

Machine Learning untuk Identifikasi Jenis Kanker Darah (*Leukemia*)

Machine Learning for Blood Cancer (*Leukemia*) Type Identification

Abdul Mahatir Najar^{1*}, I Wayan Sudarsana², M Ulul Albab³, Sultan Andhika¹

¹Laboratorium Ilmu Komputer, Jurusan Matematika, Universitas Tadulako

²Laboratorium Aljabar dan Kombinatorika, Jurusan Matematika, Universitas Tadulako

³Pendidikan Matematika, FKIP, Universitas Islam Lamongan

*Email Korespondensi: abdulmahatir@untad.ac.id

Info Artikel

Diterima : 03 Jan 21
Direvisi : 21 Jan 21
Diterbitkan : 07 Feb 22

Kata Kunci:

Kanker Darah, *Leukemia*,
Machine Learning, DNA, RNA

Cara merujuk artikel ini:

Najar, A. M., dkk. (2022).
Machine Learning untuk
Identifikasi Jenis Kanker
Darah (*Leukimia*). *Vygotsky:
Jurnal Pendidikan Matematika
dan Matematika*, 4 (1), 47-56.
Diunduh dari
[https://jurnalpendidikan.
unisla.ac.id/index.php/
VoJ/article/view/493](https://jurnalpendidikan.unisla.ac.id/index.php/VoJ/article/view/493)

Abstract

Rapid and precise method for distinguishing the types of blood cancer is very crucial for patient treatment. In this study, the blood cancer type identification is carried out by utilizing machine learning. The process of identifying begins with performing feature extraction. The feature extraction process is carried out by using the Markov Chain method. From this process, it will form a matrix which is then used as training and testing data on several machine learning algorithms. The results show that that the accuracy of the Decision Tree Classification algorithm gave the best results, namely 83%, followed by the KNN method and 50%, while the SVM method only achieved 37.5% accuracy.

Abstrak

Metode yang cepat dan tepat untuk membedakan jenis kanker darah sangat penting agar pasien kanker mendapatkan perlakuan yang sesuai. Pada penelitian ini identifikasi jenis kanker darah dilakukan dengan memanfaatkan kecerdasan buatan khususnya *machine learning*. Proses identifikasi kanker *leukemia* menggunakan *machine learning* dimulai dengan melakukan ekstraksi ciri. Proses ekstraksi ciri dilakukan dengan memanfaatkan metode Rantai Markov. Dari proses ini akan membentuk matriks yang kemudian dijadikan data *training* dan *testing* pada beberapa algoritma *machine learning*. Berdasarkan hasil *training* dan *testing* diperoleh hasil bahwa akurasi algoritma *Decision Tree Classification* memberikan hasil terbaik yaitu 83%, disusul dengan metode KNN dan sebesar 50%, sedangkan metode SVM hanya mencapai akurasi 37.5%.

PENDAHULUAN

Kanker adalah salah satu masalah kesehatan yang menjadi momok bagi hampir seluruh negara termasuk Indonesia. Setiap tahunnya jumlah penderita kanker terus meningkat, bahkan peningkatannya dapat mencapai 70% dalam kurun waktu 20 tahun (Dewi, 2017). Salah satu jenis kanker yang unik dan perlu mendapatkan perlakuan khusus adalah kanker darah yang kemudian lebih sering disebut dengan nama *leukemia*. *Leukemia* ini menjadi unik karena bentuknya yang cair sehingga sulit diidentifikasi secara fisik. *Leukemia* terjadi akibat adanya kelainan pada sumsum tulang belakang, yaitu kelainan pada sel-sel pembentuk darah muda. (Morrison & Hesdorffer, 2012).

Penyebab utama *leukemia* belum diketahui dengan pasti. Namun, secara umum kanker dimulai dengan mutasi (perubahan) bahan genetik. Bahan genetik dapat dikatakan sebagai warisan atau sesuatu yang diturunkan dari orang tua. Seiring dengan berjalannya waktu dan terjadinya pertambahan usia, sel-sel warisan ini terus membelah diri dan membentuk sel-sel baru. Selama proses pembelahan ini, sel-sel *leukemia* dapat terbentuk akibat adanya kesalahan dalam proses replikasi (Pratiwi, 2014). Terdapat 4 jenis leukemia yang dikenal yaitu *Chronic Lymphoblastic Leukemia* (CLL), *Acute Myeloblastic Leukemia* (AML), *Chronic Myeloblastic Leukemia* (CML), *Acute Lymphoblastic Leukemia* (ALL). Berdasarkan tingkat keparahannya keempat jenis *leukemia* ini dapat dikelompokkan menjadi 2 kelompok yaitu *leukemia* akut dan *leukemia* kronis (Morrison & Hesdorffer, 2012).

Identifikasi dan klasifikasi jenis kanker darah yang cepat dan tepat sangat penting agar pasien kanker mendapatkan perlakuan yang sesuai (American Cancer Society, 2019). Metode yang umum digunakan untuk identifikasi kanker adalah menggunakan biomarker. Namun metode ini memiliki kelemahan, yaitu memerlukan biaya yang cukup besar serta waktu yang diperlukan untuk memperoleh hasilnya cukup lama (Al Faroby, 2018). Sebagai alternatif banyak dikembangkan metode-metode untuk identifikasi kanker dengan memanfaatkan kecerdasan buatan.

Secara sederhana kecerdasan buatan (*artificial intelligence*) dapat didefinisikan sebagai sebuah metode untuk meniru kecerdasan yang dimiliki oleh manusia yang disimpan dan diajarkan pada mesin agar dapat berpikir seperti manusia. Pembelajaran mesin (*machine learning*) merupakan salah satu kecerdasan buatan yang saat ini banyak dikembangkan. Bidang ini berfokus untuk mengembangkan kemampuan mesin agar bisa belajar dan menganalisa data secara mandiri tanpa perlu diprogram ulang atau diperintah. Pengembangan metode-metode pembelajaran mesin melibatkan berbagai disiplin ilmu seperti matematika, ilmu komputer dan statistika (Helm et al., 2020).

Penerapan *machine learning* pada bidang kesehatan terus meningkat dari segi kuantitas dan cakupannya. Pada bidang dermatologi, *machine learning* digunakan untuk melakukan deteksi penyakit kulit berdasarkan citra (Kumar et al., 2016). *Machine learning* juga telah dimanfaatkan untuk

menentukan signifikansi kasus *pediatric acute lymphoblastic leukemia* pada anak dengan atribut klinis dan fenotip serta kondisi lingkungan (Mahmood et al., 2020), membangun model prediksi kekambuhan pasca transplantasi dan menganalisis data kekambuhan *leukemia* pasca *Hematopoietic Stem Cell Transplant* dari pasien anak ALL (Shyr et al., 2020), melakukan deteksi *retinopati diabetik* pada citra retina (Topol, 2019), mengklasifikasi empat jenis umum leukemia melalui segmentasi citra *blood smear* yang *robust* (Dese et al., 2021).

Penerapan metode *machine learning* untuk mengklasifikasikan atau mendeteksi penyakit berbasis citra secara umum memberikan hasil yang menajikan jika objeknya memiliki ciri fisik yang jelas. Namun pada kasus kanker darah, bentuknya sulit terlihat secara fisik sehingga diperlukan pendekatan lain. Pada penelitian ini, *machine learning* akan dimanfaatkan untuk mengembangkan sistem klasifikasi kanker darah berdasarkan ciri sekuens DNA/RNA. Hasil penelitian ini diharapkan dapat berkontribusi pada pengembangan metode deteksi dini *leukemia*.

METODE

Dalam penelitian data yang digunakan merupakan sekuens DNA/RNA yang diperoleh dari National Center of Biotechnology Information (NCBI), masing-masing 10 data untuk setiap jenis *leukimia*. Data-data yang digunakan ditampilkan pada **Tabel 1**.

Tabel 1. Kode Sekuens DNA/RNA *leukimia*

No	Kode Sekuens			
	ALL	AML	CML	CLL
1	M19730.1	AF135383.1	S82592.1	S73961.1
2	Y00811.1	AF135385.1	M25946.1	S73959.1
3	X07135.1	AF135384.1	HV197178.1	U22391.1
4	BC126373.1	L34598.1	BD176607.1	U22390.1
5	BC126375.1	S82690.1	BD176606.1	S64030.1
6	BC069422.1	S72604.1	MA403370.1	CA773229.1
7	BI480895.1	R73338.1	MA403367.1	S72664.1
8	XM_043355361.1	BM128504.1	MA403366.1	AY351432.1
9	D29808.1	EF413001.1	MA403364.1	CA773667.1
10	FM165199.1	BM150344.1	MA403363.1	BF432339.1

Sebelum data diolah menjadi sebuah informasi yang diinginkan, data sekuens perlu dibersihkan terlebih dahulu.

Human acute lymphocytic leukemia (ALL) BCR-ABL mRNA encoding P185-ALL-ABL protein

GenBank: M19730.1

[GenBank](#) [Graphics](#)

```
>M19730.1 Human acute lymphocytic leukemia (ALL) BCR-ABL mRNA encoding P185-ALL-ABL protein
GGAAGGGGAGGGCAAGGGCCCGCTCCTGCGCACGCAGAGCACCTCTGAGCAGGAGAAGCGCCTTACCTGG
CCCCGCAGGTCTACTCCCCCGGAGTTTGGAGATTGCGGAGGCGGTATAACCCGGACTGCAGCTCCA
ATGAGAACCCTCACCTCCAGCGAGGAGGACTTCTCCTCTGGCCAGTCCAGCCGCGTGTCCCAAGCCCCAC
CACCTACCGCATGTTCCGGGACAAAAGCCGCTCTCCCTCGCAGAACTCGCAACAGTCTTCGACAGCAGC
AGTCCCCCACGCCGAGTGCCATAAGCGGCCACCGGCACTGCCGGTTGTCGTGTCCGAGGCCACCATCG
TGGGCGTCCGCAAGACCGGGCAGATCTGGCCCAACGATGGCGAGGGCGCTTCCATGGAGACGCAGAAGC
CCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCTCAGGGTCTGAGTGAAGCCGCTCGTTGGAATCCAAG
G
```

Gambar 1. Sekuens M19730.1 Sebelum Dibersihkan

Data sekuens pada **Gambar 1** masih berformat fasta dan harus dibersihkan dari *title* data dan tanpa spasi di antara basa nitrogen sehingga hanya tinggal data sekuens berupa *text* seperti **Gambar 2** agar dapat dibaca dan diproses ke tahap selanjutnya yaitu ekstraksi ciri.



```
all1.txt x
GGAAGGGGAGGGCAAGGGCCCGCTCCTGCGCACGCAGAGCACCTCTGAGCAGGAGAAGCGCCTTACCTGG
CCCCGCAGGTCTACTCCCCCGGAGTTTGGAGATTGCGGAGGCGGTATAACCCGGACTGCAGCTCCA
ATGAGAACCCTCACCTCCAGCGAGGAGGACTTCTCCTCTGGCCAGTCCAGCCGCGTGTCCCAAGCCCCAC
CACCTACCGCATGTTCCGGGACAAAAGCCGCTCTCCCTCGCAGAACTCGCAACAGTCTTCGACAGCAGC
AGTCCCCCACGCCGAGTGCCATAAGCGGCCACCGGCACTGCCGGTTGTCGTGTCCGAGGCCACCATCG
TGGGCGTCCGCAAGACCGGGCAGATCTGGCCCAACGATGGCGAGGGCGCTTCCATGGAGACGCAGAAGC
CCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCTCAGGGTCTGAGTGAAGCCGCTCGTTGGAATCCAAG
G
```

Gambar 2. Sekuens M19730.1 Setelah Dibersihkan

Setiap data sekuens yang berupa data *string* akan diekstrak menjadi data numerik menggunakan matriks probabilitas rantai markov atau transisi (Amalutfia & Hafiyusholeh, 2020). Hasil data numerik ini digunakan sebagai data *training* dan data *testing*.

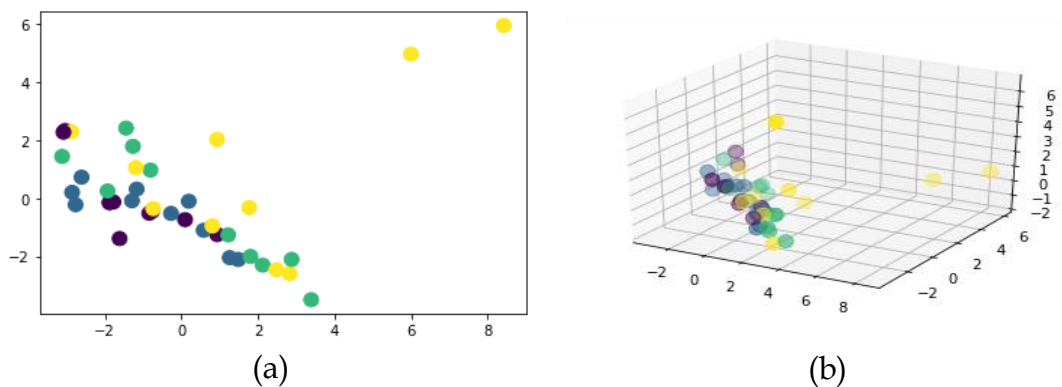
```
[0.18556701030927836, 0.25773195876288657, 0.4536082474226804,
0.10309278350515463, 0.22988505747126436, 0.3735632183908046,
0.20689655172413793, 0.1896551724137931, 0.22377622377622378,
0.36363636363636365, 0.27972027972027974, 0.13286713286713286,
0.09210526315789473, 0.42105263157894735, 0.3026315789473684,
0.18421052631578946]
```

Gambar 3. Contoh Data Numerik Hasil Ekstraksi

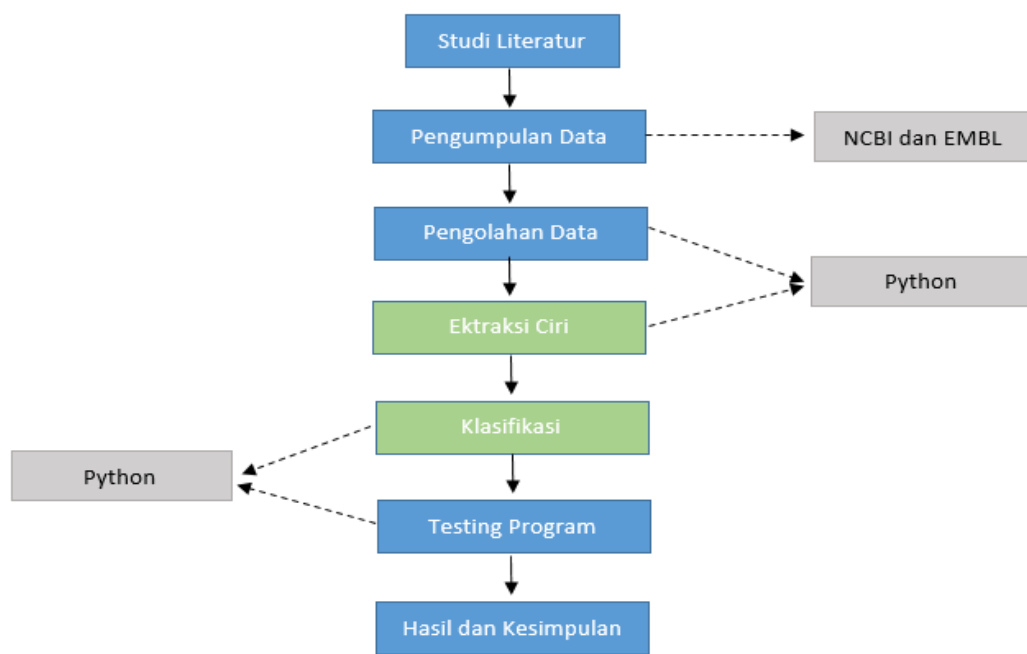
Pada **Gambar 3** ditunjukkan data numerik hasil ekstraksi sekuens DNA/RNA. Setelah semua data diekstraksi maka akan terbentuk matriks berukuran 40×16, dimana baris matriks menunjukkan banyaknya data sedangkan kolom matriks menunjukkan banyaknya fitur. Proses berikutnya adalah mereduksi ukuran matriks data numerik menjadi 2 fitur seperti **Gambar 4** (a) dan menjadi 3 fitur seperti **Gambar 4** (b) dengan

menerapkan metode PCA (Principal Component Analysis). Data hasil reduksi ini yang dimasukkan dalam klasifikasi menggunakan *machine learning*.

Berdasarkan tingkat keparahannya kanker darah (*leukemia*) dapat dikelompokkan menjadi 2, yaitu *leukemia* akut dan *leukemia* kronis. *Acute Lymphoblastic Leukemia* (ALL) dan *Acute Myeloblastic Leukemia* (AML) masuk kedalam kelompok *leukemia* akut, sedangkan *Chronic Lymphoblastic Leukemia* (CLL) dan *Chronic Myeloblastic Leukemia* (CML) masuk ke dalam kelompok *Leukemia* kronis. Berdasarkan hal tersebut maka klasifikasi dilakukan untuk mengidentifikasi tingkat keparahan kanker darah apakah masuk ke kelompok kronis atau akut. Secara umum metodologi penelitian ini dapat dilihat pada **Gambar 5**.



Gambar 4. (a) Visualisasi Data 2 Fitur; (b) Visualisasi Data 3 Fitur



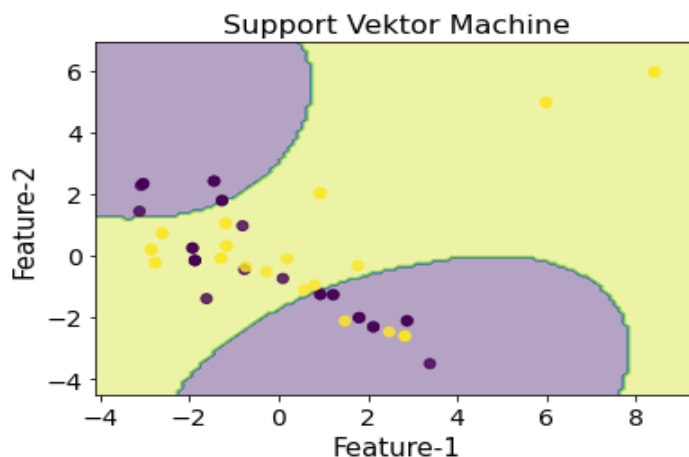
Gambar 5. Metodologi Penelitian

HASIL DAN PEMBAHASAN

Proses klasifikasi dilakukan menggunakan beberapa metode *machine learning*, hasilnya ditampilkan dalam bentuk grafik 2 dimensi.

1. *Support Vector Machine (SVM)*

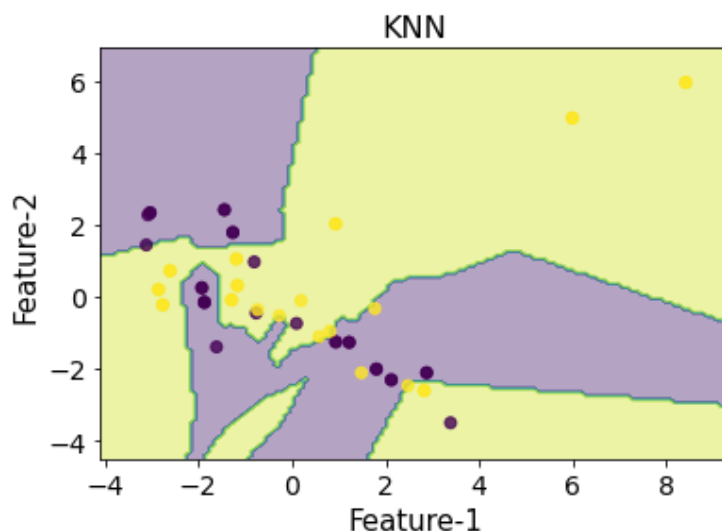
Support Vector Machine yang digunakan adalah *Support Vector Classifier (SVC)* dengan kernel *Radial Basis Function (RBF)* memberikan akurasi 0.375. Daerah hasil klasifikasi ditampilkan dalam bentuk 2 dimensi pada **Gambar 6**.



Gambar 6. Klasifikasi Jenis Leukemia dengan SVC Kernel RBF

2. *K-Neighbors Classifier*

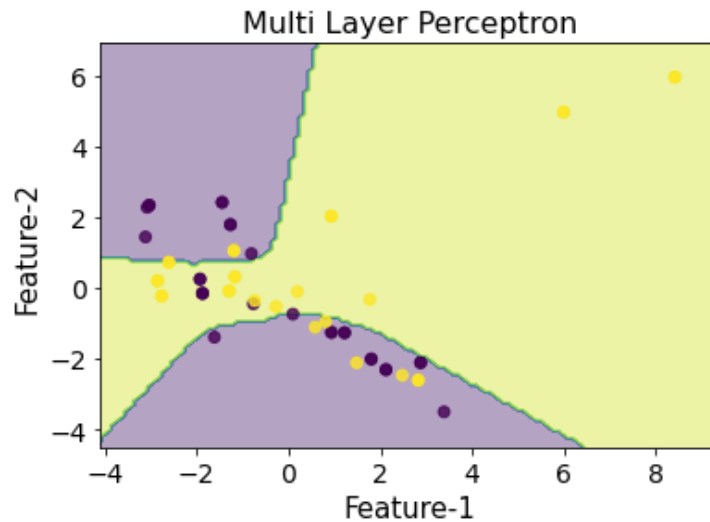
Metode *K-Neighbors Classifier* neighbors awal 3 memberikan akurasi 0.5. Daerah hasil klasifikasi ditampilkan dalam bentuk 2 dimensi pada **Gambar 7**.



Gambar 7. Klasifikasi Jenis *Leukemia* dengan *K-Neighbors Classifier*

3. *Multilayer Perceptron Classifier*

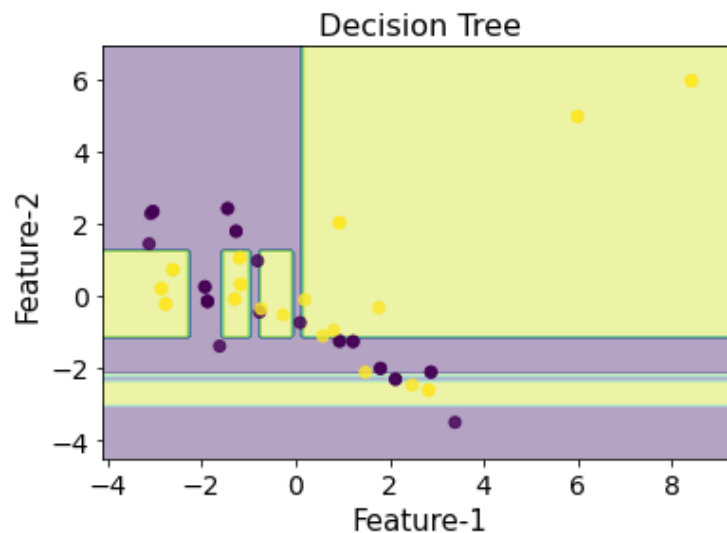
Metode *Multilayer Perceptron Classifier* dengan 1 *hidden layer* dan fungsi aktivasi *The Rectified Linear Unit (ReLU)* memberikan akurasi 0.375. Daerah hasil klasifikasi ditampilkan dalam bentuk 2 dimensi pada **Gambar 8**.



Gambar 8. Klasifikasi Jenis Leukemia dengan *Multilayer Perceptron Classifier*

4. *Decision Tree Classifier*

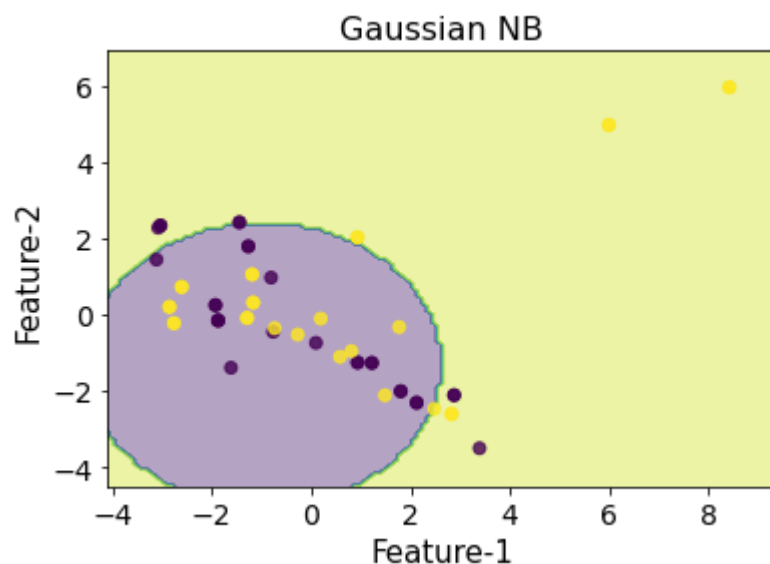
Metode *Decision Tree Classifier* memberikan akurasi 0.833. Daerah hasil klasifikasi ditampilkan dalam bentuk 2 dimensi pada **Gambar 9**.



Gambar 9. Klasifikasi Jenis Leukemia dengan *Decision Tree Classifier*

5. *Gaussian Naive Bayes*

Metode *Gaussian Naive Bayes* memberikan akurasi 0.542. Daerah hasil klasifikasi ditampilkan dalam bentuk 2 dimensi pada **Gambar 10**.



Gambar 10. Klasifikasi Jenis Leukemia dengan *Gaussian Naive Bayes*

Kemampuan metode-metode *machine learning* dalam melakukan identifikasi kanker darah untuk diklasifikasikan dalam kelompok akut atau kronis ditampilkan pada **Tabel 2**.

Tabel 2. Performa Metode Machine Learning untuk Identifikasi Kanker Darah

No	Metode	Akurasi
1.	<i>Support Vector Machine (SVM)</i>	0.375
2.	<i>KNN Classifier</i>	0.5
3.	<i>Multilayer Perceptron Classifier</i>	0.375
4.	<i>Decision Tree Classifier</i>	0.833
5.	<i>Gaussian Naive Bayes</i>	0.542

Berdasarkan **Tabel 2**, diperoleh informasi bahwa *Decision Tree Classifier* memberikan akurasi terbaik yaitu 0.833, dengan kata lain tingkat kebenaran identifikasinya mencapai 83.3%.

SIMPULAN DAN SARAN

Machine learning merupakan metode atau algoritma yang memanfaatkan komputer untuk mengenali pola-pola dari data *input*, sehingga pola tersebut dapat digunakan untuk prediksi ataupun klasifikasi data baru. *Machine learning* banyak dimanfaatkan di berbagai bidang kesehatan, diantaranya adalah identifikasi kanker. Pada penelitian ini digunakan 5 metode *machine learning* yang umum digunakan untuk klasifikasi, yaitu *Support Vector Machine (SVM)*, *KNN Classifier*, *Multilayer Perceptron Classifier*, *Decision Tree Classifier* dan *Gaussian Naive Bayes*. Dari 5

metode tersebut *Decision Tree Classifier* memberikan akurasi terbaik yaitu 0.833, dengan kata lain tingkat kebenaran identifikasinya mencapai 83.3%.

Untuk penelitian selanjutnya perlu dilakukan kolaborasi, melibatkan peneliti dari bidang lain seperti biologi, kimia, ataupun farmasi agar proses identifikasi kanker menjadi lebih akurat.

DAFTAR RUJUKAN

- Al Faroby, M. H. Z. (2018). *Identifikasi Jenis Kanker Darah (Leukemia) Terhadap Pengaruh Parameter Kernel Support Vector Machine Dan Ekstraksi Ciri Dengan Rantai Markov Orde 2*. 30–31.
- Amalutfia, S. Y., & Hafiyusholeh, M. (2020). Analisis Peramalan Nilai Tukar Rupiah Terhadap Dollar dan Yuan Menggunakan FTS-Markov Chain. *Vygotsky*, 2(2), 102. <https://doi.org/10.30736/vj.v2i2.258>
- American Cancer Society. (2019). Facts & Figures 2019. *American Cancer Society*, 1–76. <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf>
- Dese, K., Raj, H., Ayana, G., Yemane, T., Adissu, W., Krishnamoorthy, J., & Kwa, T. (2021). Accurate Machine-Learning-Based classification of Leukemia from Blood Smear Images. *Clinical Lymphoma, Myeloma and Leukemia*, 21(11), e903–e914. <https://doi.org/10.1016/j.clml.2021.06.025>
- Dewi, M. (2017). Sebaran Kanker di Indonesia, Riset Kesehatan Dasar 2007. *Indonesian Journal of Cancer*, 11(29), 1–8. <https://media.neliti.com/media/publications/197251-ID-sebaran-kanker-di-indonesia-riset-keseha.pdf>
- Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., & Ramkumar, P. N. (2020). Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. *Current Reviews in Musculoskeletal Medicine*, 13(1), 69–76. <https://doi.org/10.1007/s12178-020-09600-8>
- Kumar, Vinayshekhar Bannihatti; Kumar, Sujay S; Saboo, V. (2016). 2016 *Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR) : 19-21 Sept. 2016*. 88–93.
- Mahmood, N., Shahid, S., Bakhshi, T., Riaz, S., Ghufran, H., & Yaqoob, M. (2020). Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach. *Medical and Biological Engineering and Computing*, 58(11), 2631–2640. <https://doi.org/10.1007/s11517-020-02245-2>
- Morrison, C., & Hesdorffer, C. S. (2012). Patients' Guide to Leukemia (Panduan untuk Penderita Leukemia). In *Jakarta: PT Indeks*.
- Pratiwi, R. Y. (2014). *Aplikasi Model Fuzzy Untuk Diagnosa Jenis Kanker Darah (Leukemia)*.
- Shyr, D. C., Zhang, B. M., Parkman, R., & Brewer, S. E. (2020). Machine Learning Methods to Better Predict Post-Hematopoietic Stem Cell Transplant (HSCT) Leukemic Relapse in Pediatric Patients with Acute

Lymphoblastic Leukemia: Random Forest (RF) Classification Featuring Serial Post-Transplant Lineage-Specific Chimerism. *Blood*, 136(Supplement 1), 6-7. <https://doi.org/10.1182/blood-2020-139104>
Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56. <https://doi.org/10.1038/s41591-018-0300-7>