

Implementasi *Machine Learning* untuk Klasifikasi Penyakit Kanker Paru Menggunakan Metode *Naïve Bayes* dengan Tambahan Fitur *Chatbot*

(Implementation of *Machine Learning* for Lung Cancer Classification Using *Naïve Bayes* Method with Additional *Chatbot* Features)

Dhea Juliani*, Muhamad Soleh

Program Studi Teknik Informatika, Institut Teknologi Indonesia
Jl. Raya Puspiptek, Serpong, Kota Tangerang Selatan, Banten, 15320

Abstrak

Menurut *The Global Burden of Cancer* diperkirakan kasus baru kanker meningkat menjadi 19,3 juta pada tahun 2020, dengan sekitar 10 juta kematian. Kanker paru merupakan jenis kanker terbanyak di Indonesia, dengan sekitar 34 ribu kasus baru dan 30 ribu kematian pada tahun 2020. Salah satu bentuk kemajuan teknologi informasi adalah sistem deteksi penyakit kanker paru dan sistem *chatbot*, sistem *chatbot* digunakan untuk mempermudah pengguna dalam berkomunikasi dan mendapatkan informasi mengenai penyakit kanker paru. Tahapan yang akan dilakukan adalah pengumpulan *dataset*, *text preprocessing*, perancangan tampilan antarmuka sistem deteksi dan *chatbot*, dan klasifikasi *naïve bayes*. *Dataset* penyakit kanker paru diperoleh dari *website Kaggle* dengan jumlah 309 data dan 16 fitur diantaranya *Gender*, *Age*, *Smoking*, *Yellow Fingers*, *Anxiety*, *Peer Pressure*, *Chronic Disease*, *Fatigue*, *Allergy*, *Wheezing*, *Alcohol Consuming*, *Coughing*, *Shortness of Breath*, *Swallowing Difficulty*, dan *Chest Pain*. *Dataset* pertanyaan untuk *chatbot* merupakan data yang dimasukkan secara manual dan disimpan dalam format JSON dan untuk pemodelan pada sistem ini menggunakan *Artificial Neural Network* (ANN) dengan pendekatan *Natural Language Processing* (NLP). Hasil klasifikasi *naïve bayes* memperoleh tingkat akurasi tertinggi yang dicapai dalam penelitian ini sebesar 81%, tetapi untuk sistem *chatbot* dan sistem deteksi penyakit kanker paru masih belum terintegrasi.

Kata Kunci : Kanker paru, *Machine Learning*, *Chatbot*, *Naïve Bayes Classifier*.

Abstract

According to *The Global Burden of Cancer*, it is estimated that new cases of cancer will increase to 19.3 million by 2020, with around 10 million deaths. Lung cancer is the most common type of cancer in Indonesia, with around 34 thousand new cases and 30 thousand deaths in 2020. One form of information technology advancement is a lung cancer disease detection system and chatbot system, the chatbot system is used to facilitate users in communicating and getting information about lung cancer. The stages that will be carried out are dataset collection, text preprocessing, interface design of the detection system and chatbot, and naïve bayes classification. The lung cancer dataset is obtained from the Kaggle website with 309 data and 16 features including Gender, Age, Smoking, Yellow Fingers, Anxiety, Peer Pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness of Breath, Swallowing Difficulty, and Chest Pain. The question dataset for the chatbot is data that is manually entered and stored in JSON format and for modeling in this system using an Artificial Neural Network (ANN) with a Natural Language Processing (NLP) approach. The naïve bayes classification results obtained the highest accuracy rate achieved in this study of 81%, but the chatbot system and the lung cancer detection system are still not integrated.

Keyword : Lung Cancer, *Machine Learning*, *Chatbot*, *Naïve Bayes Classifier*.

*Penulis Korespondensi. Telp: +62 822 9521 9342
Alamat E-mail : dheajuli4171@gmail.com (Dhea Juliani)

1. Pendahuluan

Menurut The Global Burden of Cancer diperkirakan jumlah kasus baru kanker meningkat menjadi 19,3 juta dengan sekitar 10 juta kematian pada tahun 2020. Kanker paru-paru menduduki urutan pertama dengan 1.7 juta kematian, diikuti oleh kanker kolorektal dengan 935 ribu kematian, dan kanker prostat dengan 375 ribu kematian. Kanker paru merupakan jenis kanker terbanyak di Indonesia, dengan sekitar 34 ribu kasus baru dan 30 ribu kematian pada tahun 2020. [1]

Berdasarkan data dari WHO, kanker ini menempati urutan pertama di Indonesia sebagai jenis kanker yang paling banyak menyerang laki-laki dan jenis kanker urutan kelima yang paling banyak menyerang perempuan. Kebiasaan merokok merupakan penyebab utama terjadinya kanker paru, baik itu pada perokok aktif maupun perokok pasif. Hal inilah yang menjadi alasan mengapa banyak laki-laki terkena kanker paru-paru. Terdapat berbagai pemicu lain kanker paru, mulai dari tercemarnya lingkungan pasien akibat polusi udara dan unsur kimia berbahaya lainnya, atau karena keluarga pasien yang pernah atau sedang mengidap kanker atau penyakit paru-paru lainnya [2].

Teknologi telah mengalami berbagai perkembangan yang pesat seiring dengan perkembangan zaman. Salah satu bentuk teknologi yang berkembang pesat di era modern adalah kecerdasan buatan [3]. Salah satu bentuk kemajuan teknologi informasi adalah aplikasi chatbot. Chatbot adalah program komputer yang dirancang untuk melakukan percakapan atau komunikasi interaktif dengan pengguna (user) melalui teks, ucapan, atau visual [4]. Sehingga dibutuhkan sebuah sistem seperti chatting bot atau chatbot yang dapat mempermudah pengguna dalam berkomunikasi dan mendapatkan informasi mengenai penyakit kanker paru.

Pada penelitian ini akan mengklasifikasikan kanker paru berdasarkan data yang diperoleh dan mencari keakuratan data. Metode klasifikasi yang digunakan dalam penelitian ini adalah *Naïve Bayes*. Metode *Naïve Bayes* dipilih karena menghasilkan akurasi tertinggi dengan data pelatihan yang lebih sedikit [5].

Berdasarkan uraian permasalahan di atas, penelitian ini akan membuat sistem klasifikasi penyakit kanker paru dengan tambahan fitur chatbot untuk melakukan percakapan mengenai informasi dari penyakit tersebut. Kemudian gejala atau keluhan yang dirasakan oleh user akan

diklasifikasikan menggunakan salah satu algoritma *Machine Learning* yaitu *Naïve Bayes*.

2. Teori Dasar

2.1. *Machine Learning*

Secara definisi, *machine learning* adalah cabang kecerdasan buatan yang berfokus pada pengembangan serta mempelajari suatu sistem data yang diperolehnya. Menurut Arthur Samuel, *machine learning* adalah bidang studi yang memberikan kemampuan kepada program komputer untuk belajar tanpa diprogram secara eksplisit. Untuk mengimplementasikan teknik pembelajaran mesin, diperlukan data. Tanpa adanya data, algoritma pembelajaran mesin tidak dapat bekerja [6].

2.2. *Naïve Bayes*

Metode klasifikasi ini menggunakan teknik probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang masa depan berdasarkan pengalaman di masa sebelumnya, sehingga dikenal sebagai Teorema Bayes. Ciri utama dari *Naïve Bayes Classifier* adalah membuat asumsi independensi yang sangat kuat untuk masing-masing kondisi atau kejadian [7].

2.3. *Chatbot*

Chatbot adalah sebuah program komputer yang dirancang untuk berinteraksi atau berkomunikasi dengan pengguna (*user*) baik melalui teks, suara ataupun visual secara interaktif. (Hormansyah dan Utama, 2018:225).

2.4. Penyakit Kanker Paru

Kanker paru-paru merupakan penyakit ganas pada organ paru-paru yang disebabkan oleh perubahan genetika pada sel epitel saluran pernafasan yang menyebabkan perkembangbiakan sel yang tidak terkendali. Penyakit ganas ini dapat berasal dari organ paru itu sendiri (*primer*) maupun di luar paru-paru (*metastasis*). Secara klinis, kanker paru-paru *primer* diartikan sebagai tumor ganas yang muncul di epitel bronkus [8].

3. Metodologi

3.1. *Dataset*

Dataset pada penelitian ini didapat dari situs *website Kaggle* (<https://www.kaggle.com/datasets/mysarahmadb/hat/lung-cancer>). *Dataset* mengenai penyakit kanker paru-paru ini memiliki jumlah data sebanyak 309 data dalam format CSV (*Comma Separated Values*) berupa angka. Dari 309 data hanya 16 data yang ditampilkan sebagai sampel.

Dataset di atas terdiri dari 16 fitur. Berikut adalah daftar fitur pada dataset tersebut:

1. *Gender*: Jenis kelamin, M: Laki-laki, F: Perempuan;
2. *Age*: Usia dalam tahun;
3. *Smoking*: Merokok;
4. *Yellow fingers*: Perubahan warna pada jari menjadi warna kuning;
5. *Anxiety*: Kecemasan;
6. *Peer Pressure*: Tekanan sosial;
7. *Chronic Disease*: Penyakit kronis;
8. *Fatigue*: Kelelahan;
9. *Allergy*: Alergi;
10. *Wheezing*: Mengi/serak;
11. *Alcohol Consuming*: Mengonsumsi alkohol;
12. *Coughing*: Batuk;
13. *Shortness of Breath*: Sesak nafas;
14. *Swallowing Difficulty*: Kesulitan menelan;
15. *Chest Pain*: Nyeri dada;
16. *Lung Cancer*: Fitur prediksi YES/NO.

3.2. Klasifikasi Naïve Bayes

Data yang digunakan sebanyak 16 data *training*, untuk keterangan angka 1 yaitu *NO* dan angka 2 yaitu *YES*. Diketahui label *Lung Cancer* “*NO*” sebanyak 6 data dan label *Lung Cancer* “*YES*” sebanyak 10 data, maka total data = 16 data.

GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC_DISEASE	FATIGUE
M	69	1	2	2	1	1	2
M	74	2	1	1	1	2	2
F	59	1	1	2	2	1	2
M	63	2	2	2	1	1	1
F	63	1	2	1	1	1	1
F	75	1	2	1	1	2	2
M	52	2	1	1	1	1	2
F	51	2	2	2	2	1	2
F	68	2	1	2	1	1	2
M	53	2	2	2	2	2	1
F	61	2	2	2	2	2	2
M	72	1	1	1	1	2	2
F	60	2	1	1	1	1	2
M	58	2	1	1	1	1	2
M	69	2	1	1	1	1	1
F	48	1	2	2	2	2	2

Gambar 1. Tabel Data Training

ALLERGY	WHEEZING	ALCOHOL_CONSUMING	COUGHING	SHORTNESS_OF_BREATH	SWALLOWING_DIFFICULTY	CHEST_PAIN	LUNG_CANCER
1	2	2	2	2	2	2	YES
2	1	1	1	2	2	2	YES
1	2	1	2	2	2	2	NO
1	1	2	1	1	2	2	NO
1	2	1	2	2	1	1	NO
2	2	1	2	2	1	2	YES
1	2	2	2	2	2	2	YES
2	1	1	1	1	2	1	YES
1	1	1	1	1	1	1	NO
2	1	2	1	1	2	2	YES
1	2	1	2	2	2	1	YES
2	2	2	2	2	1	2	YES
1	1	1	1	2	1	1	NO
2	2	2	2	2	1	2	YES
2	2	2	2	1	1	2	NO
2	2	1	2	2	2	1	YES

Gambar 2. Tabel Data Training

Rumus Naïve Bayes :

$$P(H|X) = \frac{(P(X|H) \cdot P(H))}{(P(X))}$$

Keterangan :

P(H|X) : Probabilitas hipotesis berdasarkan kondisi X (probabilitas *posterior*);

P(X|H) : Probabilitas berdasarkan kondisi pada hipotesis;

P(H) : Probabilitas *prior*;

P(X) : Probabilitas X;

X : Fitur data yang dimasukkan;

H : Hipotesis data X.

Perhitungan probabilitas *prior* (P(H)) untuk kelas positif adalah:

$$P(YES) = \frac{10}{16} = 0,625$$

Sedangkan untuk perhitungan probabilitas *prior* (P(H)) untuk kelas negatif adalah:

$$P(NO) = \frac{6}{16} = 0,375$$

Contoh perhitungan probabilitas *posterior* dilakukan pada setiap fitur sesuai data pada sebagai berikut.

Tabel 1. Probabilitas Fitur Jenis Kelamin (Gender)

X	Probabilitas Gender		P(X Yes)	P(X No)
	Yes	No		
Laki-Laki	6	2	6/10 = 0.6	2/6 = 0.333333
Perempuan	4	4	4/10 = 0.4	2/6 = 0.666667
Jumlah	10	6	1	1

GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC_DISEASE	FATIGUE
M	69	1	2	2	1	1	2
M	74	2	1	1	1	2	2
F	59	1	1	2	2	1	2
M	63	2	2	2	1	1	1
F	63	1	2	1	1	1	1
F	75	1	2	1	1	2	2
M	52	2	1	1	1	1	2
F	51	2	2	2	2	1	2
F	68	2	1	2	1	1	2
M	53	2	2	2	2	2	1
F	61	2	2	2	2	2	2
M	72	1	1	1	1	2	2
F	60	2	1	1	1	1	2
M	58	2	1	1	1	1	2
M	69	2	1	1	1	1	1
F	48	1	2	2	2	2	2

Gambar 3. Tabel Data Testing

Menghitung nilai *Likelihood* dari hasil probabilitas setiap fitur. Nilai *Likelihood* dibagi menjadi 2 yaitu positif dan negatif.

Likelihood positif:

$$P(X|Positif) = 0.6 * 0.03194322 * 0.6 * 0.6 * 0.5 * 0.6 * 0.4 * 0.1 * 0.7 * 0.7 * 0.6 * 0.3 * 0.1 * 0.6 * 0.4 = 1.75264322E - 7 (0.0000001752)$$

Likelihood negatif:

$$P(X|Negatif) = 0.333333 * 0.0107738385 * 0.333333 * 0.333333 * 0.333333 * 0.8333333333 * 1 * 0.5 * 0.1666666667 * 0.5 * 0.5 * 0.5 * 0.5 * 0.1666666667 * 0.5 = 4.81082971E - 8 (0.0000000481)$$

Perhitungan untuk klasifikasi ke dalam kelas positif dan negatif dengan mengalikan hasil dengan rumus *naïve bayes* : $P(H|X) = P(X|H) * P(H)$

$$P(X|Positif)P(YES) = 0.0000001752 * 0.625 = 1.09500000E-7 (0.0000001095)$$

$$P(X|Negatif)P(NO) = 0.0000000481 * 0.375 = 1.8037500E-8 (0.0000000180)$$

Berdasarkan hasil perhitungan di atas dapat dibuktikan bahwa $P(X|Positif)P(Yes) > P(X|Negatif)P(No)$, maka untuk data *testing* pada



Gambar 6. Input Nilai

3. Halaman Hasil Prediksi



Gambar 7. Hasil Prediksi

4. Hasil Akurasi

Hasil pengujian dapat ditampilkan oleh sistem pada Gambar 8, dengan hasil akurasi 81%.

	precision	recall	f1-score	support
NO	0.50	0.50	0.50	6
YES	0.88	0.88	0.88	25
accuracy			0.81	31
macro avg	0.69	0.69	0.69	31
weighted avg	0.81	0.81	0.81	31

Gambar 8. Hasil Akurasi

4.2. Pemodelan Chatbot

1. Dataset Pertanyaan

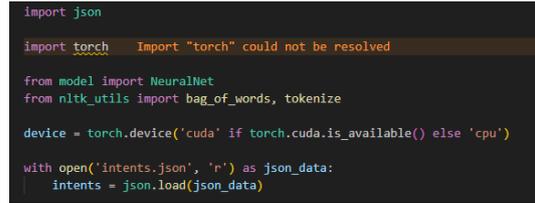
Dataset yang digunakan dalam penelitian ini merupakan dataset yang dimasukkan secara manual dan disimpan dalam sebuah file dengan format JSON (*JavaScript Object Notation*). Dataset memiliki struktur diantaranya *intens*, *tags*, *patterns*, dan *responses*. Pada Gambar 9 berikut ditampilkan potongan 7 tags/kategori pertama pada dataset dalam format JSON.



Gambar 9. Potongan Dataset

2. Import Dataset

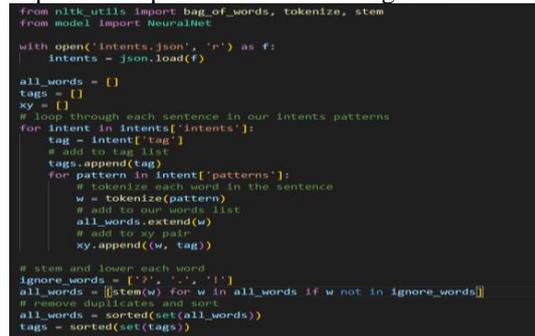
Sebelum melakukan tahapan *text preprocessing*, dataset harus diimport terlebih dahulu. Proses mengimport dataset dapat dilihat pada Gambar 10 berikut.



Gambar 10. Import Dataset

3. Text Preprocessing

Proses *text preprocessing* dibutuhkan library *Natural Language Toolkit* (NLTK) yang merupakan penerapan dari *Natural Language Processing* (NLP). Proses *text preprocessing* dapat dilihat pada Gambar 11 sebagai berikut.



Gambar 11. Proses Text Preprocessing

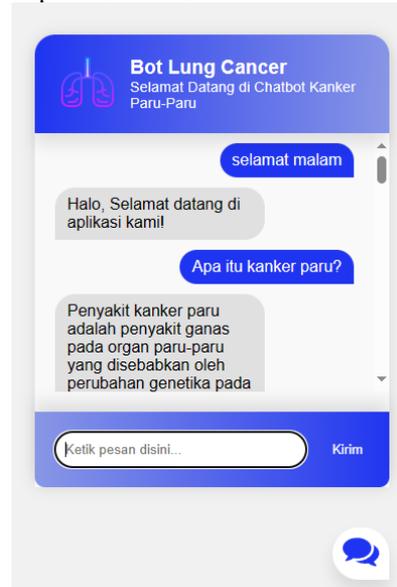
4. Hasil Text Preprocessing

Hasil dari proses *text preprocessing* didapatkan sebanyak 36 patterns, 12 tags, dan 59 token atau kata unik. Proses ini dapat dilihat pada Gambar 12 sebagai berikut.



Gambar 12. Hasil Text Preprocessing

5. Tampilan Sistem Chatbot



Gambar 13. Tampilan Sistem Chatbot

5. Kesimpulan

Berdasarkan hasil yang didapat dari keseluruhan rangkaian penelitian ini, didapatkan beberapa kesimpulan mengenai implementasi *naïve bayes* dan implementasi sistem *chatbot* ini diantaranya:

1. Klasifikasi penyakit kanker paru dengan menggunakan *naïve bayes* mendapatkan hasil akurasi yang diperoleh sebesar 81%.
2. *Chatbot* mampu memberikan respon yang sesuai. Tetapi memiliki kelemahan yaitu *user* hanya bisa menjawab dan *chatbot* hanya bisa merespon sesuai dengan yang ada pada *dataset intents.json*.
3. Sistem *chatbot* dan sistem deteksi penyakit kanker paru belum terintegrasi.

Daftar Pustaka

- [1] Sari, L., Romadloni, A., & Listyaningrum, R. (2023). Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest. *Infotekmesin*, 14(1), 155–162. <https://doi.org/10.35970/infotekmesin.v14i1.1751>
- [2] Sinaga, R. B., Widiyanto, D., Wahyono, B. T., Kunci, K., Paru, K., & Forest, R. (2022). *Deteksi Dini Penyakit Kanker Paru dengan Gabungan Algoritma Adaboost dan Random Forest*. 682–691.
- [3] Prasetyo, V. R., Benarkah, N., & Chrisintha, V. J. (2021). Implementasi Natural Language Processing Dalam Pembuatan Chatbot Pada Program Information Technology Universitas Surabaya. *Teknika*, 10(2), 114–121. <https://doi.org/10.34148/teknika.v10i2.370>
- [4] Harahap, D. W., & Fitria, L. (2020). Aplikasi Chatbot Berbasis Web Menggunakan Metode Dialogflow. *J-ICOM - Jurnal Informatika Dan Teknologi Komputer*, 1(1), 6–13. <https://doi.org/10.33059/j-icom.v1i1.2796>
- [5] Wulandari, E. (2022). Klasifikasi Kanker Paru-Paru Menggunakan Metode Naive Bayes. *International Research on Big-Data and Computer Technology: I-Robot*, 6(2), 20–24. <https://doi.org/10.53514/ir.v6i2.325>
- [6] Fikriya, Z. A., Irawan, M. I., & Soetrisno, S. (2017). Implementasi Extreme Learning Machine untuk Pengenalan Objek Citra Digital. *Jurnal Sains Dan Seni ITS*, 6(1). <https://doi.org/10.12962/j23373520.v6i1.21754>
- [7] Sihombing, J. (2021). Klasifikasi Data Antropometri Individu Menggunakan Algoritma Naïve Bayes Classifier. *BIOS : Jurnal Teknologi Informasi Dan Rekayasa Komputer*, 2(1), 1–10. <https://doi.org/10.37148/bios.v2i1.15>
- [8] Buana, I., & Harahap, D. A. (2022). Asbestos, Radon Dan Polusi Udara Sebagai Faktor Resiko Kanker Paru Pada Perempuan Bukan Perokok. *AVERROUS: Jurnal Kedokteran Dan Kesehatan Malikussaleh*, 8(1), 1. <https://doi.org/10.29103/averrous.v8i1.7088>