Journal of Information System Research (JOSH)
Volume 7, No. 3, April 2026, pp 763−773
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i3.9611

Analisis Perbandingan Algortima Support Vector Machine, Random Forest dan Naive Bayes Untuk Prediksi Penyakit Kanker Paru-Paru

Rendy Alfa Rizky*, Ahmad Fauzi, Dwi Sulistya Kusumaningrum, Hilda Yulia Novita

Teknik Informatika, Universitas Buana Perjuangan Karawang, Karawang
Jl. HS. Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat, Indonesia
Email: [1,*]if20.rendyrizky@mhs.ubpkarawang.ac.id, [2]afauzi@ubpkarawang.ac.id, [3]dwi.sulistya@ubpkarawang.ac.id, [4]hilda.yulia@ubpkarawang.ac.id
Email Penulis Korespondensi: if20.rendyrizky@mhs.ubpkarawang.ac.id
Submitted: **08/04/2026**; Accepted: **30/04/2026**; Published: **30/04/2026**

Abstrak−Paru-paru merupakan salah satu organ vital yang berfungsi dalam proses pernapasan dan sirkulasi darah, dengan kebiasaan merokok sebagai faktor utama yang memicu terjadinya kanker paru-paru. Di Indonesia, prevalensi penyakit ini terus meningkat dan menempatkannya pada urutan kedelapan di kawasan Asia Tenggara. Secara global, kanker paru-paru menyumbang sekitar 11,6% dari keseluruhan kasus kanker serta 18% dari total kematian akibat kanker. Tujuan dari penelitian ini adalah untuk menganalisis dan membandingkan performa algoritma Support Vector Machine (SVM), Random Forest, dan Naive Bayes dalam memprediksi penyakit kanker paru-paru serta menentukan algoritma dengan kinerja terbaik berdasarkan metrik akurasi, precision dan recall. Penelitian ini memanfaatkan dataset Lung Cancer Prediction yang bersumber dari Kaggle, dengan jumlah 309 data dan 16 atribut. Pendekatan yang digunakan adalah penerapan tiga algoritma machine learning, yaitu Support Vector Machine (SVM), Random Forest, dan Naïve Bayes. Tahapan penelitian mencakup pengumpulan data, proses prapengolahan, transformasi data, seleksi fitur, pembangunan model, hingga evaluasi menggunakan confusion matrix. Hasil pengujian menunjukkan bahwa SVM dan Naïve Bayes menghasilkan tingkat akurasi yang sama, yaitu sebesar 91,07%, sedangkan Random Forest memperoleh akurasi 89,28%. Dari sisi metrik evaluasi, SVM menunjukkan performa yang lebih konsisten dengan nilai precision sebesar 95% dan recall 93%, sementara Naïve Bayes memiliki keunggulan pada nilai recall sebesar 95% dengan precision 93%. Di sisi lain, Random Forest menunjukkan kelemahan dalam mengidentifikasi kelas non-kanker. Berdasarkan keseluruhan hasil tersebut, SVM dinilai sebagai metode yang paling optimal karena mampu memberikan keseimbangan kinerja yang lebih baik. Penelitian ini memperlihatkan bahwa pemanfaatan machine learning memiliki potensi yang signifikan sebagai alat bantu dalam mendukung deteksi dini kanker paru-paru secara lebih tepat dan efisien.

Kata Kunci: Kanker Paru-Paru; SVM; Random Forest; Naive Bayes; Prediksi

Abstract−The lungs are one of the vital organs responsible for the processes of respiration and blood circulation, with smoking habits being the primary factor contributing to the development of lung cancer. In Indonesia, the prevalence of this disease continues to increase, placing it eighth in the Southeast Asian region. Globally, lung cancer accounts for approximately 11.6% of all cancer cases and 18% of total cancer-related deaths. This study aims to analyze and compare the performance of Support Vector Machine (SVM), Random Forest, and Naïve Bayes algorithms in predicting lung cancer, as well as to determine the best-performing algorithm based on accuracy, precision, and recall metrics. The study utilizes the Lung Cancer Prediction dataset obtained from Kaggle, consisting of 309 instances and 16 attributes. The approach involves the implementation of three machine learning algorithms, namely Support Vector Machine (SVM), Random Forest, and Naïve Bayes. The research process includes data collection, preprocessing, data transformation, feature selection, model development, and evaluation using a confusion matrix. The experimental results show that both SVM and Naïve Bayes achieve the same accuracy of 91.07%, while Random Forest obtains an accuracy of 89.28%. In terms of evaluation metrics, SVM demonstrates more consistent performance with a precision of 95% and recall of 93%, whereas Naïve Bayes shows a higher recall of 95% with a precision of 93%. On the other hand, Random Forest exhibits limitations in identifying non-cancer cases. Based on the overall results, SVM is considered the most optimal method as it provides a better balance of performance. This study indicates that machine learning has significant potential as a supporting tool for early detection of lung cancer in a more accurate and efficient manner.

Keywords: Lung Cancer; SVM; Random Forest; Naive Bayes; Prediction

1. PENDAHULUAN

Paru-paru merupakan organ vital yang memiliki peran utama dalam sistem pernapasan serta berkontribusi dalam proses sirkulasi darah pada tubuh manusia. Salah satu faktor risiko terbesar yang memicu terjadinya kanker paru-paru adalah kebiasaan merokok. Tidak hanya perokok aktif, individu yang terpapar asap rokok sebagai perokok pasif juga tetap memiliki potensi mengalami penyakit ini, meskipun dengan tingkat risiko yang relatif lebih rendah. Dalam beberapa tahun terakhir, angka kejadian kanker paru-paru di Indonesia menunjukkan tren peningkatan yang cukup signifikan. Kondisi tersebut menempatkan Indonesia pada posisi kedelapan di kawasan Asia Tenggara dengan peningkatan sekitar 10,85% dalam kurun waktu lima tahun terakhir[1]. Secara global, kanker menjadi salah satu penyebab utama kematian dan telah berkembang menjadi permasalahan kesehatan yang serius. Data dari World Health Organization (WHO) menunjukkan bahwa pada tahun 2018 terdapat sekitar 9,6 juta kematian di dunia yang disebabkan oleh kanker[2]. Kanker paru-paru sendiri menyumbang sekitar 11,6% dari total kasus kanker secara global, dengan tingkat kematian mencapai 18%. Di Indonesia, jumlah kasus kanker paru-paru tercatat sekitar 30.023 kasus atau sebesar 8,6%, dengan angka kematian mencapai 12,6% atau sekitar 26.095 kasus[3]. Penyakit ini lebih banyak ditemukan pada laki-laki, sementara pada perempuan berada pada urutan keempat dalam jumlah kejadian. Secara medis, kanker paru-paru terjadi akibat pertumbuhan sel yang tidak terkendali pada jaringan paru, yang umumnya diawali dari terbentuknya tumor ganas pada epitel bronkus

Journal of Information System Research (JOSH)
Volume 7, No. 3, April 2026, pp 763−773
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i3.9611

(bronchogenic carcinoma)[4].

Faktor risiko utama penyakit ini adalah kebiasaan merokok, yang berkontribusi terhadap semua kasus kematian akibat kanker paru-paru. Di lain sisi paparan asap rokok baik pada perokok pasif maupun aktif, polusi udara, serta faktor lingkungan kerja turut meningkatkan risiko terjadinya penyakit. Gejala klinis yang umum dijumpai meliputi batuk berdarah, nyeri dada, suara serak, abses paru, serta sesak nafas. Mengingat tingginya tingkat risiko dan dampak yang ditimbulkan, upaya deteksi dini menjadi sangat penting untuk menekan angka kesakitan maupun kematian akibat penyakit ini. [5]. Perkembangan teknologi, khususnya dalam bidang kecerdasan buatan, memberikan peluang baru dalam mendukung proses diagnosis penyakit. Salah satu pendekatan yang banyak digunakan adalah machine learning, yaitu teknik yang memungkinkan sistem untuk mempelajari pola dari data serta menghasilkan prediksi terhadap data baru secara otomatis[6]. Dalam konteks kesehatan, machine learning mampu membantu dalam mengidentifikasi pola-pola tersembunyi pada data medis yang sulit dianalisis secara manual, sehingga dapat meningkatkan akurasi dalam proses diagnosis[7].

Seiring dengan meningkatnya pemanfaatan machine learning, berbagai algoritma telah digunakan untuk memprediksi penyakit kanker paru-paru. Beberapa metode yang sering diterapkan antara lain Support Vector Machine (SVM), Naïve Bayes, dan Random Forest. SVM bekerja dengan menentukan batas pemisah terbaik antar kelas data, Naïve Bayes menggunakan pendekatan probabilistik yang sederhana namun efektif, sedangkan Random Forest memanfaatkan kombinasi banyak pohon keputusan untuk meningkatkan stabilitas prediksi[8]. Ketiga algoritma tersebut memiliki karakteristik yang berbeda sehingga menghasilkan performa yang bervariasi tergantung pada kondisi data yang digunakan. Untuk mengukur kinerjanya, biasanya digunakan confusion matrix dengan metrik seperti akurasi, precision, dan recall. Namun, beberapa penelitian terbaru menunjukkan bahwa hasil model bisa berbeda tergantung dataset, proses preprocessing, dan pemilihan fitur, sehingga hasil antar penelitian seringkali sulit dibandingkan secara langsung . Karena itu, masih dibutuhkan penelitian yang membandingkan beberapa algoritma dalam satu eksperimen yang sama agar bisa mendapatkan model yang paling optimal untuk prediksi kanker paru-paru[9].

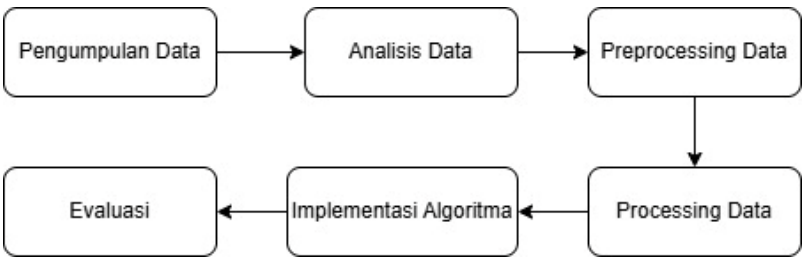
Berbagai penelitian terdahulu sudah menunjukkan bahwa algoritma Machine Learning mempunyai kinerja yang sangat baik dalam prediksi dan mendiagnosis penyakit kanker paru-paru. Dalam penelitian yang menggunakan algoritma SVM dengan dataset dari UCI Machine Learning Repository serta metode k-fold cross-validation (k=10) serta bantuan aplikasi WEKA menghasilkan tingkat akurasi 95,56%[10]. Namun, penelitian lain juga menggunakan algoritma SVM menunjukkan tingkat akurasi lebih rendah sebesar 56,69%[11]. Selain itu, penelitian pada kasus kanker payudara mepergunakan algoritma Decision Tree serta SVM menghasilkan tingkat akurasi 91,92%[12]. Penelitian yang membandingkan algoritma SVM dan K-Nearest Neighbor pada klasifikasi kanker paru-paru memperoleh akurasi sebesar 90%[13]. Sementara itu, penggunaan algoritma Naive Bayes dalam klasifikasi kanker paru-paru menunjukkan tingkat akurasi 94,62%[14]. Pada penelitian lain terkait analisis Machine Learning untuk prediksi kanker paru-paru mempergunakan algoritma random forest menghasilkan akurasi 94,7%[15]. Sedangkan penelitian lain menggunakan penerapan data mining dalam analisis prediksi kanker paru menggunakan random forest menghasilkan akurasi 98,4%[16]. Selain itu, metode Naive Bayes Classifier juga terbukti efektif dalam klasifikasi penyakit kanker paru-paru dengan tingkat akurasi mencapai 97,06%[17].

Berdasarkan hasil kajian dari penelitian sebelumnya, algoritma Machine Learning mempunyai potensi yang baik pada proses diagnosis penyakit kanker paru-paru. Oleh karena itu, penelitian ini dilakukan untuk membandingkan kinerja algoritma Support Vector Machine, Random Forest, dan Naïve Bayes dalam memprediksi penyakit kanker paru-paru. Dengan menggunakan dataset yang sama serta tahapan pengolahan data yang seragam, penelitian ini bertujuan untuk menentukan algoritma yang memiliki performa paling optimal berdasarkan metrik evaluasi yang digunakan. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan di bidang kesehatan, khususnya dalam meningkatkan akurasi deteksi dini kanker paru-paru.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Berikut Gambar 1 merupakan tahapan dari penelitian yang dilakukan



Gambar 1. Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang dirancang secara sistematis guna menghasilkan

Journal of Information System Research (JOSH)
Volume 7, No. 3, April 2026, pp 763−773
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i3.9611

model prediksi yang optimal. Setiap tahapan memiliki peran penting dalam memastikan kualitas data serta keakuratan hasil yang diperoleh. Secara umum, alur penelitian dimulai dari proses pengumpulan data hingga evaluasi model yang dihasilkan.

a. Tahap awal yang dilakukan adalah pengumpulan data. Data yang digunakan dalam penelitian ini diperoleh dari platform Kaggle melalui dataset Lung Cancer Prediction. Dataset tersebut terdiri dari 309 baris data dengan 16 atribut yang merepresentasikan berbagai faktor risiko, gejala, serta kondisi pasien yang berkaitan dengan penyakit kanker paru-paru. Data ini menjadi dasar utama dalam proses analisis dan pemodelan.
b. Selanjutnya dilakukan tahap analisis data merupakan tahap memahami karakteristik dataset sehingga dapat diproses lebih lanjut secara optimal. Beberapa aspek yang dianalisis meliputi jumlah missing value, duplikasi data, serta persentase data yang akan dihapus
    1. jumlah missing value pada tahap ini, missing value merupakan nilai yang tidak tersedia atau kosong pada suatu variabel tertentu, yang jumlahnya dapat bervariasi tergantung pada kondisi dataset.
    2. Jumlah duplikasi data pada tahap ini, setelah mencari missing value dilakukan identifikasi terhadap duplikasi data untuk mengatahui adanya data yang memiliki kesamaan data mengganti dengan nilai 0.
    3. Jumlah persentase data yang akan dihapus pada tahap ini, setelah mencari missing value dan duplikasi data, selanjutnya data yang akan disaring pada tahap ini berguna untuk menghindari data yang sama dan nilai 0
c. Preprocessing Data merupakan bagian dari tahap-tahap untuk mengolah data yang dapat diakses dan dipahami sehingga dapat digunakan untuk tahap selanjutnya dari proses. Adapun beberapa tahapan di Preprocessing Data yaitu Cleaning Data, Transformasi Data, dan Ekstrasi Fitur.
    1. Cleaning Data dalam proses ini, memperbaiki data yang tidak relevan maupun tidak lengkap, sehingga dapat meningkatkan kualitas data dan mengurangi kompleksitas dataset.
    2. Transformasi Data pada tahap transformasi, mengubah dataset mentah menjadi format yang lebih sesuai dalam menganalisis, seperti mengubah tipe data menjadi numerik atau bentuk representasi lain yang lebih mudah di interpretasikan.
    3. Ekstraksi Fitur pada tahap ini, mengidentifikasi dan menghasilkan beberapa fitur baru yang lebih memberi informasi dari data mentah, sehingga dapat meningkatkan performa model dalam melakukan prediksi
d. Processing Data merupakan proses untuk memperoleh data menjadi informasi yang siap dipakai pada pemodelan. Adapun tahapan processing data yaitu Feature Selection dan Split Data.
    1. Feature Selection pada tahap ini, dilakukan untuk memilih beberapa fitur yang relevan serta memiliki kemampuan diskriminatif tinggi dalam membedakan kelas data dengan pemilihan fitur yang optimal, kinerja model dapat ditingkatkan
    2. Split Data pada tahap ini, dilakukan dengan membagi dataset jadi beberapa bagian, seperti data training (latih), data validasi, serta data testing (uji). Yang masing-masing memiliki fungsi dalam proses pengembangan dan evaluasi model.

2.2 Objek Penelitian

Objek penelitian yang digunakan dalam studi ini adalah dataset Lung Cancer Prediction yang diperoleh dari platform Kaggle. Dataset ini terdiri dari sejumlah data yang merepresentasikan kondisi pasien berdasarkan faktor risiko, gejala, serta hasil diagnosis kanker paru-paru. Dataset tersebut dinilai cukup representatif karena memuat berbagai variabel yang berkaitan dengan kondisi kesehatan pasien, seperti usia, jenis kelamin, serta gejala yang muncul. Hal ini memungkinkan model machine learning untuk mempelajari pola hubungan antar variabel secara lebih efektif. Penggunaan dataset ini bertujuan untuk memastikan bahwa model yang dibangun mampu mengidentifikasi pola penting yang berkaitan dengan kanker paru-paru, sehingga dapat digunakan sebagai dasar dalam proses prediksi.

2.3 Kajian Pustaka

Penelitian ini memanfaatkan tiga algoritma klasifikasi dalam kerangka supervised learning, yaitu Support Vector Machine (SVM), Random Forest, dan Naïve Bayes. Ketiga pendekatan tersebut banyak diterapkan pada bidang kesehatan karena mampu mengolah data dengan karakteristik yang kompleks serta menghasilkan model prediksi yang cukup akurat dalam mendukung proses diagnosis penyakit, termasuk kanker paru-paru[18].

Support Vector Machine (SVM) merupakan salah satu metode klasifikasi yang bekerja dengan membangun batas pemisah terbaik (hyperplane) antara dua kelas data. Pendekatan yang digunakan berfokus pada upaya memperlebar margin, yaitu jarak antara batas keputusan dengan data terdekat dari masing-masing kelas. Dengan mekanisme tersebut, SVM dikenal mampu memberikan kemampuan generalisasi yang baik, terutama ketika diterapkan pada data berdimensi tinggi. Selain itu, keberadaan fungsi kernel memungkinkan model ini untuk menangani data yang tidak terpisahkan secara linear dengan cara mentransformasikannya ke ruang fitur yang lebih tinggi. Sejumlah penelitian terkini menunjukkan bahwa SVM cukup efektif dalam menangani permasalahan klasifikasi pada data medis karena mampu menangkap pola yang kompleks[19].

Random Forest merupakan metode berbasis ensemble yang mengombinasikan banyak pohon keputusan untuk meningkatkan kualitas prediksi. Proses pembentukannya dilakukan melalui teknik bootstrap aggregating, di mana setiap model dilatih menggunakan sampel data yang berbeda yang diambil secara acak. Selain itu, pemilihan fitur pada setiap node juga dilakukan secara acak, sehingga mampu mengurangi kecenderungan overfitting yang

Journal of Information System Research (JOSH)
Volume 7, No. 3, April 2026, pp 763−773
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i3.9611

sering muncul pada model tunggal. Pendekatan ini menjadikan Random Forest lebih stabil dan memiliki performa yang konsisten. Beberapa studi terbaru menunjukkan bahwa algoritma ini cukup andal dalam menangani data medis dengan jumlah fitur yang besar serta mampu memberikan tingkat akurasi yang tinggi[20].

Naïve Bayes merupakan metode klasifikasi yang didasarkan pada pendekatan probabilistik dengan mengacu pada Teorema Bayes. Metode ini mengasumsikan bahwa setiap fitur bersifat independen satu sama lain, meskipun dalam praktiknya kondisi tersebut tidak selalu terpenuhi. Namun demikian, algoritma ini tetap mampu memberikan hasil yang cukup baik pada berbagai kasus klasifikasi. Proses prediksi dilakukan dengan menghitung probabilitas posterior berdasarkan probabilitas awal serta kemungkinan kemunculan data pada masing-masing kelas. Keunggulan utama Naïve Bayes terletak pada kesederhanaan perhitungan serta efisiensi dalam pengolahan data berukuran besar[21].

Berdasarkan hal tersebut, penelitian ini dilakukan untuk membandingkan kinerja ketiga algoritma dalam memprediksi kanker paru-paru, sehingga dapat diketahui metode yang memberikan hasil paling optimal. Pendekatan komparatif ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan di bidang kesehatan, khususnya dalam meningkatkan akurasi deteksi dini penyakit.

a. SVM suatu metode klasifikasi yang banyak dipergunakan dalam analisis data, algoritma ini bekerja melalui pemisahan data kedalam 2 kelas dengan penentuan hyperplane pemisah yang optimal[22]. SVM memiliki keunggulan dalam hal kemampuan generalisasi yang tinggi bahkan mampu memberikan model klasifikasi yang optimal walaupun dilatih menggunakan jumlah data yang relatif terbatas[23]

$$f(x) = \sum_{i=1}^{\infty} ai\ yi\ K\ (x.x') + b \quad (1)$$

berdasarkan persamaan tersebut ai merupakan koefisien lagrange pada data ke-i, yi menunjukkan label kelas data, x merupakan data uji, x' adalah data latih, K(x,x') merupakan fungsi kernel yang digunakan untuk menghitung kedekatan antar data, sedangkan b adalah bias yang digunakan untuk menentukan posisi hyperplane.

b. Random Forest metode ensemble learning yang mengombinasikan sejumlah besar decision tree (pohon keputusan) untuk meningkatkan akurasi klasifikasi. Algoritma ini menggunakan teknik Bagging (Bootstrap Aggregating), di mana semua pohon dilatih menggunakan sampel acak yang diambil dengan teknik bootstrap dari kata pelatihan[24].

$$Gini(S_i) = 1 \sum_{i=0}^{c-1} P_i^2 \quad (2)$$

Dengan keterangan Pi adalah hasil dari frekuensi relative kelas Ci pada dataset. Ci adalah kelas untuk I = 1,....,c-1 dan c adalah jumlah kelas yang telah ditentukan.

c. Naive Bayes teknik klasifikasi yang memanfaatkan pendekatan probabilistik dan berlandaskan pada Teorema Bayes yang diperkenalkan Thomas Bayes. Metode tersebut tergolong sederhana, tetapi memiliki tingkat efektivitas yang tinggi dalam melakukan pengelompokan data, terutama dataset berdimensi besar[25].

$$P(H|X) = \frac{P(H|X).P(H)}{P(X)} \quad (3)$$

Keterangan dari persamaan tersebut adalah H merupakan hipotesis kelas, X adalah data yang akan diklasifikasikan, P(H) merupakan probabilitas awal (prior), P(H|X) adalah probabilitas posterior, P(X) adalah probabilitas data, dan P(X|H) adalah likelihood peluangan kemunculan data terhadap hipotesis.

2.4 Evaluasi Model

Evaluasi dilaksanakan untuk mengukur kinerja model algoritma SVM, Random Forest, serta Naive Bayes hasilkan. Metode yang digunakan adalah confusion matrix, yaitu representasi hasil klasifikasi dalam bentuk matriks yang menunjukkan jumlah prediksi yang benar maupun kesalahan klasifikasi[26].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}\ X\ 100\% \quad (4)$$

$$Precision\ \frac{TP}{TP+FP}\ X\ 100\% \quad (5)$$

$$Recall\ \frac{TP}{TP+FN}\ X\ 100\% \quad (6)$$

Dalam persamaan tersebut, TN menunjukkan jumlah data negatif yang diprediksi dengan benar, TP menunjukkan jumlah data positif yang diprediksi dengan benar, FN merupakan jumlah data positif yang salah diklasifikasikan sebagai negatif, dan FP merupakan jumlah data negatif yang salah diklasifikasikan sebagai positif

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Hasil analisis didasarkan pada dataset yang diakses melalui laman Lung Cancer Prediction (kaggle.com). dataset tersebut terdiri dari 4944 data dengan 16 kolom dan 309 baris. Contoh lima data teratas ditampilkan pada Tabel 1,

Journal of Information System Research (JOSH)
Volume 7, No. 3, April 2026, pp 763−773
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i3.9611

yang memperlihatkan struktur awal data sebelum dilakukan proses pengolahan. Pada tabel tersebut terlihat beberapa atribut utama sepeti AGE, GENDER, SMOKING, dan LUNG_CANCER yang digunakan sebagai variabel dalam penelitian. Data yang ditampilkan menunjukkan bahwa dataset memiliki kombinasi antara variabel numerik dan kategorikal, sehingga diperlukan proses preprocessing sebelum digunakan dalam pemodelan.

Tabel 1. Dataset Lung Cancer

	GENDER	AGE	SMOKING	ANXIETY	PEER_PREASURE	CHRONIC DISEASE	FATIGUE
0	M	69	1	2	2	1	2
1	M	74	2	1	1	1	2
2	F	59	1	1	1	2	2
3	M	63	2	2	2	1	1
4	F	63	1	2	1	1	1

Tabel 2. Dataset Lung Cancer (lanjutan)

ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH
1	2	2	2	2
2	1	1	1	2
1	2	1	2	2
1	1	2	1	1
1	2	1	2	2

Tabel 3. Dataset Lung Cancer (lanjutan)

SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
2	2	YES
2	2	YES
1	2	NO
2	2	NO
1	1	NO

3.2 Analisis Data

Hasil analisis awal data yang ditampilkan pada Gam.bar 3 menggambarkan kondisi dataset sebelum memasuki tahap preprocessing. Dari hasil tersebut dapat diketahui bahwa dataset tidak mengandung missing value, yang menunjukkan bahwa seluruh variabel telah terisi secara lengkap tanpa adanya nilai yang hilang. Kondisi ini menjadi keuntungan tersendiri karena tidak diperlukan penanganan khusus seperti imputasi data. Namun demikian, ditemukan sebanyak 33 data duplikat yang berpotensi menurunkan kualitas analisis apabila tidak segera ditangani.

```
Jumlah Missing Value: 0
Jumlah Duplikasi Data: 33
Persentase Data yang akan dihapus: 0.6674757281553397%
```

Gambar 2. Analisis Data

Selain itu, teridentifikasi bahwa sebesar 0,66% dari total data akan dihapus. Persentase ini berasal dari keberadaan data yang tidak relevan maupun data yang terduplikasi, sehingga perlu dieliminasi untuk menjaga konsistensi dan keakuratan dataset. Meskipun jumlahnya relatif kecil, keberadaan data duplikat dapat menyebabkan distorsi dalam proses pembelajaran model, terutama karena model dapat memberikan bobot berlebih pada pola data yang sama.

3.3 Preprocessing Data

Hasil dari tahapan preprocessing data secara mendetail dapat dilihat pada Gambar 4. Pada fase krusial ini, dilakukan prosedur pembersihan data sistematis yang mencakup pemeriksaan terhadap missing value serta identifikasi duplikasi data. Berdasarkan visualisasi tersebut, dapat dikonfirmasi bahwa dataset berada dalam kondisi optimal dengan jumlah nilai kosong dan data ganda sebesar nol.

```
Jumlah Missing Value: 0
Jumlah Duplikasi Data: 0
```

Gambar 3. Preprocessing Data

Pencapaian ini menunjukkan bahwa integritas informasi tetap terjaga tanpa adanya redundansi yang berisiko mengaburkan pola statistik. Dengan kondisi data yang telah terverifikasi bersih dan konsisten, dataset kini memenuhi standar kualitas tinggi serta validitas yang diperlukan untuk dilanjutkan ke tahap analisis mendalam dan pemodelan prediktif, guna menjamin hasil yang lebih akurat

Journal of Information System Research (JOSH)
Volume 7, No. 3, April 2026, pp 763−773
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i3.9611

Setelah tahap sebelumnya, Hasil transformasi variabel AGE_Level menjadi format numerik ditunjukkan secara detail pada Gambar 5. Langkah ini merupakan bagian esensial dari rekayasa fitur (feature engineering), di mana kategori usia pada dataset dikonversi menjadi representasi angka melalui kolom AGE_Level_Numeric. Prosedur ini diambil sebagai langkah krusial untuk memenuhi persyaratan teknis pemodelan, mengingat sebagian besar algoritma pembelajaran mesin (machine learning) hanya mampu memproses input data dalam bentuk kuantitatif.

	AGE	AGE_Level	AGE_Level_Numeric
0	69	High	3
1	74	High	3
2	59	High	3
3	63	High	3
4	63	High	3
..	...	...	...
271	59	High	3
272	59	High	3
273	55	High	3
274	46	High	3
275	60	High	3

[276 rows x 3 columns]

Gambar 4. Transformasi Age Level Numeric

Melalui transformasi ini, setiap tingkatan usia kini memiliki nilai representatif yang memungkinkan komputer untuk membaca dan mengenali pola data dengan lebih mudah. Penyesuaian format tersebut tidak hanya sekadar mengganti label, tetapi juga bertujuan untuk memastikan bahwa proses pelatihan model berjalan tanpa kendala teknis, sehingga hasil prediksi atau analisis yang dihasilkan nantinya jauh lebih presisi dan konsisten. Dengan demikian, dataset telah memenuhi syarat teknis untuk masuk ke tahapan komputasi yang lebih kompleks.

Langkah berikutnya Setelah melakukan transformasi Age Level Numeric, selanjutnya melakukan Transformasi variabel GENDER dan LUNG_CANCER ditampilkan pada Gambar 6, memperlihatkan hasil konversi variabel GENDER dan LUNG_CANCER ke dalam format angka. Penyesuaian ini dilakukan agar fitur-fitur tersebut dapat dibaca oleh algoritma, mengingat model machine learning pada umumnya memerlukan input numerik untuk melakukan perhitungan. Dengan mengubah data kategorikal menjadi representasi angka, pola dalam dataset menjadi lebih mudah dikenali selama proses pelatihan model.

```
['YES' 'NO']
```

	GENDER	AGE	SMOKING	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	\
0	1	69	1	1	1	2	1	
1	1	74	2	1	2	2	2	
2	0	59	1	2	1	2	1	
3	1	63	2	1	1	1	1	
4	0	63	1	1	1	1	1	

	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	\
0	2	2	2	2	
1	1	1	1	2	
2	2	1	2	2	
3	1	2	1	1	
4	2	1	2	2	

	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER	AGE_Level_Numeric
0	2	2	1	3
1	2	2	1	3
2	1	2	0	3
3	2	2	0	3
4	1	1	0	3

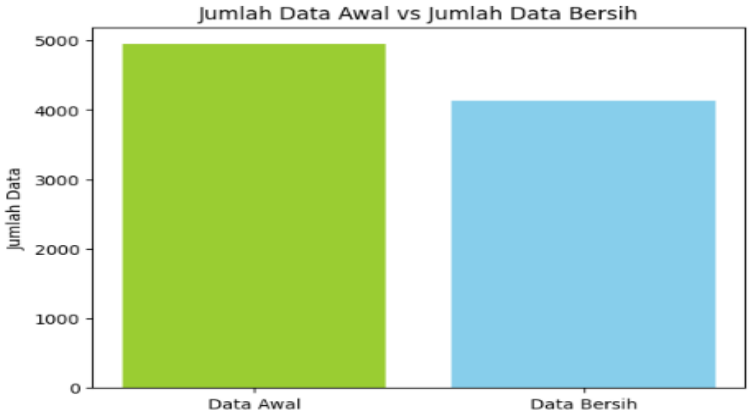
Gambar 5. Transformasi Gender dan Lung Cancer

Dengan mengubah nilai-nilai tersebut menjadi angka, struktur dataset menjadi lebih konsisten sehingga memudahkan model dalam menangkap pola-pola penting selama fase pelatihan. Proses ini tidak hanya meminimalkan risiko kesalahan teknis saat komputasi, tetapi juga membantu mempercepat waktu eksekusi program. Secara keseluruhan, tahapan ini memastikan bahwa variabel-variabel kunci dalam penelitian ini telah siap sepenuhnya untuk digunakan dalam membangun model prediksi yang lebih presisi dan efisien.

Setelah melewati seluruh rangkaian pembersihan dan seleksi, perbandingan volume data sebelum dan sesudah tahap preprocessing dapat dilihat pada Gambar 6. Grafik ini menunjukkan adanya penyusutan jumlah baris yang cukup jelas; dari yang awalnya sebanyak 4.944 data, kini menjadi 4.140 data siap pakai. Pengurangan ini terjadi karena adanya tindakan tegas dalam membuang data ganda serta mengatasi informasi yang tidak lengkap agar tidak merusak kualitas dataset. Walaupun angka akhirnya berkurang, prosedur ini sangat penting untuk

Journal of Information System Research (JOSH)
Volume 7, No. 3, April 2026, pp 763−773
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i3.9611

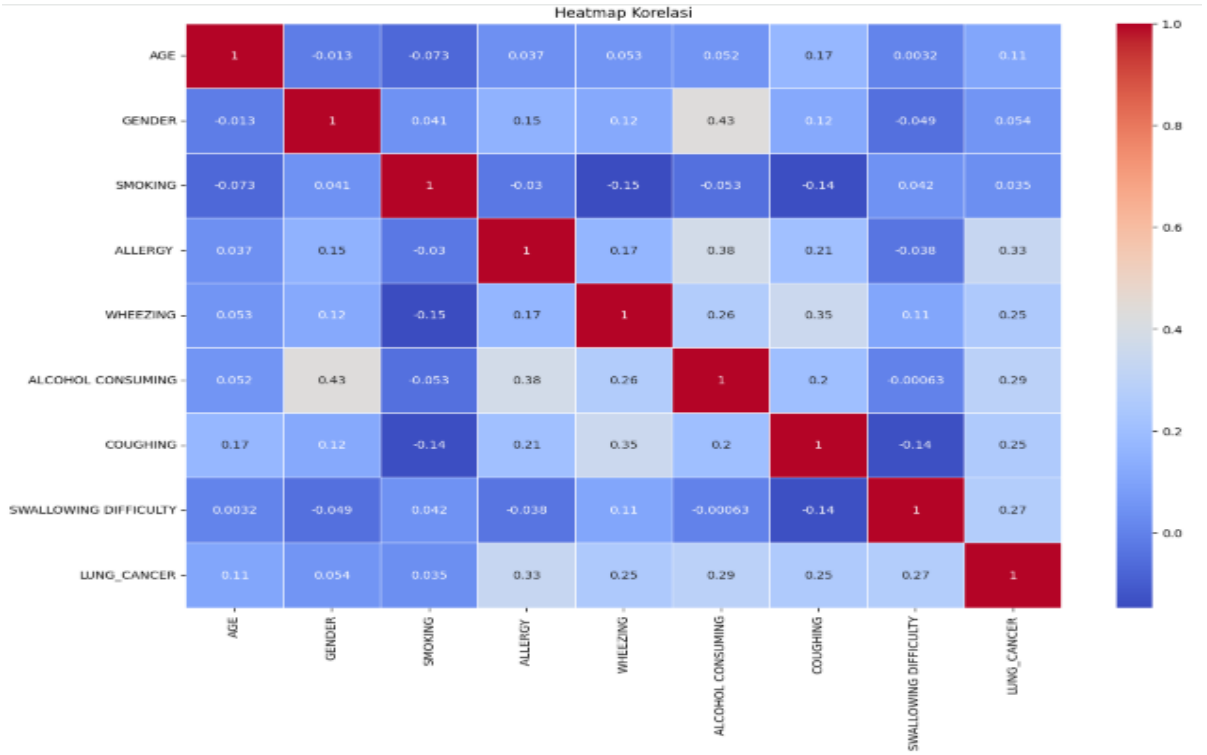
memastikan bahwa model hanya mengolah informasi yang valid. Dengan membuang entri yang cacat atau berulang sehingga hasil analisis nantinya jauh lebih akurat.



Gambar 6. Data Awal dan Data Bersih

3.4 Processing Data

Setelah menyelesaikan tahapan preprocessing, yang meliputi penghapusan missing values, baris kosong, kolom kosong, serta transformasi variabel seperti AGE_Level_Numeric, GENDER, dan LUNG_CANCER, dilakukan analisis korelasi antar variabel dalam dataset menggunakan heatmap. Visualisasi ini bertujuan untuk menunjukkan hubungan antara berbagai faktor risiko, gejala, dan diagnosis kanker paru-paru. Berdasarkan hasil korelasi pada heatmap, terdapat beberapa hubungan yang cukup menonjol terhadap variabel target LUNG_CANCER. Variabel ALLERGY memiliki korelasi tertinggi dengan nilai sebesar 0.33, yang menunjukkan adanya hubungan positif yang cukup kuat terhadap kejadian kanker paru-paru. Selanjutnya, variabel ALCOHOL CONSUMING menunjukkan korelasi sebesar 0.29, diikuti oleh SWALLOWING DIFFICULTY sebesar 0.27, serta WHEEZING dan COUGHING yang masing-masing memiliki nilai korelasi sebesar 0.25. Selain itu, variabel AGE memiliki korelasi yang relatif rendah terhadap LUNG_CANCER, yaitu sebesar 0.11, sementara GENDER dan SMOKING menunjukkan hubungan yang sangat lemah dengan nilai masing-masing sebesar 0.054 dan 0.035. Hal ini mengindikasikan bahwa dalam dataset ini, gejala klinis lebih berpengaruh dibandingkan faktor demografis terhadap prediksi kanker paru-paru. Secara keseluruhan, heatmap ini menunjukkan bahwa variabel yang berkaitan dengan gejala pernapasan dan kondisi tubuh memiliki kontribusi yang lebih signifikan dalam mendeteksi kanker paru-paru dibandingkan variabel lainnya, sebagaimana ditunjukkan pada Gambar 7.

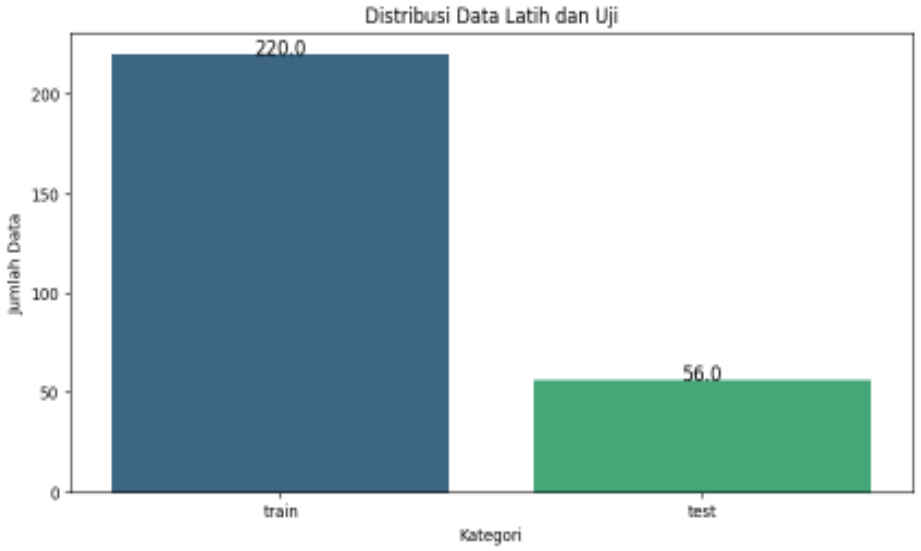


Gambar 7. Feature Selection

Journal of Information System Research (JOSH)
Volume 7, No. 3, April 2026, pp 763−773
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i3.9611

3.5 Split Data

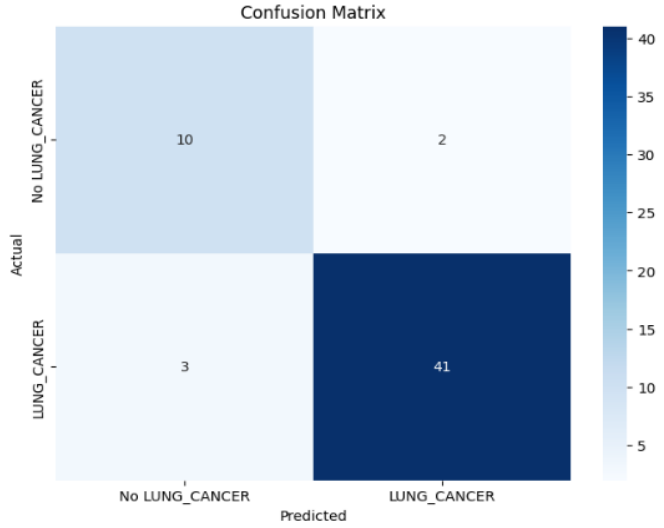
Pada tahap ini, setelah data melewati proses pre-processing serta processing, langkah berikutnya adalah melakukan pembagian dataset menjadi data latih serta data uji. Proses pembagian ini dilaksanakan dengan menerapkan metode split validation melalui perbandingan 80% data latih serta 20% data uji. Pembagian dataset tersebut mempunyai tujuan untuk memastikan bahwa model yang dibangun mempergunakan algoritma machine learning bisa dilatih dengan optimal bahkan diuji performanya secara objektif. Berdasarkan pembagian yang dilakukan, jumlah data latih sebanyak 220 data, serta data uji berjumlah 56 data. Ilustrasi pembagian dataset tersebut bisa diamati di Gambar 8.



Gambar 8. Split Data

3.6 Implementasi Algoritma SVM

Pada tahap ini merupakan confusion matrix dari hasil evaluasi model dalam mengklasifikasi data kanker paru-paru. Berdasarkan visualisasi model berhasil mengklasifikasi 41 pasien yang benar-benar menderita kanker paru-paru dengan tepat (true positive) dan 10 pasien yang tidak menderita kanker (true negative). Namun, masih terdapat 2 kasus false positive, di mana pasien sehat diprediksi menderita kanker, serta 3 kasus false negative, di mana pasien yang sebenarnya menderita kanker tidak terdeteksi oleh model. Secara keseluruhan, hasil ini memperlihatkan bahwa model mempunyai performa yang baik untuk mengenali kedua kelas, terutama pada kasus positif kanker paru-paru, dengan tingkat kesalahan prediksi yang tergolong rendah bisa diamati Gambar 9.

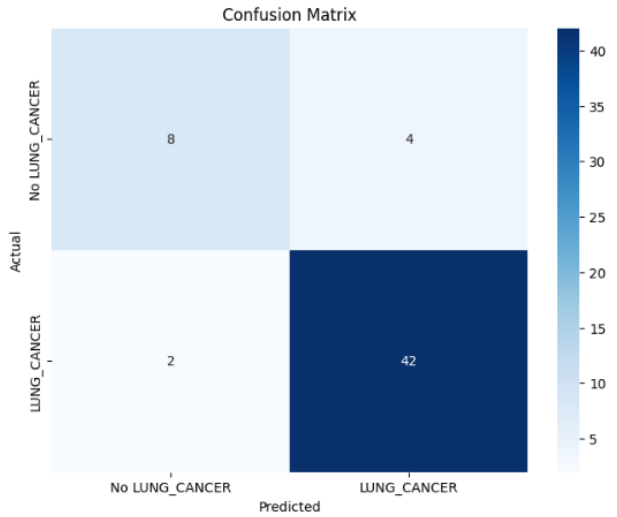


Gambar 9. Confusion Matrix SVM

3.7 Implementasi Algoritma Random Forest

Pada tahapan ini, data sudah melalui serangkaian proses seperti preprocessing, processing, serta split data. Pada tahap implementasi algoritma Random Forest untuk menghasilkan prediksi kanker paru-paru, visualisasi hasilnya dapat dilihat pada confusion matrix.

Journal of Information System Research (JOSH)
Volume 7, No. 3, April 2026, pp 763−773
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i3.9611

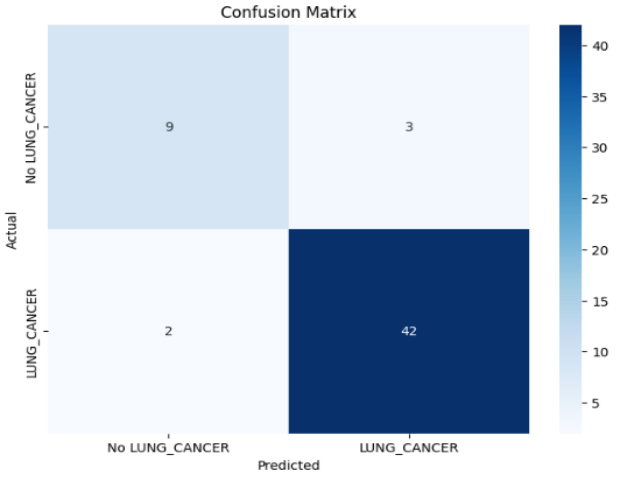


Gambar 10. Confusion Matrix Random Forest

Menunjukkan confusion matrix dari hasil prediksi model terhadap klasifikasi kanker paru-paru. Model berhasil mengklasifikasikan 42 pasien yang benar-benar menderita kanker paru-paru secara tepat (true positive) dan 8 pasien yang tidak menderita kanker (true negative). Namun demikian, terdapat 4 kasus false positive, yang mana pasien yang sebenarnya tidak menderita kanker diprediksi menderita kanker, serta 2 kasus false negative, yaitu pasien yang menderita kanker tetapi tidak terdeteksi oleh model. Hasil tersebut memperlihatkan bahwa model mempunyai kemampuan sangat baik untuk mengenali pasien yang menderita kanker paru-paru, namun masih perlu ditingkatkan dalam mengurangi kesalahan prediksi pada pasien yang sehat dapat dilihat pada Gambar 10.

3.8 Implementasi Algoritma Naive Bayes

Pada tahapan ini, data sudah melalui serangkaian proses seperti preprocessing, processing, serta split data. Pada tahap implementasi algoritma Naïve Bayes untuk menghasilkan prediksi kanker paru-paru, visualisasi hasilnya dapat dilihat pada confusion matrix.



Gambar 11. Confusion Matrix Naive Bayes

Confusion matrix di atas memperlihatkan performa model dalam mengklasifikasikan data kanker paru-paru. Model berhasil mengidentifikasi dengan benar 42 pasien yang menderita kanker paru-paru (true positive) dan 9 pasien yang tidak menderita kanker (true negative). Namun, terdapat 3 kasus false positive, yaitu pasien yang sebenarnya sehat namun diprediksi menderita kanker, serta 2 kasus false negative, yaitu pasien yang menderita kanker namun tidak terdeteksi oleh model. Hasil tersebut mengindikasikan bahwa model mempunyai kemampuan klasifikasi yang begitu baik, khususnya untuk mendeteksi kasus positif kanker paru-paru. Tingkatan kesalahan yang rendah juga menunjukkan bahwa model cukup andal dan layak untuk digunakan sebagai alat bantu diagnosis awal dapat dilihat pada Gambar 11.

3.9 Evaluasi

Pada tahap evaluasi model, dilakukan pengujian terhadap tiga algoritma klasifikasi yakni SVM, Random Forest Classifier, serta Naïve Bayes. Evaluasi dilaksanakan dengan menerapkan metrik recall, precision, akurasi serta f1-score terhadap dua kelas, yaitu (0) tidak menderita kanker paru-paru dan (1) menderita kanker paru-paru. Hasil

Journal of Information System Research (JOSH)
Volume 7, No. 3, April 2026, pp 763−773
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i3.9611

evaluasi disajikan sebagai berikut. Berdasarkan Tabel 2 menampilkan hasil pengujian tiga model klasifikasi, yaitu SVM, Random Forest (RMC), dan Naive Bayes (NB), dengan menggunakan metrik accuracy, precision, dan recall pada masing-masing kelas (No = 0 dan Yes = 1). Berdasarkan hasil, SVM memperlihatkan performa yang paling stabil dengan nilai accuracy sebesar 91% pada kedua kelas. Pada kelas positif (Yes = 1), SVM juga unggul dengan precision mencapai 95% dan recall sebesar 93%, yang menunjukkan kemampuannya dalam mendeteksi kasus kanker secara tepat sekaligus menekan kesalahan prediksi.

Tabel 4. Hasil Evaluasi

Model and	No (0)			Yes (1)		
Accuracy Score	Accuracy	Precision	Recall	Accuracy	Precision	Recall
SVM (91.07%)	91%	77%	83%	91%	95%	93%
RMC (89.28)	89%	80%	67%	89%	91%	95%
NB (91.07)	91%	82%	75%	91%	93%	95%

Naive Bayes memiliki performa yang cukup kompetitif dengan accuracy yang sama, yaitu 91%. Model ini mencatat recall tertinggi pada kelas positif sebesar 95%, yang berarti sangat baik dalam menemukan kasus kanker. Namun demikian, nilai precision sebesar 93% masih sedikit di bawah SVM, sehingga potensi kesalahan pada prediksi positif tetap ada. Di sisi lain, juga memperlihatkan bahwa Random Forest (RMC) memiliki kinerja yang lebih rendah dibandingkan kedua model lainnya. Hal ini terlihat dari accuracy sebesar 89% serta nilai recall pada kelas negatif yang hanya 67%, sehingga model ini kurang optimal dalam mengidentifikasi data non-kanker. Secara keseluruhan, SVM dapat dinilai sebagai model yang paling optimal karena mampu menjaga keseimbangan antara precision dan recall, terutama pada kelas positif. Hal ini menunjukkan bahwa SVM lebih andal dalam menghasilkan prediksi yang akurat dan konsisten untuk deteksi kanker paru-paru

4. KESIMPULAN

Berdasarkan hasil penelitian, ketiga algoritma machine learning yang digunakan, yaitu SVM, Random Forest, dan Naïve Bayes, terbukti dapat dimanfaatkan untuk memprediksi kanker paru-paru dengan performa yang cukup baik. Hal ini ditunjukkan dari nilai akurasi yang relatif tinggi pada masing-masing model. SVM dan Naïve Bayes mencatat akurasi tertinggi sebesar 91,07%, sementara Random Forest sedikit berada di bawahnya dengan nilai 89,28%. Meski demikian, jika ditinjau lebih mendalam melalui metrik precision dan recall pada tiap kelas, SVM menunjukkan kinerja yang lebih stabil dan seimbang. Model ini mampu menjaga ketepatan prediksi sekaligus memiliki kemampuan deteksi yang baik, terutama dalam mengidentifikasi kasus kanker paru-paru. Kondisi ini menjadi penting karena dalam konteks medis, kesalahan dalam mendeteksi kasus positif perlu diminimalkan. Di sisi lain, Random Forest masih memiliki keterbatasan, terutama dalam mengenali data non-kanker yang tercermin dari nilai recall yang lebih rendah. Sementara itu, Naïve Bayes sebenarnya menunjukkan performa yang cukup bersaing, khususnya dalam hal kemampuan mendeteksi kasus positif, namun masih cenderung menghasilkan kesalahan klasifikasi yang lebih tinggi dibandingkan SVM. Dengan mempertimbangkan seluruh hasil evaluasi tersebut, SVM dapat dinilai sebagai metode yang paling sesuai dalam penelitian ini. Model ini tidak hanya memberikan akurasi yang tinggi, tetapi juga mampu menjaga keseimbangan antara precision dan recall. Oleh karena itu, SVM berpotensi untuk diterapkan sebagai model prediksi yang andal dalam mendukung deteksi dini kanker paru-paru. Secara umum, penelitian ini juga menegaskan bahwa pemanfaatan machine learning dapat memberikan kontribusi yang signifikan dalam meningkatkan kualitas diagnosis di bidang kesehatan.

REFERENCES

[1] K. Jainudin and A. Abdullah, “Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Metode Decision Tree C4.5,” Justek Jurnal Sains dan Teknologi, vol. 8, no. 3, pp. 232–240, 2025, doi: 10.31764/justek.v8i3.31981.
[2] I. F. Rosyid and H. Pramaditya, “Visual Interpretation of Machine Learning Models (Random Forest) for Lung Cancer Risk Classification Using Explainable Artificial Intelligence (SHAP & LIME),” JUTIF Jurnal Teknik Informatika, vol. 6, no. 4, pp. 2187–2206, 2025, doi: 10.52436/1.jutif.2025.6.4.4925.
[3] T. D. Putra, E. Utami, and M. P.Kurniawan, “Klasifikasi penderita kanker Paru Paru Menggunakan Algoritma Artificial Neural Network (ANN),” Explore, vol. 12, no. 2, p. 13, 2022, doi: 10.35200/explore.v12i2.568.
[4] 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, pp. 15–19, 2023, doi: 10.57152/malcom.v3i1.591.
[5] L. Sari, A. Romadloni, and R. Listyaningrum, “Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest,” Infotekmesin, vol. 14, no. 1, pp. 155–162, 2023, doi: 10.35970/infotekmesin.v14i1.1751.
[6] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, “Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung,” Inform. J. Ilmu Komput., vol. 18, no. 3, p. 239, 2022, doi: 10.52958/iftk.v18i3.4694.
[7] Y. Li, X. Wu, P. Yang, G. Jiang, and Y. Luo, “Machine Learning for Lung Cancer Diagnosis , Treatment , and Prognosis,” Genomics. Proteomics Bioinformatics, vol. 20, no. 5, pp. 850–866, 2022, doi: 10.1016/j.gpb.2022.11.003.
[8] B. Shafa, H. H. Handayani, S. Arum, and P. Lestari, “Prediksi Kanker Paru dengan Normalisasi menggunakan

Journal of Information System Research (JOSH)
Volume 7, No. 3, April 2026, pp 763−773
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i3.9611

Perbandingan Algoritma Random Forest , Decision Tree dan Naïve Bayes," DECODE Jurnal Pendidikan Teknologi Informasi, vol. 4, no. 3, pp. 1057–1070, 2024, doi: 10.51454/decode.v4i3.779.

[9] S. P. Maurya, P. S. Sisodia, R. Mishra, and D. Pratap, "Performance of machine learning algorithms for lung cancer prediction : a comparative approach," Sci. Rep., pp. 1–11, 2024, doi: 10.1038/s41598-024-58345-8.

[10] S. Muawanah, U. Muzayanah, M. G. R. Pandin, M. D. S. Alam, and J. P. N. Trisnaningtyas, "Stress and Coping Strategies of Madrasah's Teachers on Applying Distance Learning During COVID-19 Pandemic in Indonesia," Qubahan Acad. J., vol. 3, no. 4, pp. 206–218, 2023, doi: 10.48161/Issn.2709-8206.

[11] T. M. T. A. Hamid, R. Sallehuddin, Z. M. Yunos, and 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, no. May, p. 100054, 2021, doi: 10.1016/j.mlwa.2021.100054.

[12] T. A. Assegie and S. S. J., "A Support Vector Machine and Decision Tree Based Breast Cancer Prediction," Int. J. Eng. Adv. Technol., vol. 9, no. 3, pp. 2972–2976, 2020, doi: 10.35940/ijeat.a1752.029320.

[13] A. Desiani et al., "Perbandingan Klasifikasi Penyakit Kanker Paru-Paru menggunakan Support Vector Machine dan K-Nearest Neighbor," J. Process., vol. 18, no. 1, pp. 54–62, 2023, doi: 10.33998/processor.2023.18.1.700.

[14] E. Wulandari, "Klasifikasi Kanker Paru-Paru Menggunakan Metode Naive Bayes," Int. Res. Big-Data Comput. Technol. I-Robot, vol. 6, no. 2, pp. 20–24, 2022, doi: 10.53514/ir.v6i2.325.

[15] A. N. Am, M. Nurkholifah, and F. K. Oktorina, "Analisa Penyakit Jantung Menggunakan Algoritma Naïve Bayes," J. Syst. Comput. Eng., vol. 4, no. 1, pp. 26–36, 2023, doi: 10.47650/jsce.v4i1.671.

[16] M. Y. Iskandar and H. W. Nugroho, "Comparative Evaluation of Decision Tree and Random Forest for Lung Cancer Prediction Based on Computational Efficiency and Predictive Accuracy," JUTIF Jurnal Teknologi Informatika, vol. 6, no. 5, pp. 3392–3404, 2025, doi: 10.52436/1.jutif.2025.6.5.4877.

[17] M. Y. Haffandi, E. Haerani, F. Syafria, and L. Oktavia, "Klasifikasi Penyakit Paru-Paru Dengan Menggunakan Metode Naïve Bayes Classifier," J. Tek. Inf. dan Komput., vol. 5, no. 2, p. 176, 2022, doi: 10.37600/tekinkom.v5i2.649.

[18] M. Amine et al., "Heliyon Early heart disease prediction using feature engineering and machine learning algorithms," Heliyon, vol. 10, no. 19, p. e38731, 2024, doi: 10.1016/j.heliyon.2024.e38731.

[19] F. S. Gomiasti, E. Kartikadarma, J. Gondohanindijo, and D. R. I. Moses, "Enhancing Lung Cancer Classification Effectiveness Through Hyperparameter-Tuned Support Vector Machine," Journal of Computing Theories and Applications, vol. 1 no. 4, pp 396-406, 2024, doi: 10.62411/jcta.10106.

[20] C. M. Lauw, H. Hairani, I. Saifudin, J. X. Guterres, and M. M. Huda, "Combination of Smote and Random Forest Methods for Lung Cancer Classification," IJECSA International Journal of Engineering and Computer Science Applications, vol. 2, no. 2, pp. 63–70, 2023, doi: 10.30812/IJECSA.v2i2.3333.

[21] A. P. Aulia and Q. Adelia, "Lung Disease Risk Prediction Using Machine Learning Algorithms," PREDATECS Public Research Journal of Engineering Data Technology and Computer Science, vol. 3, no. July, pp. 70–79, 2025, doi: 10.57152/predatecs.v3i1.1858

[22] I. A. Purnomo, J. Indra, E. E. Awal, and T. Rohana, "Analisis Prediksi Banjir di Indonesia Menggunakan Algoritma Support Vector Machine dan Random Forest," vol. 6, no. 1, pp. 219–228, 2026, doi: 10.47065/josh.v6i1.5958.

[23] I. Nurul Hassanah, S. Faisal, A. Mutoi Siregar, U. Buana Perjuangan Karawang Jl HSRonggo Waluyo, T. Timur, and J. Barat, "Perbandingan Algoritma Support Vector Machine Dengan Decision Tree Pada Aplikasi Ruang Guru," Kumpul. J. Ilmu Komput., vol. 10, no. 1, pp. 39–50, 2023.

[24] A. Masruriyah, H. Novita, C. Sukmawati, A. Ramadhan, S. Arif, and B. Dermawan, "Pengukuran Kinerja Model Klasifikasi dengan Data Oversampling pada Algoritma Supervised Learning untuk Penyakit Jantung," Comput. Sci., vol. 4, no. 1, pp. 62–70, 2024, doi: 10.31294/coscience.v4i1.2389.

[25] N. C. Ramadhan, H. H. H, T. Rohana, and A. M. Siregar, "Optimasi Algoritma Machine Learning Menggunakan Seleksi Fitur Xgboost Untuk Klasifikasi Kanker Payudara." TIN : Terapan Informatika Nusantara vol. 5, no. 2, pp. 162–171, 2024, doi: 10.47065/tin.v5i2.5408.

[26] I. P. Rahayu, A. Fauzi, and J. Indra, "Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Naive Bayes Dan Support Vector Machine," J. Sist. Komput. dan Inform., vol. 4, no. 2, p. 296, 2022, doi: 10.30865/json.v4i2.5381.