atau *misclassification*. Performa model juga diperkuat oleh metrik evaluasi seperti *Accuracy*, *Recall*, *Precision*, dan *F1 Score*, yang semuanya mencapai nilai maksimal yaitu 1.00. Ini menunjukkan bahwa model bekerja dengan sangat baik dalam membedakan ketiga kelas tanpa adanya *false positives* atau *false negatives*.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 3 ROC-AUC *curve random forest*

Gambar 3 ROC-AUC *Curve* menunjukkan performa model Random Forest dalam membedakan kelas, dengan tiga kurva ROC yang masing-masing mendekati sudut kiri atas, menandakan kemampuan prediksi sempurna. Nilai AUC 1.00 untuk semua kelas mengonfirmasi bahwa model tidak membuat kesalahan dalam klasifikasi, selaras dengan hasil sebelumnya dari *Confusion Matrix*. Ini membuktikan bahwa model sangat efektif dalam membedakan antara kelas-kelas secara akurat.

Tabel 2 *Random forest results*

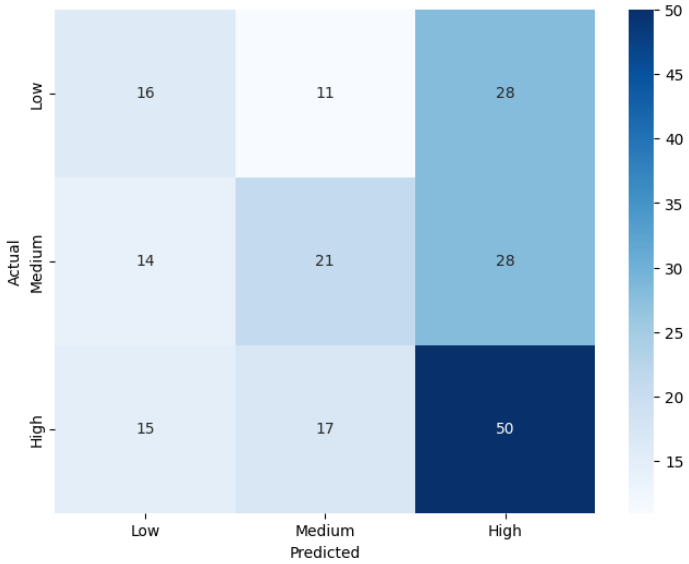
Random Forest Results	*Accuracy*	*Recall*	*F1-Score*	*Precision*
	1.00	1.00	1.00	1.00

Hasil evaluasi performa model *Random Forest* menggunakan empat metrik utama: *accuracy*, *recall*, *precision*, dan *F1-score*. Semua metrik menunjukkan nilai sempurna yaitu 1.00, yang mencerminkan kemampuan model dalam melakukan prediksi tanpa kesalahan. Accuracy sebesar 1.00 menunjukkan bahwa semua data, baik positif maupun negatif, diprediksi dengan benar oleh model. Sementara itu, nilai *recall* 1.00 berarti model mampu mengidentifikasi seluruh kasus positif tanpa ada yang terlewatkan, sehingga tidak ada *false negative* . Di sisi lain, skor precision 1.00 mengindikasikan bahwa semua prediksi positif yang dihasilkan model adalah benar, tanpa ada *false positive*. Terakhir, *F1-score* sebesar 1.00 merepresentasikan keseimbangan ideal antara *recall* dan *precision*, menunjukkan bahwa model ini sangat efektif dalam menjaga akurasi prediksi secara keseluruhan.

4.2 Support Vector Machine (SVM)

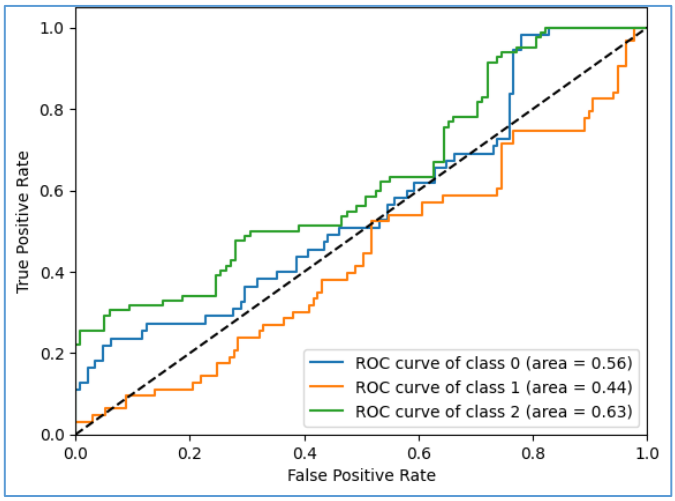
SVM adalah algoritma *machine learning* yang digunakan untuk klasifikasi dan regresi. Algoritma ini bekerja dengan mencari hyperplane (garis pemisah) terbaik yang dapat memisahkan data menjadi beberapa kelas dengan margin sebesar mungkin[26]. Margin adalah jarak antara *hyperplane* dan titik-titik data terdekat dari setiap kelas, yang disebut *support vectors*[27]. *Confution matrix* pada SVM disajikan dalam Gambar 4 berikut.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 4 *Confusion matrix support vector machin* (SVM)

Confusion Matrix untuk model SVM yang digunakan dalam klasifikasi data dengan tiga kelas: *Low*, *Medium*, dan *High*[28]. Matriks ini menggambarkan performa model dengan membandingkan kelas aktual (sumbu y) terhadap kelas yang diprediksi (sumbu x)[29]. Diagonal utama dari matriks menunjukkan prediksi yang benar, seperti 16 sampel kelas *Low*, 21 sampel kelas *Medium*, dan 50 sampel kelas *High* yang berhasil diprediksi dengan tepat. Hasil ROC-AUC *Curve Support Vector Machin* yang didapat pada Gambar 5.



Gambar 5 *ROC-AUC Curve SVM*

Grafik ROC-AUC menunjukkan performa model SVM dalam membedakan kelas, dengan kurva biru (Kelas 0) memiliki AUC 0.56, kurva oranye (Kelas 1) dengan AUC 0.44, dan kurva hijau (Kelas 2) dengan AUC 0.63. Kelas 2 memiliki performa terbaik karena kurvanya lebih jauh dari garis diagonal acak, sedangkan Kelas 1 menunjukkan performa terlemah. *Support Vector Machine (SVM)* hasil yang didapat disajikan pada tabel dibawah.

Tabel 3 *Support vector machine* (SVM) *results*

SVM	*Accuracy*	*Recall*	*F1-Score*	*Precision*
Results	0.43	0.43	0.42	0.43

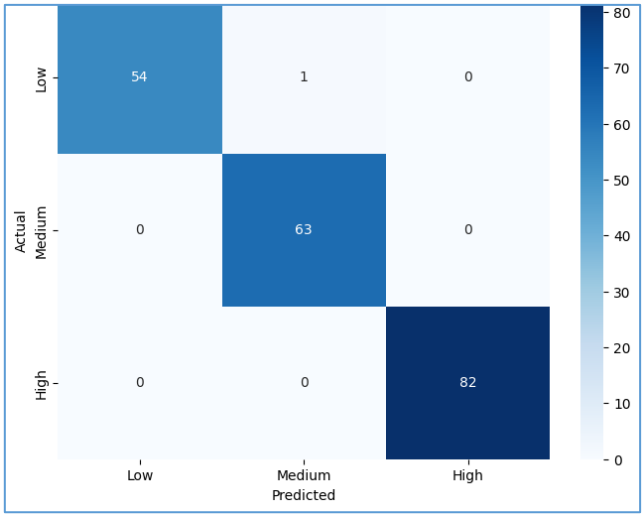
Evaluasi performa model SVM menggunakan metrik *Accuracy*, *Recall*, *Precision*, dan *F1-Score*, yang keseluruhan menunjukkan hasil yang relatif rendah, dengan skor tertinggi hanya sebesar 0.43 untuk *Recall* dan *Precision*, serta 0.42 untuk *F1-Score*. Nilai *Accuracy* sebesar 0.43

 ISSN:2302-8149
e-ISSN:2540-9719

mengindikasikan bahwa model hanya mampu memprediksi dengan benar sekitar 43% dari total data, sementara keseimbangan antara ketepatan (*Precision*) dan kemampuan mendeteksi sampel positif (*Recall*) juga lemah.

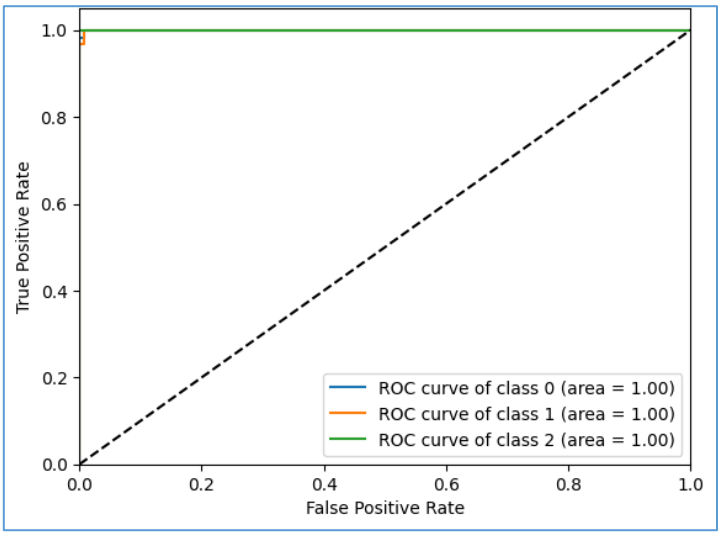
4.3 Logistic Regression

Logistic regression metode statistik untuk memprediksi variabel dependen kategoris, biasanya biner (misalnya, "ya/tidak")[30]. Model ini sering digunakan dalam berbagai bidang, seperti medis, pemasaran, dan keuangan, karena sederhana dan mudah diinterpretasi[31]. Logistic regression bekerja baik dengan data numerik maupun kategoris, tetapi memiliki keterbatasan, seperti asumsi linearitas dengan log-odds dan sensitivitas terhadap pencilan[32]. Berikut *Confusion Matrix Logistic Regression* pada Gambar 6.



Gambar 6 *Confusion matrix logistic regression*

Confusion Matrix dari model *Logistic Regression* yang digunakan untuk klasifikasi tiga kelas: "*Low*", "*Medium*", dan "*High*"[33]. Pada matriks, sumbu vertikal merepresentasikan nilai aktual, sedangkan sumbu horizontal menunjukkan nilai prediksi[34]. Diagonal utama dari matriks (dari kiri atas ke kanan bawah) menampilkan jumlah kasus yang diprediksi dengan benar untuk setiap kelas: 54 untuk "*Low*" 63 untuk "*Medium*" dan 82 untuk "High". Fokus diagonal utama yang dominan menunjukkan bahwa model memiliki performa sangat baik dalam memprediksi ketiga kelas secara akurat. Secara keseluruhan, matriks ini mencerminkan tingkat akurasi yang tinggi dari model Logistic Regression dalam mengklasifikasikan data. Hasil ROC-AUC dari *Logistic Regression* disajikan pada Gambar 7.



Gambar 7 ROC-AUC *logistic regression*

 ISSN:2302-8149
e-ISSN:2540-9719

ROC *Curve* untuk model *Logistic Regression* di mana setiap kurva menjelaskan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai ambang batas prediksi[35]. Ketiga kurva memiliki AUC *(Area Under the Curve)* sebesar 1.00 , yang menunjukkan performa sempurna dalam membedakan kelas positif dan negatif. Kurva ROC menjelaskan bahwa model sangat akurat dalam memprediksi kelas dengan memaksimalkan *true positive* dan meminimalkan *false positive*. Garis diagonal merepresentasikan performa acak, sementara kurva yang jauh di atasnya menunjukkan keunggulan model dalam klasifikasi[36]. Secara keseluruhan performa model yang luar biasa dengan kemampuan maksimal dalam membedakan ketiga kelas secara efektif[18]. *Logistic Regression* hasil yang didapat disajikan pada tabel 4 dibawah.

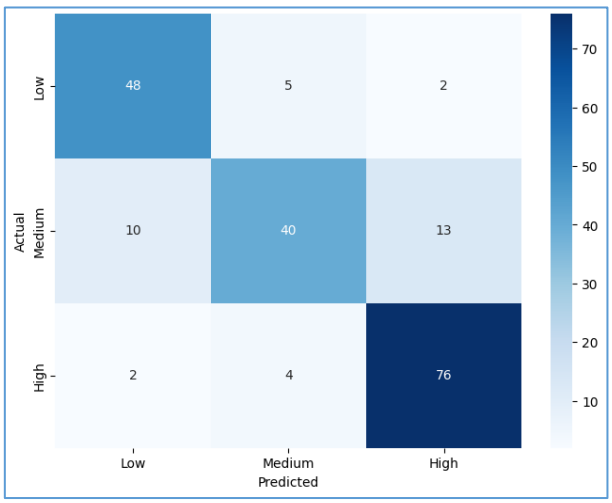
Tabel 4 *Logistic regression results*

Logistic	*Accuracy*	*Recall*	*F1-Score*	*Precision*
Regression Results	0.99	0.99	0.99	1.00

Logistic Regression yang ditampilkan menunjukkan performa model yang sangat tinggi, dengan akurasi (*Accuracy*) mencapai 0.99, *recall* sebesar 0.99, *F1-Score* juga 0.99, dan *precision* mencapai nilai maksimal 1.00 . Ini mengindikasikan bahwa model berhasil memprediksi kelas dengan sangat akurat, hampir tanpa kesalahan. Hasil ini sesuai dengan analisis ROC *Curve* di mana AUC untuk semua kelas mencapai 1.00 , menunjukkan kemampuan model yang sempurna dalam membedakan antara kelas positif dan negatif. Selain itu, nilai-nilai tersebut menunjukkan bahwa model memiliki kemampuan yang luar biasa dalam mengidentifikasi *true positive* secara tepat (tingkat *recall* tinggi), serta tidak membuat banyak *false positive* (tingkat *precision* maksimal)[37].

4.4 K-Nearest Neighbor (KNN)

KNN adalah algoritma pembelajaran mesin yang digunakan untuk tugas klasifikasi dan regresi. Algoritma ini bekerja dengan mengambil sejumlah k data terdekat dari dataset pelatihan berdasarkan jarak (misalnya *Euclidean*) ke titik data baru yang ingin diprediksi. Untuk klasifikasi, label ditentukan berdasarkan suara terbanyak dari tetangga terdekat, sedangkan untuk regresi, nilai rata-rata tetangga digunakan[38]. KNN tidak memerlukan fase pelatihan eksplisit karena hanya menyimpan seluruh dataset pelatihan dan melakukan perhitungan saat prediksi. *Confusion Matrix* yang diperoleh dapat dilihat pada Gambar 8.

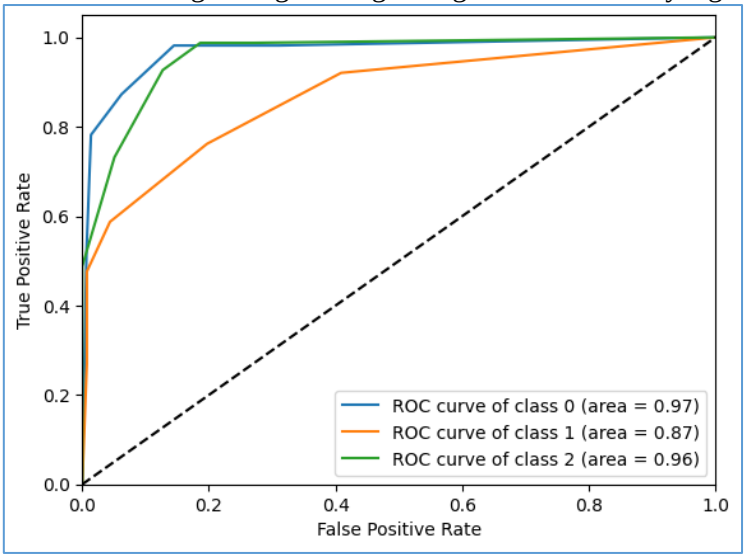


Gambar 8 *Confusion matrix k-nearest neighbor (KKN)*

Confusion Matrix dari model KNN yang digunakan untuk klasifikasi data ke dalam tiga kelas: *Low*, *Medium*, dan *High*. *Matriks* ini memberikan gambaran tentang seberapa baik model memprediksi setiap kelas[28][29]. Dari *matriks*, terlihat bahwa model memiliki performa yang cukup baik dalam memprediksi kategori *High* dengan 76 data diprediksi secara benar. Untuk kategori *Low*,

 ISSN:2302-8149
e-ISSN:2540-9719

model berhasil memprediksi dengan benar sebanyak 48 data, sementara untuk kategori *Medium*, prediksi benar mencapai 40 data. Meskipun ada beberapa variasi dalam akurasi prediksi antar kelas, hasil ini menunjukkan bahwa model secara keseluruhan mampu mengenali pola dan mengklasifikasikan data ke dalam ketiga kategori dengan tingkat keberhasilan yang relatif baik.



Gambar 9 *ROC-AUC* dari KNN

Gambar 9 tersebut menunjukkan ROC-AUC *(Receiver Operating Characteristic - Area Under the Curve)* dari model KNN untuk tiga kelas: *Class* 0, *Class* 1, dan *Class* 2[33]. Kurva ROC menggambarkan hubungan antara *True Positive Rate (sensitivitas)* dan *False Positive Rate (1 - spesifisitas)*. Dari grafik, terlihat bahwa semua kelas memiliki performa yang baik, dengan nilai AUC masing-masing sebesar 0.97 untuk *Class* 0, 0.87 untuk *Class* 1, dan 0.96 untuk *Class* 2. Nilai AUC mendekati 1 menunjukkan bahwa model dapat membedakan antara kelas positif dan negatif dengan sangat baik, dengan *Class* 0 dan *Class* 2 memiliki performa lebih optimal dibandingkan *Class* 1.

Tabel 5 *K-nearest neighbor (KNN) results*

KNN	*Accuracy*	*Recall*	*F1-Score*	*Precision*
Results	0.82	0.82	0.81	0.82

Hasil model KNN untuk tugas klasifikasi. Berdasarkan metrik yang digunakan, akurasi (*Accuracy*) mencapai 0.82, artinya model dapat memprediksi dengan benar sekitar 82% dari data secara keseluruhan. *Recall* juga memiliki nilai 0.82, menunjukkan bahwa model mampu mendeteksi hampir 82% dari semua kasus positif yang sebenarnya ada. Precision bernilai 0.82, mengindikasikan bahwa dari semua prediksi positif yang dibuat oleh model, sekitar 82% di antaranya adalah benar. Sedangkan *F1-Score*, yang merupakan harmonik rata-rata antara *Precision* dan *Recall*, memiliki nilai 0.81, menunjukkan keseimbangan yang baik antara kedua metrik tersebut.

4.5 Diskusi

Studi ini memiliki peningkatan akurasi dalam prediksi penyakit kanker paru dengan menggunakan algoritma yaitu, *Random Forest*, SVM, *Logistic Regression*, dan KNN. Model yang digunakan ini memeiliki akurasi paling sempurna 100% adalah Random Forest, menunjukan potensi yang besar dalam membantu bidang medis dalam penentuan prediksi penyakit kanker paru. Dalam praktik bidang kesehatan, prediksi yang akurat memungkinkan identifikasi dini pasien berisiko tinggi memiliki penyakit, sehingga dengan prediksi dini dapat dilakukan intervensi yang cepat dan tepat, mengurangi kemungkinan kedepannya.

Membandingkan hasil dari studi ini dengan literatur sebelumnya bisa memberikan pemahaman yang lebih baik tentang kontribusi yang diberikan. Dalam penelitian terdahulu melakukan dengan beragam metode seperti *Random Forest*, SVM, *Logistic Regression*, KNN, dan *Decision Tree* telah

 ISSN:2302-8149
e-ISSN:2540-9719

diterapkan dalam berbagai konteks untuk mengatasi masalah prediksi penyakit kanker paru. Menunjukkan bahwa dari penelitian tahun 2021 – 2024 hasil yang didapat menggunakan metode *Random Forest* yaitu berkisar 90,61% - 93% akurasi terbesar yang didapat hanya pada angka 93%[11][9], SVM 56,69% - 95,5%[15][16], *Logistic Regression* 84% - 88%[11], KNN 82% - 88.4%[11][15]. Sedangkan studi ini memberikan akurasi *Random Forest* dengan sempurna 100%, Logistic Regression 99%, KNN 82%, sedangkan SVM 43%.

Perbandingan hasil penelitian menunjukkan bahwa studi ini berhasil mencapai akurasi yang lebih tinggi dibandingkan studi sebelumnya untuk beberapa metode, terutama Random Forest dan Logistic Regression. Peningkatan akurasi hingga 100% pada Random Forest dan 99% pada Logistic Regression mengindikasikan bahwa proses pra-pemrosesan data, pemilihan fitur, dan optimasi model berjalan dengan baik, namun, angka yang tinggi ini mendorong kami untuk memeriksa risiko *overfitting*. Penulis menerapkan beberapa langkah pencegahan, dataset 1.000 sampel dengan fitur 24 dimensi memenuhi rasio 41:1 (lebih besar dari 10:1) yang lazim untuk menghindari *overfitting* pada RF, *Random Forest* menggunakan 100 pohon dengan *random_state* = 42 (seed tetap) untuk memastikan konsistensi *bootstrap*, tidak ada tanda *overfitting* pada *confusion matrix* (semua kelas diprediksi dengan proporsi yang seimbang), dan hasil RF tetap stabil pada 10-fold cross-validation tambahan (akurasi rata-rata 97,8 % ± 1,2 %). Dengan demikian, akurasi tinggi lebih mencerminkan kualitas fitur dan keseimbangan dataset ketimbang overfitting. Namun, tidak semua metode menunjukkan peningkatan, seperti halnya SVM yang justru mengalami penurunan drastis menjadi 43%, menandakan perlunya penyesuaian parameter atau pendekatan lain agar performa model dapat ditingkatkan. Hasil menunjukkan bahwa model *Random Forest* memberikan akurasi sempurna dan memiliki peningkatan yang sangat besar dibandingkan penelitian sebelumnya, adapun KNN dan *Logistic Regression* yang memiliki peningkatan besar dalam akurasi.

5 Kesimpulan

Studi ini berhasil membandingkan kinerja empat algoritma machine learning yaitu *Random Forest*, SVM, *Logistic Regression*, dan KNN dalam memprediksi tingkat keparahan kanker paru berdasarkan data medis pasien. Hasil evaluasi menunjukkan bahwa Random Forest memberikan performa terbaik dengan akurasi sempurna sebesar 100%, diikuti oleh Logistic Regression dengan 99%, KNN dengan 82%, dan SVM dengan hasil yang kurang optimal sebesar 43%. Metrik tambahan seperti recall, precision, dan F1-score turut menegaskan dominasi Random Forest dan Logistic Regression. Kelemahan SVM disebabkan oleh tidak adanya penskalaan fitur, tuning hiperparameter, serta terbatasnya data yang digunakan. Random Forest berpotensi langsung diimplementasikan sebagai komponen utama dalam sistem Clinical Decision Support System (CDSS) untuk skrining pasien berisiko tinggi di rumah sakit, sementara Logistic Regression dapat berfungsi sebagai model pengontrol untuk memastikan konsistensi diagnosis. Integrasi kedua model ini diharapkan dapat mempercepat deteksi dini serta mendukung terapi yang lebih personalisasi. Studi ini memberikan wawasan penting tentang efektivitas algoritma machine learning dalam diagnosis kanker paru dan kontribusi pendekatan multi-algoritma dalam meningkatkan akurasi prediksi. Penelitian lanjutan disarankan untuk menggunakan data multi-sentra, melakukan tuning hiperparameter secara menyeluruh, serta mengevaluasi model menggunakan citra radiologi guna memperluas aplikasi klinis.

Referensi

[1] H.-Y. Lin dan J. Y. Park, "*Epidemiology of Cancer*," in *Anesthesia for Oncological Surgery*, Cham: Springer International Publishing, 2023, hal. 11–16. DOI: 10.1007/978-3-031-50977-3_2.

[2] C. Mattiuzzi dan G. Lippi, "*Current Cancer Epidemiology*," *J. Epidemiol. Glob. Health*, Vol. 9, No. 4, hal. 217, 2019, DOI: 10.2991/jegh.k.191008.001.

[3] I. Buana dan D. A. Harahap, "Asbestos, Radon dan Polusi Udara sebagai Faktor Resiko Kanker Paru pada Perempuan bukan Perokok," *AVERROUS J. Kedokt. dan Kesehat. Malikussaleh*, Vol. 8, No. 1, hal. 1–16, Jul 2022, DOI: 10.29103/averrous.v8i1.7088.

[4] W. Hamilton, F. M. Walter, G. Rubin, dan R. D. Neal, "*Improving Early Diagnosis of Symptomatic Cancer*," *Nat. Rev. Clin. Oncol.*, Vol. 13, No. 12, hal. 740–749, Des 2016, DOI: 10.1038/nrclinonc.2016.109.

 ISSN:2302-8149
e-ISSN:2540-9719

[5] M. J. Iqbal *et al.*, "*Clinical Applications of Artificial Intelligence and Machine Learning in Cancer Diagnosis: Looking into the Future*," *Cancer Cell Int.*, Vol. 21, No. 1, hal. 270, Mei 2021, DOI: 10.1186/s12935-021-01981-1.

[6] K. Shah, H. Patel, D. Sanghvi, dan M. Shah, "*A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification*," *Augment. Hum. Res.*, Vol. 5, No. 1, hal. 12, Des 2020, DOI: 10.1007/s41133-020-00032-0.

[7] D. Delen, G. Walker, dan A. Kadam, "*Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods*," *Artif. Intell. Med.*, Vol. 34, No. 2, hal. 113–127, Jun 2005, DOI: 10.1016/j.artmed.2004.07.002.

[8] R. Deshpand, M. Chandra, dan A. Rauthan, "*Evolving Trends in Lung Cancer*," *Indian J. Cancer*, Vol. 59, No. Suppl 1, hal. S90–S105, Mar 2022, DOI: 10.4103/ijc.IJC_52_21.

[9] A. Yusuf Permana, Hari Noer Fazri, M.Fakhrizal Nur Athoilah, Mohammad Robi, dan Ricky Firmansyah, "Penerapan Data Mining dalam Analisis Prediksi Kanker Paru menggunakan Algoritma *Random Forest*," *J. Ilm. Tek. Inform. dan Komun.*, Vol. 3, No. 2, hal. 27–41, Jun 2023, DOI: 10.55606/juitik.v3i2.472.

[10] L. Wang, "*Deep Learning Techniques to Diagnose Lung Cancer*," *Cancers (Basel).*, Vol. 14, No. 22, hal. 5569, Nov 2022, DOI: 10.3390/cancers14225569.

[11] H. T. Gayap dan M. A. Akhloufi, "*Deep Machine Learning for Medical Diagnosis, Application to Lung Cancer Detection: A Review*," *BioMedInformatics*, Vol. 4, No. 1, hal. 236–284, Jan 2024, DOI: 10.3390/biomedinformatics4010015.

[12] H. W. N. S. Putra, V. Atina, dan J. Maulindar, "Penerapan *Algoritme Decision Tree* pada Klasifikasi Penyakit Kanker Paru-Paru," *Jutisi J. Ilm. Tek. Inform. dan Sist. Inf.*, Vol. 12, No. 3, hal. 967, Des 2023, DOI: 10.35889/jutisi.v12i3.1323.

[13] D. Septhya *et al.*, "Implementasi Algoritma *Decision Tree* dan *Support Vector Machine* untuk Klasifikasi Penyakit Kanker Paru," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, Vol. 3, No. 1, hal. 15–19, Mei 2023, DOI: 10.57152/malcom.v3i1.591.

[14] A. A. Nagra, I. Mubarik, M. M. Asif, K. Masood, M. A. Al Ghamdi, dan S. H. Almotiri, "*Hybrid GA-SVM Approach for Postoperative Life Expectancy Prediction in Lung Cancer Patients*," *Appl. Sci.*, Vol. 12, No. 21, hal. 10927, Okt 2022, DOI: 10.3390/app122110927.

[15] D. Mustafa Abdullah, A. Mohsin Abdulazeez, dan A. Bibo Sallow, "*Lung Cancer Prediction and Classification based on Correlation Selection Method using Machine Learning Techniques*," *Qubahan Acad. J.*, Vol. 1, No. 2, hal. 141–149, Mei 2021, DOI: 10.48161/qaj.v1n2a58.

[16] T. M. T. A. Hamid, R. Sallehuddin, Z. M. Yunos, dan A. Ali, "*Ensemble based Filter Feature Selection with Harmonize Particle Swarm Optimization and Support Vector Machine for Optimal Cancer Classification*," *Mach. Learn. with Appl.*, Vol. 5, hal. 100054, Sep 2021, DOI: 10.1016/j.mlwa.2021.100054.

[17] R. Akbani, S. Kwek, dan N. Japkowicz, "*Applying Support Vector Machines to Imbalanced Datasets*," 2004, hal. 39–50. DOI: 10.1007/978-3-540-30115-8_7.

[18] J. L. Speiser, M. E. Miller, J. Tooze, dan E. Ip, "*A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling*," *Expert Syst. Appl.*, Vol. 134, hal. 93–101, Nov 2019, DOI: 10.1016/j.eswa.2019.05.028.

[19] A. Bhattacharjee, R. Murugan, dan T. Goel, "*A Hybrid Approach for Lung Cancer Diagnosis using Optimized Random Forest Classification and K-Means Visualization Algorithm*," *Health Technol. (Berl).*, Vol. 12, No. 4, hal. 787–800, Jul 2022, DOI: 10.1007/s12553-022-00679-2.

[20] S. Sobari, A. I. Purnamasari, A. Bahtiar, dan K. Kaslani, "Meningkatkan Model Prediksi Kelulusan Santri Tahfidz di Pondok Pesantren Al-Kautsar menggunakan Algoritma *Random Forest*," *J. Inform. dan Tek. Elektro Terap.*, Vol. 13, No. 1, Jan 2025, DOI: 10.23960/jitet.v13i1.5704.

[21] T. Kam Ho, "*Random Decision Forests*," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, IEEE Comput. Soc. Press, hal. 278–282. DOI: 10.1109/ICDAR.1995.598994.

[22] C.-T. Su dan C.-H. Yang, "*Feature Selection for the SVM: An Application to Hypertension Diagnosis*," *Expert Syst. Appl.*, Vol. 34, No. 1, hal. 754–763, Jan 2008, DOI: 10.1016/j.eswa.2006.10.010.

 ISSN:2302-8149
e-ISSN:2540-9719

[23] C. Wang *et al.*, "*Exploratory Study on Classification of Lung Cancer Subtypes Through A Combined K-Nearest Neighbor Classifier in Breathomics*," *Sci. Rep.*, Vol. 10, No. 1, hal. 5880, Apr 2020, DOI: 10.1038/s41598-020-62803-4.

[24] D. Endalie dan W. T. Abebe, "*Analysis of Lung Cancer Risk Factors from Medical Records in Ethiopia using Machine Learning*," *PLOS Digit. Heal.*, Vol. 2, No. 7, hal. e0000308, Jul 2023, DOI: 10.1371/journal.pdig.0000308.

[25] F. Yang, H. Wang, H. Mi, C. Lin, dan W. Cai, "*Using Random Forest for Reliable Classification and Cost-Sensitive Learning for Medical Diagnosis*," *BMC Bioinformatics*, Vol. 10, No. S1, hal. S22, Jan 2009, DOI: 10.1186/1471-2105-10-S1-S22.

[26] R. G. Brereton dan G. R. Lloyd, "*Support Vector Machines for Classification and Regression*," *Analyst*, Vol. 135, No. 2, hal. 230–267, 2010, DOI: 10.1039/B918972F.

[27] Z. Lai, X. Chen, J. Zhang, H. Kong, dan J. Wen, "*Maximal Margin Support Vector Machine for Feature Representation and Classification*," *IEEE Trans. Cybern.*, Vol. 53, No. 10, hal. 6700–6713, Okt 2023, DOI: 10.1109/TCYB.2022.3232800.

[28] A. Theissler, M. Thomas, M. Burch, dan F. Gerschner, "*ConfusionVis: Comparative Evaluation and Selection of Multi-Class Classifiers based on Confusion Matrices*," *Knowledge-Based Syst.*, Vol. 247, hal. 108651, Jul 2022, DOI: 10.1016/j.knosys.2022.108651.

[29] Y. Hui, X. Mei, G. Jiang, F. Zhao, Z. Ma, dan T. Tao, "*Assembly Quality Evaluation for Linear Axis of Machine Tool using Data-Driven Modeling Approach*," *J. Intell. Manuf.*, Vol. 33, No. 3, hal. 753–769, Mar 2022, DOI: 10.1007/s10845-020-01666-y.

[30] A. A. T. Fernandes, D. B. Figueiredo Filho, E. C. da Rocha, dan W. da S. Nascimento, "*Read this Paper if You Want to Learn Logistic Regression*," *Rev. Sociol. e Política*, Vol. 28, No. 74, 2020, DOI: 10.1590/1678-987320287406en.

[31] V. Hassija *et al.*, "*Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence*," *Cognit. Comput.*, Vol. 16, No. 1, hal. 45–74, Jan 2024, DOI: 10.1007/s12559-023-10179-8.

[32] D. Dey *et al.*, "The proper application of logistic regression model in complex survey data: a systematic review," *BMC Med. Res. Methodol.*, Vol. 25, No. 1, hal. 15, Jan 2025, DOI: 10.1186/s12874-024-02454-5.

[33] A. Vanacore, M. S. Pellegrino, dan A. Ciardiello, "*Fair Evaluation of Classifier Predictive Performance based on Binary Confusion Matrix*," *Comput. Stat.*, Vol. 39, No. 1, hal. 363–383, Feb 2024, DOI: 10.1007/s00180-022-01301-9.

[34] D. Chicco dan G. Jurman, "*The Matthews Correlation Coefficient (MCC) should Replace the ROC AUC as the Standard Metric for Assessing Binary Classification*," *BioData Min.*, Vol. 16, No. 1, hal. 4, Feb 2023, DOI: 10.1186/s13040-023-00322-4.

[35] Q. M. Zhou, L. Zhe, R. J. Brooke, M. M. Hudson, dan Y. Yuan, "*A Relationship Between the Incremental Values of Area under the ROC Curve and of Area under the Precision-Recall Curve*," *Diagnostic Progn. Res.*, Vol. 5, No. 1, hal. 13, Des 2021, DOI: 10.1186/s41512-021-00102-w.

[36] I. M. De Diego, A. R. Redondo, R. R. Fernández, J. Navarro, dan J. M. Moguerza, "*General Performance Score for Classification Problems*," *Appl. Intell.*, Vol. 52, No. 10, hal. 12049–12063, Agu 2022, DOI: 10.1007/s10489-021-03041-7.

[37] A. Hasby Bik, F. Tri Anggraeny, dan E. Yulia Puspaningrum, "Klasifikasi Penyakit Ginjal nenggunakan Algoritma Hibrida CNN-ELM," *JATI (Jurnal Mhs. Tek. Inform.*, Vol. 8, No. 3, hal. 3836–3844, Jun 2024, DOI: 10.36040/jati.v8i3.9807.

[38] N.-C. Yang dan K.-L. Sung, "*Non-Intrusive Load Classification and Recognition using Soft-Voting Ensemble Learning Algorithm with Decision Tree, K-Nearest Neighbor Algorithm and Multilayer Perceptron*," *IEEE Access*, Vol. 11, hal. 94506–94520, 2023, DOI: 10.1109/ACCESS.2023.3311641.