Jurnal Qua Teknika, Vol. 15 No. 2 September 2025
ISSN 2088-2424 (cetak); ISSN 2527-3892 (elektronik)
Fakultas Teknik Universitas Islam Balitar, Blitar
Https://ejournal.unisbablitar.ac.id/index.php/qua; Email: quateknika@unisbablitar.ac.id

Bifadhlillah Marsheila Islami[1)], Sucipto[2)], Arie Nugroho [3)],
ANALISIS ALGORITMA KNN DAN PENERAPAN SMOTE DALAM DETEKSI DINI KANKER PARU-PARU
Jurnal *Qua Teknika*, (2025), 15 (2): 38-50

ANALISIS ALGORITMA KNN DAN PENERAPAN SMOTE DALAM DETEKSI DINI KANKER PARU-PARU

Bifadhlillah Marsheila Islami[1)], Sucipto[2)], Arie Nugroho[3)]
[1, 2, 3] Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri
Jl. Ahmad Dahlan No.76, Mojoroto, Kec. Mojoroto, Kota Kediri, Jawa Timur 64112
email: biff6167@gmail.com[1)], sucipto@unpkediri.ac.id[2)],
arienugroho@unpkediri.ac.id[3)]

ABSTRAK

Kanker paru-paru merupakan salah satu jenis kanker paling mematikan di dunia dengan tingkat kematian tinggi akibat keterlambatan dalam deteksi dini. Latar belakang dari penelitian ini adalah kebutuhan mendesak terhadap sistem klasifikasi yang akurat guna membantu diagnosis dini kanker paru-paru secara non-invasif. Tujuan dari penelitian ini adalah untuk menganalisis kinerja algoritma *K-Nearest Neighbors* (KNN) yang dikombinasikan dengan teknik *Synthetic Minority Oversampling Technique* (SMOTE) dalam meningkatkan akurasi klasifikasi tingkat keparahan kanker paru-paru. Penelitian ini menggunakan dataset dari Kaggle yang terdiri atas 1000 data pasien dengan 26 fitur klinis dan demografis. Metodologi yang digunakan adalah CRISP-DM yang mencakup enam tahap yaitu *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment*. Hasil dari penerapan algoritma KNN dengan nilai k = 3 setelah SMOTE menunjukkan akurasi sebesar 99,50%, serta precision, recall, dan F1-score yang sangat tinggi pada semua kelas. Kesimpulannya, integrasi algoritma KNN dengan SMOTE terbukti efektif dalam mengatasi ketidakseimbangan kelas dan meningkatkan performa klasifikasi secara signifikan, sehingga memiliki potensi untuk dikembangkan lebih lanjut sebagai sistem pendukung keputusan dalam bidang kesehatan.

Kata kunci: kanker paru-paru, KNN, SMOTE, deteksi dini, klasifikasi.

ABSTRACT

Lung cancer is one of the deadliest types of cancer in the world with a high mortality rate due to delays in early detection. The background of this study is the urgent need for an accurate classification system to assist in the early diagnosis of lung cancer non-invasively. The purpose of this study is to analyze the performance of the K-Nearest Neighbors (KNN) algorithm combined with the Synthetic Minority Oversampling Technique (SMOTE) technique in improving the accuracy of lung cancer severity classification. This study uses a dataset from Kaggle consisting of 1000 patient data with 26 clinical and demographic features. The methodology used is CRISP-DM which includes six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The results of the application of the KNN algorithm with a value of k = 3 after SMOTE showed an accuracy of 99.50%, as well as very high precision, recall, and F1-score in all classes. In conclusion, the integration of the KNN algorithm with SMOTE has proven effective in overcoming class imbalance and significantly improving classification performance, thus having the potential to be further developed as a decision support system in the health sector.

Keywords: lung cancer, KNN, SMOTE, early detection, classification.

PENDAHULUAN

Kanker merupakan salah satu penyakit mematikan yang menjadi masalah kesehatan global. Menurut data dari Organisasi Kesehatan Dunia (WHO), penyakit ini menyebabkan sekitar 9,6 juta kematian setiap tahunnya di seluruh dunia[1]. Salah satu betuk kanker yang paling mematikan adalah kanker paru-paru, yang mencakup sekitar 13% dari total kasus kanker secara global [2]. WHO mencatat bahwa kanker paru-paru merupakan penyebab utama kematian akibat kanker, baik pada laki-laki maupun perempuan, tanpa memandang usia.

Data Globocan tahun 2018 menunjukkan bahwa tingkat kejadian kanker paru-paru mencapai 11,6% dengan angka kematian mencapai 18,4%. Di Indonesia, insiden kanker paru tercatat sebesar 8,6%, atau sekitar 30.023 kasus, dengan tingkat kematian sebesar 12,6%, yaitu sekitar 26.095 kasus [3]. Berdasarkan penelitian sebelumnya yang berjudul "Analisa Kanker Paru-paru Menggunakan Algoritma K-Nearest Neighbor" menunjukkan bahwa algoritma K-NN dapat diterapkan dalam klasifikasi penyakit kanker paru-paru dengan memanfaatkan data medis dan faktor risiko terkait. Penelitian ini menggunakan dataset dari Kaggle.com yang terdiri dari 1000 data pasien dengan berbagai parameter seperti usia, jenis kelamin, riwayat kebiasaan merokok, paparan polusi udara, hingga gejala klinis. Tahapan penelitian meliputi pengumpulan data, preprocessing, transformasi data, penerapan algoritma K-NN, serta evaluasi model. Hasil menunjukkan akurasi sebesar 80,40% dengan penerapan confusion matrix sebagai metrik evaluasi. Nilai k pada algoritma K-NN menjadi parameter penting yang turut memengaruhi performa model. Penelitian tersebut membuktikan bahwa pendekatan data mining, khususnya K-NN, memiliki potensi untuk mendukung deteksi dini kanker paru-paru secara lebih efektif[4].

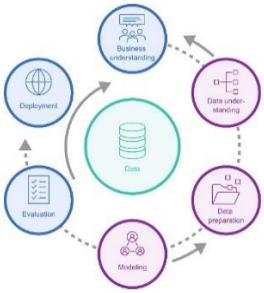
Permasalahan utama yang melatarbelakangi penelitian ini adalah belum tersedianya sistem deteksi dini kanker paru-paru yang non-invasif dan dapat diimplementasikan secara praktis menggunakan data klinis. Dataset kanker paru-paru yang digunakan dalam penelitian ini dengan 1000 pasien dan 26 fitur klinis dan demografis menunjukkan adanya ketidakseimbangan kelas (*imbalanced data*), di mana label risiko "*High*" lebih dominan dibandingkan "*Medium*" dan "*Low*". Ketidakseimbangan ini cenderung menurunkan performa klasifikasi dan menyebabkan bias terhadap kelas mayoritas.

Seiring kemajuan teknologi, berbagai pendekatan berbasis *machine learning* telah dikembangkan untuk meningkatkan akurasi deteksi dini kanker paru-paru. Salah satu algoritma yang menunjukkan hasil yang menjanjikan adalah K-Nearest Neighbors (KNN). Algoritma ini merupakan metode klasifikasi sederhana namun efektif yang bekerja berdasarkan prinsip bahwa objek dengan karakteristik serupa cenderung berada dalam kategori yang sama [5]. Pendekatan berbasis algoritma seperti KNN sangat relevan dalam pengenalan dini kanker paru-paru. Penelitian terkini menunjukkan bahwa KNN memiliki kinerja yang kompetitif dalam klasifikasi kanker paru berdasarkan data citra histopatologi, ekspresi gen, dan biomarker klinis [6], [7]. Selain itu, integrasi KNN dengan teknologi non-invasif seperti *electronic nose system* dan *Raman spectroscopy* memungkinkan deteksi dini yang lebih cepat dan akurat, sekaligus mengurangi ketergantungan pada prosedur invasif [8][9].

Penelitian ini menerapkan algoritma K-Nearest Neighbors (KNN) untuk deteksi dini kanker paru-paru menggunakan Jupyter Notebook, yang menawarkan fleksibilitas tinggi dalam *preprocessing, tuning parameter*, dan visualisasi data. Berbeda dengan penelitian sebelumnya oleh Teguh Abdi Mangun yang hanya menggunakan klasifikasi biner tanpa penanganan data tidak seimbang atau evaluasi komprehensif, penelitian ini mengintegrasikan pendekatan multikelas *(Low, Medium, High*), teknik SMOTE untuk penyeimbangan kelas, serta kerangka kerja CRISP-DM yang sistematis. Hasilnya diharapkan dapat mendukung pengembangan sistem diagnosis berbasis AI dalam lingkungan klinis.

METODE PENELITIAN

Penelitian ini menggunakan kerangka kerja CRISP-DM (*Cross-Industry Standard Process for Data Mining*) yang merupakan metodologi umum dalam pengembangan solusi berbasis data mining dan *machine learning*. Pendekatan ini terbagi menjadi enam tahapan utama yaitu:



Gambar 1. Alur CRISP-DM

A. Bussiness Understanding

Pada tahap ini, tujuan bisnis utama adalah meningkatkan deteksi dini kanker paru-paru melalui klasifikasi tingkat keparahan pasien menggunakan algoritma *K-Nearest Neighbors* (KNN). Permasalahan yang ingin diselesaikan adalah akurasi rendah akibat ketidakseimbangan kelas data serta kurangnya representasi sampel dari kelas minoritas. Oleh karena itu, digunakan teknik *oversampling* SMOTE untuk memperbaiki distribusi data.

B. Data Understanding

Dataset yang digunakan merupakan himpunan data pasien kanker paru yang mencakup berbagai fitur klinis dan demografis seperti usia, tingkat pernapasan, frekuensi batuk, dan konsumsi alkohol. Analisis eksplorasi dilakukan terhadap distribusi label (Level) untuk mengetahui derajat ketidakseimbangan antar kelas. Visualisasi awal memperlihatkan dominasi kelas mayoritas, sehingga diperlukan upaya balancing.

Penggunaan dataset dari Kaggle untuk keperluan klasifikasi kanker paru telah dilakukan pula dalam penelitian oleh Rifa'i dan Prabowo, yang menggunakan pendekatan fuzzy logic untuk diagnosis berbasis data gejala [10].

C. Data Preparation

Tahap ini mencakup sejumlah proses penting dalam pra-pemrosesan data untuk memastikan kualitas dan kesiapan data sebelum dilakukan pemodelan. Langkah-langkah yang dilaksanakan meliputi:

1. Penghapusan Atribut

   Penghapusan atribut yang tidak memiliki kontribusi informatif terhadap proses klasifikasi, seperti Patient Id dan index, guna menghindari bias dan redundansi informasi dalam model.
2. Encode Label

   Kolom level menunjukkan Tingkat keparahan kanker paru-paru yang dikategorikan ke dalam tiga kelas, yaitu *Low, Medium* dan *High.* Oleh karena label tersebut masih dalam format *string,* maka dilakukan proses transformasi (*encode*) ke dalam format numerik dengan menggunakan *Label Encoding.*
3. Pemisahan Data

   Dataset dibagi ke dalam dua subset, yaitu data latih dan data uji, menggunakan metode train_test_split untuk mempertahankan proporsi distribusi kelas yang representatif pada masing-masing subset. Hal ini penting untuk menjaga validitas evaluasi model.
4. Penyeimbangan Kelas Menggunakan SMOTE

   Ketidakseimbangan kelas dalam data latih diatasi dengan menerapkan teknik Synthetic Minority Over-sampling Technique (SMOTE), yang secara sintetik menghasilkan data baru dari kelas minoritas guna meningkatkan representasi dan mengurangi bias prediktif terhadap kelas mayoritas. Teknik SMOTE telah terbukti efektif dalam meningkatkan representasi kelas minoritas pada dataset medis, sebagaimana ditunjukkan dalam penelitian klasifikasi kanker paru-paru [11].

D. Modelling

Model yang digunakan adalah algoritma *K-Nearest Neighbors* (KNN), dipilih karena kesederhanaannya, efisiensi dalam klasifikasi berbasis jarak, dan performa yang baik dalam banyak kasus medis. Model dilatih menggunakan data hasil SMOTE dengan parameter *n_neighbors=3* yang diperoleh berdasarkan pendekatan eksperimen awal.

E. Evaluasi Model

Evaluasi dilakukan untuk mengukur kinerja model klasifikasi. Evaluasi ini menggunakan metrik *Confusion Matrix*, yang menggambarkan perbandingan antara hasil prediksi dengan nilai aktual. Confusion Matrix digunakan untuk menilai seberapa akurat hasil dari suatu proses klasifikasi. Akurasi menunjukkan persentase prediksi yang benar dari total keseluruhan prediksi yang dilakukan. Untuk menghitung nilai akurasi, presisi, dan *recall* [12]

F. Deployment

Tahap terakhir adalah merancang implementasi model ke dalam lingkungan nyata atau sistem pendukung keputusan klinis. Meskipun pada penelitian ini tahap deployment masih bersifat konseptual, hasil dan model yang diperoleh dapat diintegrasikan ke dalam aplikasi berbasis web atau sistem klinik sebagai alat bantu diagnosis non-invasif. Dengan demikian, proses CRISP-DM memastikan bahwa seluruh tahapan penelitian dilakukan secara sistematis dan dapat dipertanggungjawabkan secara ilmiah.

HASIL DAN PEMBAHASAN

A. Business Understanding

Pada tahap ini, tujuan utama penelitian adalah meningkatkan deteksi dini kanker paru-paru melalui klasifikasi tingkat keparahan pasien menggunakan algoritma *K-Nearest Neighbors* (KNN). Masalah yang ingin diselesaikan adalah rendahnya akurasi prediksi akibat ketidakseimbangan distribusi kelas dalam dataset serta kurangnya representasi data pada kelas minoritas. Untuk mengatasi hal tersebut, diterapkan teknik *Synthetic Minority Oversampling Technique* (SMOTE) sebagai metode penyeimbangan kelas.

Dengan pendekatan ini, diharapkan model mampu memberikan prediksi yang lebih adil dan akurat terhadap tiga tingkatan risiko kanker paru-paru, yaitu *Low* (Rendah), *Medium* (Sedang), dan *High* (Tinggi). Hasil prediksi ini dapat menjadi dasar bagi sistem pendukung keputusan medis dalam mendeteksi potensi risiko kanker secara dini, khususnya di wilayah dengan keterbatasan akses layanan medis.

B. Data Understanding

a. Collect Initial Data

Data yang digunakan adalah dataset publik yang didapat dari *Kaggle* dan dapat diakses melalui tautan berikut ini: *https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link*. Dataset ini memiliki 1000 sampel data dan 26 fitur, Dapat dilihat pada tabel 1.

Tabel 1. Dataset

NO	Dataset Lung Cancer						
	Index	Patient Id	Age	…..	Dry Cough	Snoring	Level
1	0	P01	33	…..	3	4	Low
2	1	P10	17	…..	7	2	Medium
3	2	P100	35	…..	7	2	High
4	3	P1000	37	…..	7	5	High
…	…...	…...	…...	…...	…...	…...	…...
997	996	P996	37	…..	1	4	High
998	997	P997	25	…..	7	2	High
999	998	P998	18	…..	2	3	High
1000	999	P999	47	…..	7	2	High

b. Describe Data

Untuk memahami karakteristik awal dari data yang digunakan, dilakukan eksplorasi terhadap tipe data, distribusi nilai, serta nilai statistik deskriptif dari masing-masing atribut. Langkah ini penting untuk memperoleh gambaran umum mengenai pola dan kecenderungan data sebelum dilakukan pemodelan lebih lanjut.

Tabel 2. Deskripsi Data

No	Dataset Lung Cancer		
	Nama Fitur	Tipe Data	Keterangan
1	Index	Numerik	Nomor urut untuk setiap pasien dalam dataset.
2	Patient Id	Kategorikal	Identitas unik untuk setiap pasien (misal: P1, P2, dst.).
3	Age	Numerik	Usia pasien (dalam tahun).

No	Dataset Lung Cancer		
	Nama Fitur	**Tipe Data**	**Keterangan**
4	Gender	Kategorikal	Jenis kelamin pasien (1 = Laki-laki, 2 = Perempuan).
5	Air Pollution	Numerik	Tingkat paparan polusi udara (skala ordinal, nilai lebih tinggi menunjukkan paparan lebih besar).
6	Alcohol Use	Numerik	Frekuensi atau intensitas konsumsi alkohol (skala ordinal).
7	Dust Allergy	Numerik	Tingkat alergi debu (skala ordinal).
8	Occupational Hazards	Numerik	Risiko kerja terkait kesehatan (skala ordinal).
9	Genetic Risk	Numerik	Risiko genetik terhadap penyakit (skala ordinal).
10	Chronic Lung Disease	Numerik	Penyakit paru-paru kronis (skala ordinal).
11	Balanced Diet	Numerik	Pola makan seimbang (skala ordinal).
12	Obesity	Numerik	Indeks obesitas (skala ordinal).
13	Smoking	Numerik	Frekuensi merokok (skala ordinal).
14	Passive Smoker	Numerik	Paparan asap rokok pasif (skala ordinal).
15	Chest Pain	Numerik	Frekuensi rasa nyeri dada (skala ordinal).
16	Coughing of Blood	Numerik	Frekuensi batuk darah (skala ordinal).
17	Fatigue	Numerik	Tingkat kelelahan (skala ordinal).
18	Weight Loss	Numerik	Berat badan yang hilang (skala ordinal).
19	Shortness of Breath	Numerik	Kesulitan bernapas (skala ordinal).
20	Wheezing	Numerik	Frekuensi suara serak saat bernapas (skala ordinal).
21	Swallowing Difficulty	Numerik	Kesulitan menelan (skala ordinal).
22	Clubbing of Finger Nails	Numerik	Perubahan bentuk kuku (skala ordinal).
23	Frequent Cold	Numerik	Frekuensi pilek (skala ordinal).
24	Dry Cough	Numerik	Batuk kering (skala ordinal).
25	Snoring	Numerik	Frekuensi menggebrak saat tidur (skala ordinal).
26	Level	Kategorikal	Tingkat risiko kanker (Low, Medium, High).

c. Explore Data

Pada tahap ini dilakukan eksplorasi awal terhadap dataset untuk memahami karakteristik data, distribusi fitur, serta potensi hubungan antar variabel. Dataset yang digunakan berasal dari file *cancer patient data sets.csv* , berisi 1000 sampel dan 26 kolom , termasuk label kelas bernama Level yang menunjukkan tingkat risiko kanker (*Low, Medium, High*). Setelah kolom *index* dan *Patient Id* dihapus karena tidak relevan sebagai fitur prediktif, tersisa 24 kolom numerik yang menjadi input model *machine learning*.

Tabel 3. Eksplorasi Data

NO	Dataset Lung Cancer				
	Nama Fitur	**Mean**	**Standar Deviasi**	**Minimum**	**Maximum**
1	Age	37.174	12.005	14.0	73.0
2	Gender	1.402	0.491	1.0	2.0
3	Air Pollution	3.840	2.030	1.0	8.0
4	Alcohol Use	4.563	2.620	1.0	8.0
5	Dust Allergy	5.165	1.981	1.0	8.0
6	Occupational Hazards	4.840	2.108	1.0	8.0
7	Genetic Risk	4.580	2.127	1.0	7.0
8	Chronic Lung Disease	4.380	1.849	1.0	7.0
9	Balanced Diet	4.491	2.136	1.0	7.0
10	Obesity	4.465	2.125	1.0	7.0
11	Smoking	3.948	2.496	1.0	8.0
12	Passive Smoker	4.195	2.312	1.0	8.0
13	Chest Pain	4.438	2.280	1.0	9.0
14	Coughing of Blood	4.859	2.428	1.0	9.0
15	Fatigue	3.856	2.245	1.0	9.0
16	Weight Loss	3.855	2.207	1.0	8.0
17	Shortness of Breath	4.240	2.285	1.0	9.0
18	Wheezing	3.777	2.042	1.0	8.0
19	Swallowing Difficulty	3.746	2.270	1.0	8.0
20	Clubbing of Finger Nails	3.923	2.388	1.0	9.0
21	Frequent Cold	3.536	1.833	1.0	7.0
22	Dry Cough	3.853	2.039	1.0	7.0
23	Snoring	2.926	1.475	1.0	7.0
24	Level	~1.17	~0.73	0.0	2.0

C. Data Preparation

a. Penghapusan Atribut

Pada tahap awal, dilakukan penghapusan atribut yang tidak memiliki kontribusi signifikan terhadap proses klasifikasi. Dalam hal ini, kolom *index* dan *patient id* dihapus dari dataset karena hanya berisi informasi identitas pasien dan tidak memiliki pengaruh terhadap hasil prediksi.

Tabel 4. Perbandingan Dataset Sebelum dan Sesudah Penghapusan Atribut

NO	Dataset Lung Cancer	
	Atribut Sebelum Dihapus	**Atribut Setelah Dihapus**
1	Index	Age
2	Patient Id	Gender
3	Age	Air Pollution
4	Gender	Alcohol Use
5	Air Pollution	Dust Allergy
6	Alcohol Use	Occupational Hazards
7	Dust Allergy	Genetic Risk
8	Occupational Hazards	Chronic Lung Disease
9	Genetic Risk	Balanced Diet
10	Chronic Lung Disease	Obesity
11	Balanced Diet	Smoking
12	Obesity	Passive Smoker
13	Smoking	Chest Pain
14	Passive Smoker	Coughing of Blood
15	Chest Pain	Fatigue
16	Coughing of Blood	Weight Loss
17	Fatigue	Shortness of Breath
18	Weight Loss	Wheezing
19	Shortness of Breath	Swallowing Difficulty
20	Wheezing	Clubbing of Finger Nails
21	Swallowing Difficulty	Frequent Cold
22	Clubbing of Finger Nails	Dry Cough
23	Frequent Cold	Snoring
24	Dry Cough	Level
25	Snoring	
26	Level	

b. Encode Label

Kolom level menunjukkan Tingkat keparahan kanker paru-paru yang dikategorikan ke dalam tiga kelas, yaitu *Low, Medium* dan *High.* Oleh karena label tersebut masih dalam format *string,* maka dilakukan proses transformasi (*encode*) ke dalam format numerik dengan menggunakan *Label Encoding.*

Tabel 5. Encode Label

Encode Label	
Label Sebelum Encode	**Label Setelah Encode**
High	0
Low	1
Medium	2

c. Pemisahan Data

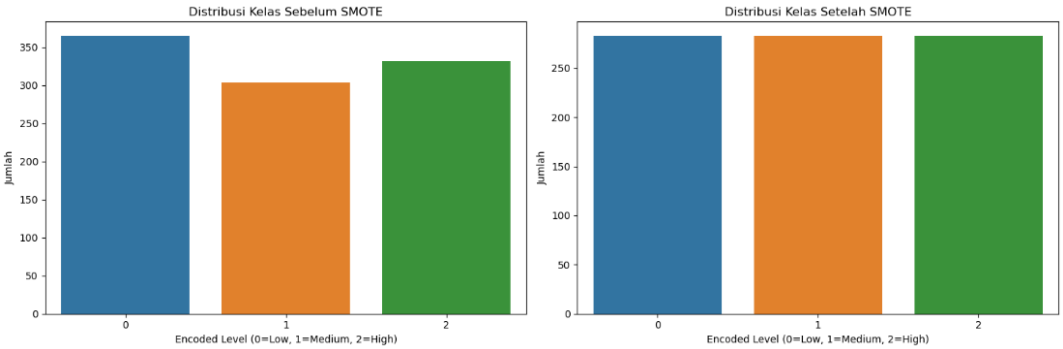
Dataset dibagi ke dalam dua subset, yaitu data latih dan data uji, menggunakan metode train_test_split untuk mempertahankan proporsi distribusi kelas yang representatif pada masing-masing subset. Hal ini penting untuk menjaga validitas evaluasi model.

Tabel 6. Pemisahan Data

Pemisahan Data		
Jenis Data	**Jumlah Data**	**Persentase**
Data Latih	800	80%
Data Uji	200	20%

d. Penyeimbangan Kelas Menggunakan SMOTE

Ketidakseimbangan kelas dalam data latih diatasi dengan menerapkan teknik *Synthetic Minority Over-sampling Technique* (SMOTE), yang secara sintetik menghasilkan data baru dari kelas minoritas guna meningkatkan representasi dan mengurangi bias prediktif terhadap kelas mayoritas. Perbandingan jumlah kelas dari sebelum dan sesudah SMOTE dapat dilihat pada gambar 2.



Gambar 2. Hasil Dari SMOTE

D. Modelling

a. Inisialisasi dan Pelatihan Model

Model *K-Nearest Neighbors* (KNN) diinisialisasi dengan parameter *n_neighbors=3*, yang berarti model akan mempertimbangkan tiga tetangga terdekat dalam menentukan kelas suatu sampel. Pemilihan nilai k=3 bertujuan untuk menjaga keseimbangan antara bias dan varians dalam proses klasifikasi. Setelah inisialisasi, model dilatih menggunakan data latih hasil *resampling* oleh metode SMOTE, yang bertujuan untuk mengatasi ketidakseimbangan kelas pada dataset kanker paru-paru. Proses pelatihan ini memungkinkan model untuk mempelajari pola hubungan antara fitur-fitur klinis dan tingkat keparahan kanker (label Level). Dengan pendekatan berbasis jarak antar titik dalam ruang fitur terstandarisasi, model KNN menghasilkan prediksi berdasarkan mayoritas kelas dari tetangga terdekat setiap sampel uji.

b. Prediksi Data Uji

Model *K-Nearest Neighbors* (KNN) yang telah dilatih kemudian digunakan untuk melakukan prediksi terhadap data uji (testing set). Proses ini bertujuan untuk mengevaluasi kemampuan model dalam mengenali pola hubungan antara fitur-fitur klinis dan tingkat keparahan kanker paru-paru pada data baru yang belum pernah dilihat sebelumnya oleh model. Output dari proses ini berupa prediksi status kesehatan pasien yang direpresentasikan dalam label Level (0 = *High*, 1 = *Low*, 2 = *Medium*), berdasarkan nilai-nilai fitur yang tersedia seperti usia, kebiasaan merokok, paparan polusi udara, dan kondisi kesehatan lainnya. Dengan pendekatan berbasis kedekatan jarak antar sampel dalam ruang fitur, model menentukan kelas setiap pasien berdasarkan mayoritas kelas dari tiga tetangga terdekatnya dalam data latih.

c. Output model

Hasil keluaran dari model *K-Nearest Neighbors* (KNN) dievaluasi menggunakan sejumlah metrik evaluasi standar untuk klasifikasi, yaitu *confusion matrix, classification report* (yang mencakup *precision, recall,* dan *f1-score*), serta accuracy score. Evaluasi ini dilakukan pada data uji yang belum pernah digunakan dalam proses pelatihan, guna memastikan objektivitas penilaian performa model.

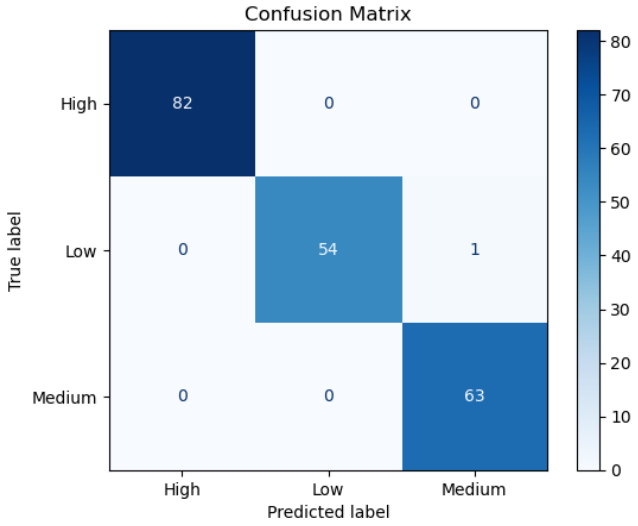
Berdasarkan hasil evaluasi, model KNN menunjukkan kinerja klasifikasi yang baik dalam membedakan tingkat keparahan kanker paru-paru (*High, Low, Medium*). Meskipun model bersifat sederhana dan berbasis jarak, hasil evaluasi menunjukkan bahwa kombinasi KNN dengan teknik *oversampling* SMOTE berhasil meningkatkan akurasi dan sensitivitas terhadap kelas minoritas, yang sebelumnya kurang terwakili dalam data asli.

E. Evaluation

Evaluasi model bertujuan untuk mengukur seberapa baik algoritma klasifikasi mampu memprediksi status kesehatan pasien. Untuk memberikan gambaran yang komprehensif mengenai kinerja model, digunakan beberapa metrik evaluasi, antara lain:

a. Confusion Matrix

Confusion matrix digunakan untuk menunjukkan jumlah prediksi benar dan salah yang dilakukan oleh model. Berdasarkan hasil evaluasi, model menghasilkan *confusion matrix* yang dapat dilihat pada gambar 3.



Gambar 3. Confusion Matrix

Evaluasi kinerja model klasifikasi dilakukan dengan menggunakan confusion matrix, yang merupakan representasi visual dari hasil prediksi model terhadap kelas yang sebenarnya. Gambar 3 menunjukkan confusion matrix untuk tiga kelas, yaitu *High, Medium*, dan *Low*.

Berdasarkan *confusion matrix* tersebut, model mampu mengklasifikasikan data secara akurat dengan tingkat kesalahan yang sangat rendah. Sebanyak 82 data yang sebenarnya termasuk dalam kelas *High* berhasil diprediksi dengan benar sebagai *High*. Demikian pula, model memprediksi dengan benar

54 dari 55 data pada kelas *Low*, dengan hanya satu data yang salah diklasifikasikan sebagai *Medium.* Sementara itu, seluruh 63 data dari kelas *Medium* berhasil diprediksi secara tepat.

Hasil ini menunjukkan bahwa model memiliki performa yang sangat baik dalam mengklasifikasikan data ke dalam masing-masing kelas. Nilai-nilai pada diagonal utama *matrix* mengindikasikan jumlah prediksi yang benar, sedangkan nilai-nilai di luar diagonal menunjukkan kesalahan klasifikasi. Kesalahan klasifikasi yang terjadi sangat minimal, yang menandakan bahwa model memiliki tingkat presisi dan akurasi yang tinggi untuk ketiga kelas tersebut.

b. Classification Report

Akurasi: 99.50%
Classification Report:

	precision	recall	f1-score	support
High	1.00	1.00	1.00	82
Low	1.00	0.98	0.99	55
Medium	0.98	1.00	0.99	63
accuracy			0.99	200
macro avg	0.99	0.99	0.99	200
weighted avg	1.00	0.99	0.99	200

Gambar 4. Classification Report

Berdasarkan gambar 4 evaluasi model klasifikasi menggunakan algoritma *K-Nearest Neighbors* (KNN) dengan pendekatan SMOTE untuk menangani ketidakseimbangan kelas, diperoleh performa yang sangat baik. Akurasi model mencapai 99.50% , menunjukkan kemampuan prediksi yang sangat tinggi. *Precision, Recall,* dan *F1-Score* untuk masing-masing kelas menunjukkan nilai yang sangat mendekati 1.00. Kelas "*High*" memiliki nilai sempurna pada ketiga metrik, sedangkan kelas "*Low*" dan "*Medium*" juga menunjukkan kinerja yang sangat baik dengan hanya sedikit penyimpangan kecil, terutama pada recall untuk kelas *"Low*" dan *precision* untuk kelas "Medium".

Secara keseluruhan, model memberikan hasil yang seimbang antara kelas mayoritas dan minoritas, membuktikan bahwa penerapan SMOTE efektif dalam meningkatkan representasi kelas minoritas tanpa mengorbankan kualitas prediksi. Nilai *macro average* dan *weighted average* yang hampir sempurna menunjukkan bahwa model tidak hanya unggul dalam mengklasifikasikan kelas dominan, tetapi juga mampu mengenali kelas dengan jumlah sampel lebih sedikit secara akurat. Hasil ini menegaskan bahwa model dapat digunakan secara efektif untuk prediksi risiko kanker berdasarkan data yang tersedia.

c. Accuracy Score

Model yang diuji mencapai tingkat akurasi sebesar 99.50% , yang menunjukkan performa prediksi yang sangat tinggi dalam mengklasifikasikan data ke dalam kelas-kelas yang ditentukan. Tingkat akurasi ini mengindikasikan bahwa model hanya menghasilkan sedikit kesalahan, jika ada, dalam proses klasifikasi. Dalam konteks aplikasi medis, di mana ketelitian dan keandalan adalah faktor kritis, tingkat akurasi sebesar 99.50% dapat dianggap luar biasa. Hal ini menunjukkan bahwa model memiliki kemampuan yang sangat baik untuk membedakan antara kelas "*High", "Low*", dan "*Medium*" dengan sangat presisi.

d. Macro dan Weighted Average

Hasil evaluasi menunjukkan bahwa model mencapai nilai *Precision , Recall* , dan *F1-Score* yang sangat tinggi untuk masing-masing kelas ("*High", "Low*", dan "*Medium*"), dengan sebagian besar metrik mendekati atau mencapai nilai sempurna (1.00). Nilai *macro average* untuk *Precision, Recall*, dan *F1-Score* tercatat sebesar 0.99, menunjukkan performa yang seimbang di seluruh kelas tanpa memperhatikan jumlah sampel pada tiap kelas. Sementara itu, nilai *weighted average* mencapai angka 1.00 untuk *Precision*, serta 0.99 untuk *Recall* dan *F1-Score*, mengindikasikan bahwa model tidak hanya unggul dalam klasifikasi secara keseluruhan, tetapi juga efektif dalam menangani ketidakseimbangan kelas. Secara keseluruhan, hasil ini membuktikan bahwa model memiliki stabilitas dan akurasi yang sangat tinggi dalam memprediksi kategori risiko pasien.

e. Perbandingan Hasil Penelitian Terdahulu

Untuk mengevaluasi efektivitas pendekatan yang digunakan dalam penelitian ini, dilakukan perbandingan dengan salah satu jurnal lain yang juga membahas deteksi kanker paru-paru menggunakan algoritma *K-Nearest Neighbor* (KNN). Perbandingan ini mencakup aspek metodologi, teknik pra-pemrosesan, akurasi model, serta penerapan teknik penyeimbangan data. Tabel berikut menyajikan ringkasan perbandingan antara penelitian ini dengan jurnal yang ditulis oleh Teguh Abdi Mangun

Tabel 7 Perbandingan Dengan Peneliti Terdahulu

Perbandingan Dengan Peneliti Terdahulu		
Aspek	**Penelitian Ini**	**Peneliti Terdahulu**
Algoritma	KNN (k=3) dengan integrasi SMOTE	KNN (tanpa SMOTE, nilai k tidak dijelaskan secara rinci)
Teknik Penyeimbangan Data	Ya, menggunakan SMOTE	Tidak digunakan
Metodologi	CRISP-DM lengkap	Alur umum
Tools	Jupyter Notebook	RapidMiner atau Excel tools
Nilai Akurasi	99.50%	80.40%
Evaluasi Model	Confusion matrix, precision, recall, F1-score, macro & weighted average	Hanya akurasi dan confusion matrix
Output Kelas	Multiclass: Low, Medium, High	Binary: High dan Low
Inovasi Tambahan	Penggunaan SMOTE, evaluasi menyeluruh, konsep implementasi sistem klinis	Tidak ada inovasi tambahan
Kelebihan	Akurasi tinggi, analisis menyeluruh, potensi integrasi sistem klinis	Implementasi sederhana, cocok untuk studi awal
Kekurangan	Perlu validasi lanjutan pada dataset berbeda	Tidak mengatasi ketidakseimbangan data, akurasi lebih rendah

Berdasarkan tabel 7, dapat disimpulkan bahwa pendekatan yang digunakan dalam penelitian ini terbukti lebih unggul dalam hal akurasi dan ketelitian klasifikasi, khususnya berkat penerapan teknik SMOTE yang mampu mengatasi ketidakseimbangan kelas. Selain itu, metodologi yang digunakan juga lebih terstruktur dengan mengikuti tahapan CRISP-DM secara lengkap. Sementara itu, jurnal pembanding memberikan pendekatan dasar yang lebih sederhana, namun tetap memberikan kontribusi sebagai acuan awal dalam penerapan algoritma KNN untuk klasifikasi kanker paru-paru. Evaluasi ini menegaskan bahwa pemilihan metode yang tepat serta pengolahan data yang baik sangat berpengaruh terhadap performa akhir model.

Selain itu, penerapan teknik oversampling seperti SMOTE terbukti sangat efektif dalam mengatasi masalah data yang tidak seimbang. Dalam penelitian lain yang dilakukan oleh Nugroho dan Harini (2024), kombinasi antara algoritma Random Forest dan SMOTE berhasil meningkatkan akurasi model hingga 97,5% serta memberikan presisi yang seimbang pada setiap kelas. Penelitian tersebut menegaskan bahwa penyeimbangan data merupakan faktor kunci dalam meningkatkan kinerja model klasifikasi pada dataset medis, terutama ketika terdapat dominasi kelas mayoritas yang signifikan [13].

F. Deployment

Tahap *deployment* merupakan implementasi akhir dari model *machine learning* ke dalam lingkungan nyata sebagai bagian dari sistem pendukung keputusan klinis. Meskipun pada penelitian ini tahap *deployment* masih bersifat konseptual, model yang telah dikembangkan memiliki potensi besar untuk diintegrasikan ke dalam sistem informasi klinik atau rumah sakit sebagai alat bantu diagnosis *non-invasif*.

Model KNN yang telah dilatih dengan data yang diseimbangkan menggunakan teknik SMOTE dapat dikemas dalam bentuk antarmuka pengguna sederhana berbasis aplikasi web atau desktop. Dalam penerapannya, tenaga medis hanya perlu memasukkan nilai-nilai fitur klinis pasien seperti usia, riwayat merokok, paparan polusi udara, serta gejala fisik lainnya. Sistem kemudian akan memberikan prediksi terhadap tingkat risiko kanker paru-paru secara *real-time*, yaitu dalam kategori "*High", "Medium*", atau *"Low"*.

Integrasi model ke dalam sistem pendukung keputusan dapat membantu tenaga medis dalam melakukan deteksi dini secara lebih cepat dan objektif, terutama di daerah dengan keterbatasan akses layanan kesehatan spesialis. Selain itu, sistem ini juga dapat digunakan sebagai alat *skrining* awal sebelum dilakukan pemeriksaan lebih lanjut seperti *rontgen* dada, *CT-scan*, atau biopsi jaringan paru-paru.

Dengan demikian, proses CRISP-DM yang telah dilakukan secara lengkap dan sistematis menjamin bahwa seluruh tahapan penelitian, mulai dari pemahaman bisnis hingga evaluasi model, dilakukan secara ilmiah dan dapat dipertanggungjawabkan. Hasil akhir penelitian ini tidak hanya menghasilkan model prediksi yang akurat, tetapi juga membuka peluang pengembangan sistem berbasis *artificial intelligence* dalam mendukung diagnosis medis secara praktis dan efisien di dunia nyata.

SIMPULAN

Penelitian ini menunjukkan bahwa kombinasi algoritma *K-Nearest Neighbors* (KNN) dan teknik penyeimbangan data SMOTE (*Synthetic Minority Oversampling Technique*) mampu meningkatkan performa model dalam melakukan klasifikasi tingkat keparahan kanker paru-paru. Dengan pendekatan metodologi CRISP-DM , seluruh tahapan analisis data dilakukan secara sistematis, mulai dari pemahaman bisnis, eksplorasi data, persiapan data, pemodelan, evaluasi, hingga konsep implementasi sistem.

Penerapan SMOTE berhasil mengatasi permasalahan ketidakseimbangan kelas pada dataset sehingga memungkinkan model memberikan prediksi yang lebih adil dan akurat untuk setiap kelas risiko ("High", "Low", dan "Medium"). Hasil evaluasi menunjukkan performa yang sangat baik, dengan akurasi sebesar 99.50% serta nilai *precision , recall* , dan F1-*score* yang mendekati sempurna pada semua kelas. Hal ini membuktikan bahwa model tidak hanya mampu mengklasifikasikan kelas mayoritas dengan baik, tetapi juga efektif dalam mengenali kelas minoritas.

Meskipun hasil yang diperoleh sangat memuaskan, perlu dilakukan validasi lebih lanjut untuk memastikan kemampuan generalisasi model terhadap data baru. Disarankan untuk melakukan pengujian menggunakan dataset lain, validasi silang (*cross-validation*), serta membandingkan model ini dengan algoritma lain guna mengevaluasi stabilitas dan robustness sistem prediksi.

Dengan demikian, integrasi antara algoritma KNN dan teknik SMOTE dapat dijadikan sebagai pendekatan yang andal dalam membangun sistem deteksi dini kanker paru-paru berbasis data klinis. Penelitian ini diharapkan menjadi fondasi bagi pengembangan sistem pendukung keputusan medis yang akurat, efisien, dan memiliki aplikasi nyata dalam dunia kedokteran.

REFERENSI

[1] S. A. Naufal, A. Adiwijaya, and W. Astuti, "**Analisis Perbandingan Klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) untuk Deteksi Kanker dengan Data Microarray**" *JURIKOM (Jurnal Riset Komputer)*, vol. 7, no. 1, p. 162, Feb. 2020, doi: 10.30865/jurikom.v7i1.2014.

[2] A. Reynaldi, Y. Trisyani, and D. Adiningsih, "**KUALITAS HIDUP PASIEN KANKER PARU STADIUM LANJUT**" Jun. 2020.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "**Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries**" *CA Cancer J Clin*, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[4] T. Abdi Mangun, O. Nurdiawan, and A. Irma Purnamasari, "**LUNG CANCER ANALYSIS USING K-NEARST NEIGHBOR ALGORITHM**" 2023. [Online]. Available: https://ejournal.ubibanyuwangi.ac.id/index.php/jurnal_tinsika

[5] J. Han, M. Kamber, and J. Pei, "**Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)**" 2017.

[6] P. Gulande and R. Awale, "**A Hybrid mRMR-RSA Feature Selection Approach for Lung Cancer Diagnosis Using Gene Expression Data**" *Biomedical and Pharmacology Journal*, vol. 18, pp. 257–270, Mar. 2025, doi: 10.13005/bpj/3086.

[7] K. K. V, S. Balaji B, S. K R, and A. Najat Ahmed, **"Enhanced Lung Cancer Prediction Using Ensemble Machine Learning Algorithms**" in *2024 International Conference on Emerging Research in Computational Science (ICERCS)*, Dec. 2024, pp. 1–5. doi: 10.1109/ICERCS63125.2024.10894971.

[8] R. Ullah, K. Parveen, I. Rehan, and S. Khan, "**Enhancing lung cancer diagnostics through Raman spectroscopy and machine learning**" *Phys Scr*, vol. 100, no. 4, p. 046015, 2025, doi: 10.1088/1402-4896/adc214.

[9] Y. Lin *et al.*, "**A fast, non-invasive auxiliary screening algorithm for lung cancer based on electronic nose system**" *Sens Actuators A Phys*, vol. 389, p. 116490, 2025, doi: https://doi.org/10.1016/j.sna.2025.116490.

[10] A. Rifa'i and Y. Prabowo, "**Diagnosis Kanker Paru-Paru dengan Sistem Fuzzy**" vol. 10, no. 1, pp. 19–28, 2022, doi: 10.32832/kreatif.v10i1.6317.

[11] M. Yunianto, F. Anwar, D. Nur Septianingsih, T. Dwi Ardyanto, and R. Farits Pradana, "**KLASIFIKASI KANKER PARU PARU MENGGUNAKAN NAÏVE BAYES DENGAN VARIASI FILTER DAN EKSTRAKSI CIRI GRAY LEVEL CO-OCCURANCE MATRIX (GLCM)**" *Indonesian Journal of Applied Physics*, vol. 11, no. 2, 2021.

[12] S. Sucipto, D. Dwi Prasetya, and T. Widiyaningtyas, "**Educational Data Mining: Multiple Choice Question Classification in Vocational School**" *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 2, pp. 379–388, Mar. 2024, doi: 10.30812/matrik.v23i2.3499.

[13] A. Nugroho and D. Harini, "**Teknik Random Forest untuk Meningkatan Akurasi Data Tidak Seimbang**" *JSITIK*, vol. 2, no. 2, 2024, doi: 10.53624/jsitik.v2i2.XX.