

Efektivitas K-Means Clustering dan GMM dalam Menentukan Kluster Organ dan Jaringan Kanker Pada Paru-paru Berdasarkan Nilai HU CT Thorax

Rafi Achmad Fahreza^{1*}
Erviana Widia Astuti¹

AFILIASI :

¹⁾ Jurusan Fisika, Fakultas
Matematika dan Ilmu
Pengetahuan, Universitas
Jember

ALAMAT:

Universitas Jember, Jalan Kalimantan
Tegal Boto, Nomor 37, Jember, Jawa
Timur 68121

KORESPONDENSI:

Rafi Achmad Fahreza¹
211810201054@mail.unej.ac.id

KATA KUNCI:

Adenokarsinoma,
Hounsfield Unit (HU),
Tomografi terkompulasi,
Klusterisasi.

JEI

<https://journal.unej.ac.id/JEI>
jei@unej.ac.id
FMIPA UNIVERSITAS JEMBER
ISSN:3032 3398

ABSTRAK

Penelitian ini membandingkan efektivitas dua algoritma klusterisasi, Gaussian Mixture Model (GMM) dan K-Means, dalam mengelompokkan jaringan pada citra CT scan toraks dengan diagnosis adenokarsinoma berdasarkan nilai Hounsfield Unit (HU). Dalam penelitian ini, nilai HU yang mewakili berbagai jenis jaringan, seperti udara, lemak, otot, dan tulang, diekstraksi dan diklasifikasikan menggunakan kedua algoritma tersebut. GMM menggunakan pendekatan probabilistik yang lebih fleksibel dalam menangani variasi densitas jaringan, sementara K-Means bekerja dengan memisahkan data berdasarkan jarak terdekat dari pusat kluster. Hasil menunjukkan bahwa GMM memberikan performa klusterisasi yang lebih unggul dengan Silhouette Score 0,7447, dibandingkan dengan K-Means yang memperoleh skor 0,70. GMM mampu memisahkan jaringan yang lebih kompleks dengan akurasi lebih tinggi, khususnya pada jaringan dengan nilai HU yang tumpang tindih, seperti jaringan otot dan adenokarsinoma. Oleh karena itu, GMM dinilai lebih efektif dan andal untuk analisis segmentasi pada data medis yang kompleks, seperti diagnosis kanker paru-paru berbasis citra CT scan

PENDAHULUAN

Kanker paru-paru merupakan salah satu kanker paling umum dan menjadi penyebab kematian utama yang diakibatkan oleh kanker baik pada pria maupun wanita. Setelah kanker prostat pada pria dan kanker payudara pada wanita, kanker paru-paru dan bronkus adalah jenis kanker yang paling sering didiagnosis [1], dengan perkiraan 2 juta kasus baru dan 1,76 juta kematian setiap tahun [2]. Kanker paru-paru dapat didiagnosis menggunakan berbagai teknologi, termasuk *Magnetic Resonance Imaging* (MRI), X-ray, dan *Computed Tomography* (CT). Dua metode pencitraan anatomi yang paling umum digunakan untuk mendiagnosis berbagai penyakit paru-paru adalah radiografi dada dengan X-ray dan CT [3].

Hasil pencitraan anatomi berdasarkan metode CT telah banyak digunakan oleh peneliti untuk mengembangkan diagnosis penyakit yang lebih akurat, salah satunya kanker paru-paru [1] [3] [4] [5] [6] [7]. Dalam praktiknya, analisis CT scan sering kali memerlukan teknik pengolahan citra lanjutan untuk mempermudah identifikasi jenis dan karakteristik kanker. Salah satu langkah penting dalam analisis citra CT adalah mengklasterisasi atau mengelompokkan jenis kanker paru-paru berdasarkan nilai *Hounsfield Unit* (HU) yang diperoleh dari CT image [8]. HU merupakan satuan yang menunjukkan densitas jaringan dan menjadi penanda penting dalam membedakan berbagai jenis jaringan pada citra CT [8].

Pendekatan berbasis *machine learning* dapat membantu dalam mengidentifikasi dan mengelompokkan berbagai jaringan dan kanker pada paru-paru berdasarkan pola densitas jaringan, yang selanjutnya dapat mendukung proses diagnosis dan terapi [6] [9]. Namun, beberapa peneliti banyak menggunakan pendekatan *machine learning* dalam bentuk klasifikasi. Salah satunya yang dilakukan oleh (Yunianto dkk., 2021) menggunakan metode naive bayes [10]. Dalam penelitian ini akan menggunakan

Algoritma *K-Means Clustering* dan *Gaussian Mixture Model* (GMM) yang merupakan dua teknik dalam *machine learning* dengan pendekatan klasterisasi yang umum digunakan dalam pemrosesan citra medis.

K-Means bekerja dengan cara membagi data ke dalam kelompok berdasarkan jarak terdekat dari pusat klaster [11], sementara GMM menggunakan pendekatan probabilistik untuk memodelkan distribusi data dalam bentuk campuran *gaussian* [12]. Meskipun kedua metode ini memiliki tujuan yang sama, yaitu melakukan klasterisasi, perbedaan dasar dalam pendekatannya dapat mempengaruhi akurasi dan efektivitas dalam mengklasterisasi jenis kanker paru-paru. Penelitian ini bertujuan untuk menganalisis efektivitas *K-Means Clustering* dan *Gaussian Mixture Model* (GMM) dalam mengklasterisasi jenis kanker paru-paru berdasarkan nilai HU dari CT image. Melalui perbandingan ini, diharapkan dapat diketahui metode mana yang lebih efektif dalam memberikan hasil klasterisasi yang akurat dan dapat diandalkan dalam mendukung diagnosis medis kanker paru-paru.

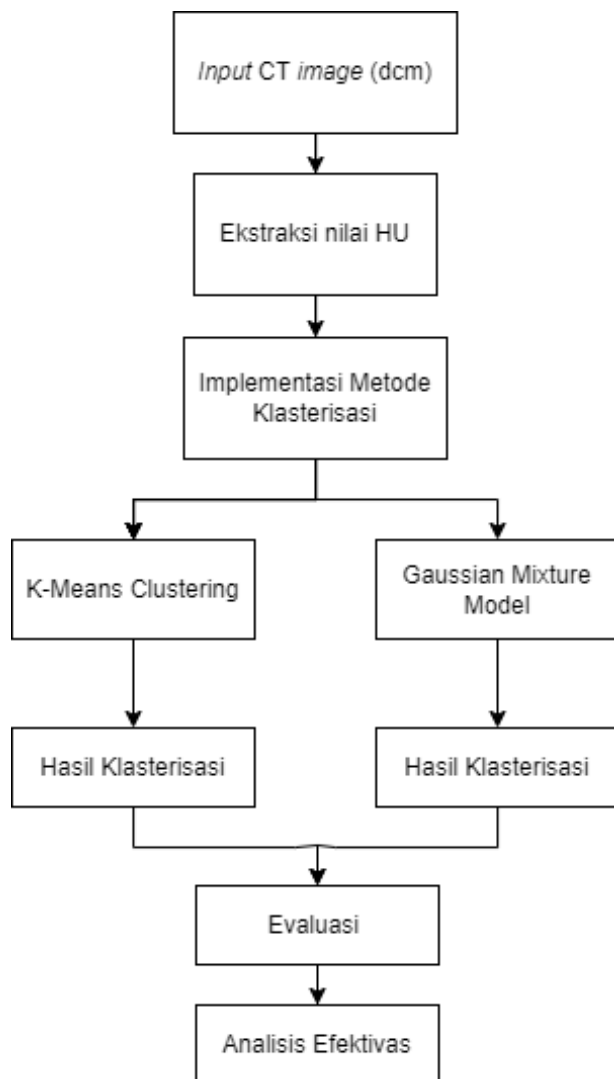
METODE

Penelitian ini merupakan klasterisasi jaringan yang terdapat dalam paru-paru dan mendeteksi kanker paru-paru berdasarkan nilai HU yang diperoleh dari CT image. Penelitian ini menggunakan dua jenis pendekatan *Machine learning* yaitu *K-Means Clustering* dan GMM (*Gaussian Mixture Model*) yang bertujuan untuk membandingkan hasil yang lebih akurat dari kedua metode tersebut dalam klasterisasi berdasarkan nilai HU.

Desain Penelitian

Akuisisi data CT scan dengan format DICOM menjadi tahap awal dalam penelitian ini. Tahap selanjutnya yaitu ekstraksi nilai HU dari CT image untuk mendapatkan informasi densitas jaringan. Berikutnya implementasi algoritma klasterisasi dilakukan dengan

menggunakan dua metode yaitu *K-Means Clustering* dan *Gaussian Mixture Model (GMM)*. Kedua algoritma ini diterapkan secara terpisah pada data untuk mengklasterisasi jaringan dan kanker pada paru-paru berdasarkan pola densitas yang teridentifikasi dari nilai HU. Setelah proses klasterisasi selesai, hasil dari kedua algoritma dievaluasi dan dibandingkan dengan menghitung beberapa metrik evaluasi, seperti *Silhouette Score Index*. Metrik-metrik ini memberikan gambaran tentang kualitas dan efektivitas hasil klasterisasi dari masing-masing metode. Tahapan berikutnya yaitu analisis efektivitas berdasarkan evaluasi yang telah dilakukan. Tahapan-tahapan tersebut dalam dilihat pada



Gambar 1. Diagram Alir Penelitian

Data

Objek pada penelitian ini yaitu *CT image* paru-paru yang merupakan data sekunder yang diperoleh melalui *Cancer Imaging Archive* dengan laman <https://nbia.cancerimagingarchive.net/nbia-search/>. Variabel bebas pada penelitian ini adalah metode klasterisasi yang digunakan, di antaranya *K-Means Clustering* dan *GMM (Gaussian Mixture Model)*. Variabel terikat pada penelitian ini adalah ketepatan dalam memunculkan nilai HU dan mendeteksi kelainan yang ada. Variabel kontrol penelitian ini adalah nilai HU *reference* yang dicantumkan.

Analisa

Analisa keefektifan metode *K-Means* dan *GMM* dilakukan dengan menggunakan *Silhouette Score* untuk mengukur kekompakan dan keterpisahan klaster. *Silhouette Score* memberikan gambaran seberapa baik data dalam klaster tertentu berkumpul erat dan terpisah jelas dari klaster lain. Skor dihitung berdasarkan perbedaan rata-rata jarak antara data di dalam satu klaster dan jarak antara data di klaster terdekat lainnya. *Silhouette Score* dinyatakan dalam rentang -1 hingga 1, dengan skor mendekati 1 menunjukkan hasil klasterisasi yang baik. Semakin tinggi *Silhouette Score* maka semakin baik hasil klasterisasi. Analisis ini digunakan untuk menilai keefektifan metode *K-Means* dan *GMM* dalam mengelompokkan organ dan kanker pada paru-paru berdasarkan nilai HU dari *CT image*

HASIL DAN PEMBAHASAN

Penelitian ini melakukan klasterisasi nilai *Hounsfield Unit (HU)* dari citra *CT scan* toraks dengan diagnosis adenokarsinoma menggunakan algoritma *Gaussian Mixture Model (GMM)* dan *K-Means*. Tahap awal yang dilakukan sebelum mengembangkan model *Unsupervised learning* yaitu dengan melakukan *Preprocessing*. Data yang digunakan berupa citra *CT scan* toraks dengan diagnosis adenokarsinoma sangat kotor, sehingga diperlukan *Preprocessing*

data. *Preprocessing* data yang akan dilakukan dibagi menjadi beberapa bagian *Exploratory Data Analysis* (EDA). Pada tahapan EDA, dilakukan untuk mengetahui karakteristik dari sebuah data, sehingga mempermudah ketika melakukan pengembangan *unsupervised machine learning* dengan *K-Means* dan *GMM*. Gambar 2 menampilkan metadata dari citra dicom adenokarsinoma. Citra yang digunakan memiliki metadata yang lengkap dengan minimal HU -1024, maksimal HU bernilai 2987, dan rata-rata HU bernilai -675,31 dengan standar deviasi 449,79. Berdasarkan data tersebut terlihat bahwa mulai dari udara, jaringan lunak, air, dan jaringan keras

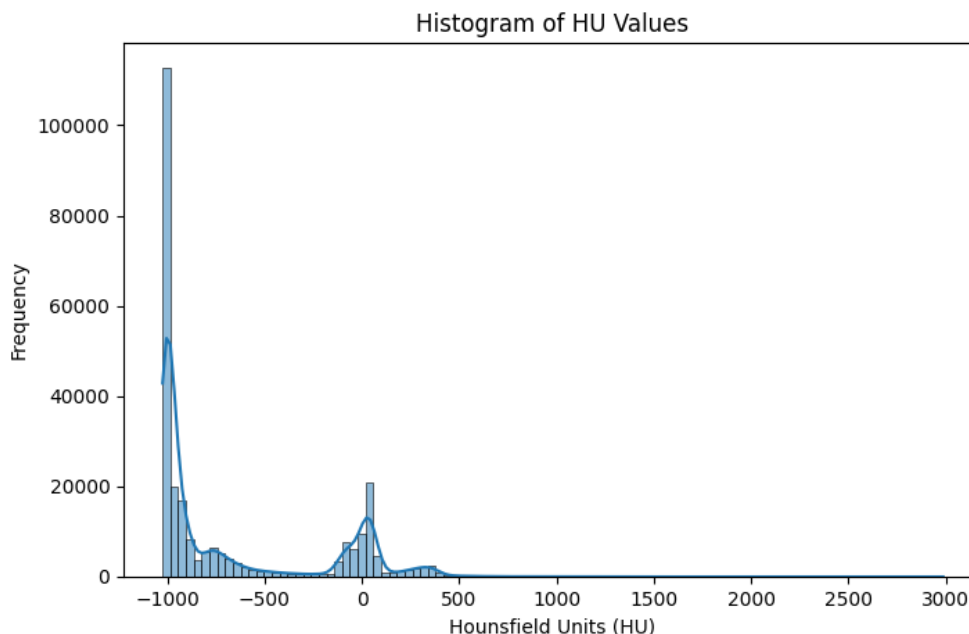
dapat dicitrakan menggunakan *CT scan* oleh manufaktur SIEMENS.

Tahapan EDA selanjutnya yaitu melakukan plot histogram untuk mengetahui bagaimana karakteristik dari nilai HU pada citra *CT scan* yang digunakan. Gambar 3 merupakan histogram visualisasi fitur hasil ekstraksi nilai HU. Histogram tersebut menunjukkan nilai distribusi HU dalam *CT image* dengan rentang nilai -1000 sampai dengan 3000. Frekuensi nilai HU paling rendah berada pada nilai -1000 yang merupakan nilai HU udara atau ruang kosong atau udara, hal ini menunjukkan bahwa area udara sangat mendominasi dalam citra tersebut.

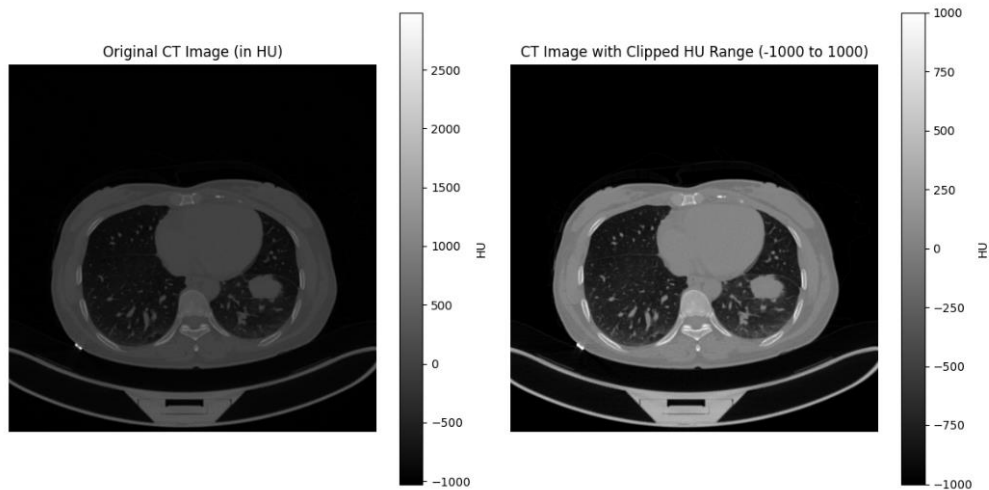
```
(0028,1053) Rescale Slope          DS: '1'
(0028,1054) Rescale Type          LO: 'HU'
(0028,1055) Window Center & Width Explanation LO: ['WINDOW1', 'WINDOW2']
(0029,0010) Private Creator       LO: 'SIEMENS CSA HEADER'
(0029,0011) Private Creator       LO: 'SIEMENS MEDCOM HEADER'
(7FE0,0010) Pixel Data            OW: Array of 524288 elements

Basic Statistics for HU values:
Min HU: -1024.0
Max HU: 2987.0
Mean HU: -675.31
Standard Deviation HU: 449.79
```

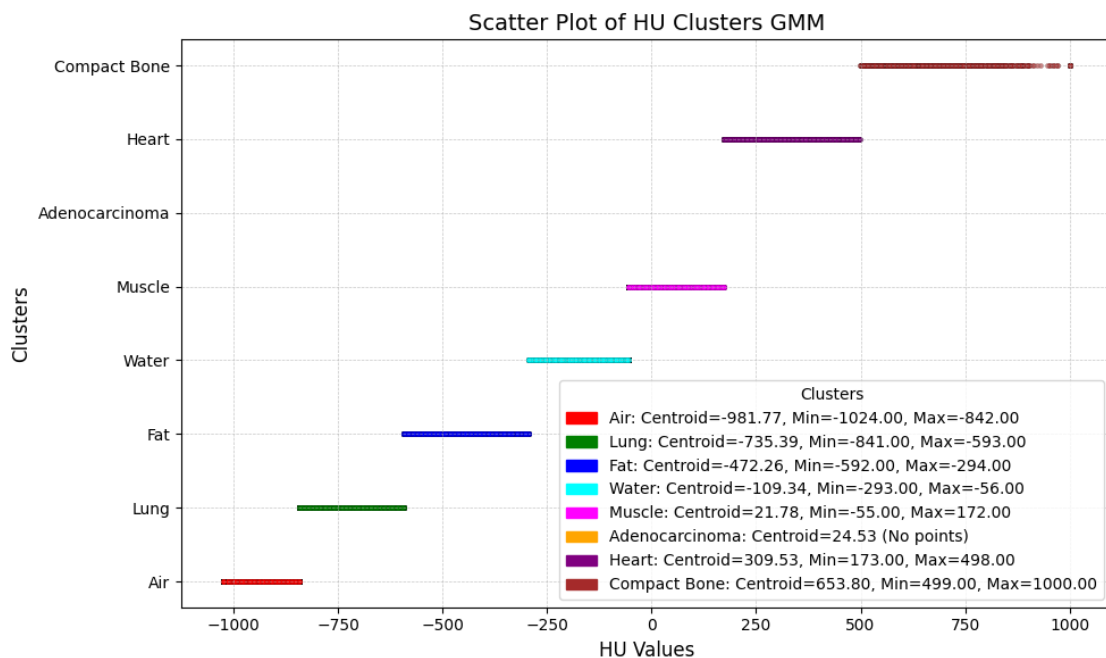
Gambar 2 Metadata dicom CT Adenokarsinoma



Gambar 3. Sebaran Nilai HU



Gambar 4 Hasil citra dengan threshold HU -1000 sampai 1000

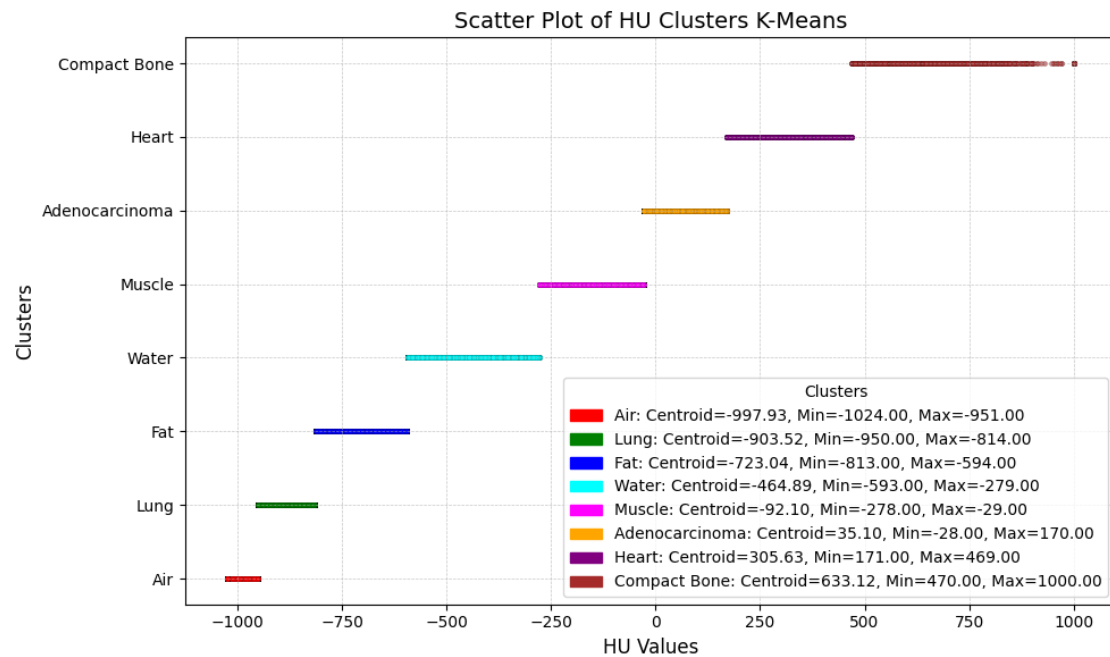


Gambar 5. Hasil Klasterisasi Nilai HU Menggunakan Gaussian Mixture Model atau GMM

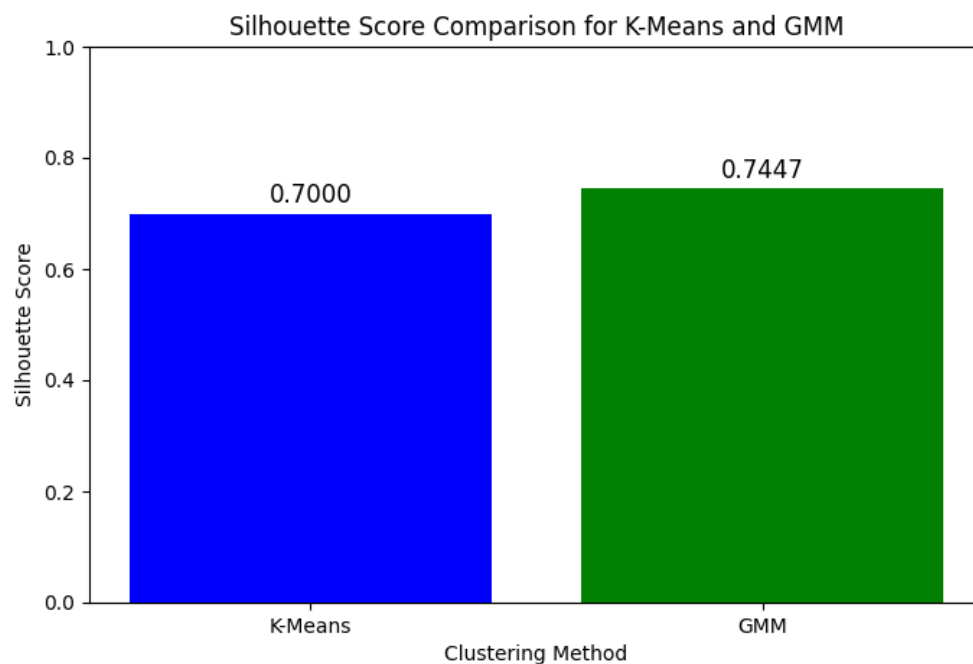
Puncak kedua terdapat pada rentang HU sekitar 0—10 yang merupakan nilai HU untuk air. Frekuensi kemunculan nilai HU di atas 5000 sangat rendah, hingga tidak terlihat pada gambar. Hal tersebut menunjukkan bahwa hanya sedikit area piksel yang memiliki nilai HU sangat tinggi.

Tahapan *preprocessing* selanjutnya setelah melakukan EDA yaitu dengan memberikan ambang batas (*threshold*) untuk rentang HU yang memiliki nilai di atas 1000, maka akan dikembalikan ke nilai HU 1000. Hal ini dilakukan

untuk mencegah *machine learning* kesulitan dalam melakukan klasterisasi karena data pada rentang nilai HU di atas 1000 sangat sedikit. Selain itu, penetapan *threshold* HU ke nilai 1000 juga berdasarkan pada nilai HU untuk tulang yang juga memiliki nilai 1000. Hasil yang didapatkan terlihat pada Gambar 4 yang menunjukkan citra yang sudah diberi *threshold* terlihat lebih jelas dan mudah dianalisis oleh model *unsupervised machine learning* seperti K-Means dan GMM.



Gambar 6. Hasil Klasterisasi Nilai HU Menggunakan K-Means Clustering



Gambar 7 Silhouette Score K-Means dan GMM

Gambar 5 menampilkan hasil klasterisasi nilai *Hounsfield Unit* (HU) dari citra *CT scan* toraks dengan diagnosis adenokarsinoma menggunakan algoritma *Gaussian Mixture Model* (GMM). Pada klasterisasi GMM,

delapan jenis jaringan dapat diklasifikasi dengan *centroid* dan rentang HU menurut karakteristik HU yang sesuai untuk setiap jenis jaringan. Jaringan udara memiliki *centroid* terendah -981.77 HU dan rentang -1024

hingga -842 HU, sementara jaringan tulang padat memiliki *centroid* tertinggi di 653.80 HU dengan rentang 499 – 1000 HU. Jaringan lain yang termasuk paru-paru, lemak, air, otot, juga memiliki rentang HU masing-masing yang unik dan sesuai dengan tipe jaringan kontrol. Namun, pada klaster adenokarsinoma, dengan *centroid* di 24.53 HU, tidak ada data masuk ke dalam klaster. Hal ini disebabkan oleh ketidaksesuaian area rentang kanker pada citra yang dianalisis. GMM memiliki keunggulan dalam klasterisasi bentuk klaster yang dengan lebih fleksibel, yang membuat segmentasi lebih akurat untuk jaringan dengan variasi densitas dan rentang. Dari distribusi nilai HU, model GMM mampu membuat pola variasi jaringan dengan lebih fleksibel. Hal ini bisa dilihat dari variasi rentang HU dalam setiap jaringan referensi seperti udara, paru-paru, lemak, tulang padat yang dapat diklasifikasi dengan baik oleh GMM karena mampu memisahkan berbagai rentang HU pada jaringan ini. GMM juga mampu memberikan batasan klaster yang mencerminkan distribusi HU dengan baik. Misalnya, otot dan air memiliki HU yang relatif sama, tetapi tetap mampu dipisahkan oleh GMM, menandakan model mampu mendeteksi perbedaan kecil pada distribusi densitas jaringan.

Gambar 6 menampilkan hasil klasterisasi menggunakan algoritma *K-Means*, *centroid* dan rentang HU untuk setiap jaringan menunjukkan perbedaan dibandingkan hasil GMM sebelumnya. *Centroid* udara berada pada -997.93 HU dengan rentang yang lebih luas, yaitu dari -1024 hingga -951 HU, sementara jaringan tulang padat memiliki *centroid* pada 633.12 HU dengan rentang sedikit lebih rendah dibandingkan GMM. Klaster adenokarsinoma memiliki *centroid* 35.10 HU dan berada berdekatan dengan jaringan otot dan air, yang menimbulkan potensi tumpang tindih antar klaster, terutama pada area dengan perbedaan densitas tipis. Sebagai contoh, jaringan otot dan adenokarsinoma, yang memiliki HU serupa, cenderung tidak terpisah dengan jelas menggunakan *K-Means*, yang mengarah pada potensi salah klasifikasi. Hal ini dapat mempengaruhi ketepatan dalam

analisis jaringan, terutama pada jaringan kanker seperti adenokarsinoma yang membutuhkan identifikasi akurat. Selain itu, *K-Means* cenderung kurang optimal dalam menangani variasi alami densitas jaringan, sehingga model ini lebih cocok untuk data set dengan klaster yang jelas terpisah. Meskipun hasil yang didapatkan terdapat beberapa yang tumpang tindih, *K-Means* memberikan hasil yang lebih cepat secara komputasi, namun batas antar klaster yang dihasilkan kurang optimal dalam menangani distribusi HU yang saling tumpang tindih.

Berdasarkan hasil *Silhouette Score* pada Gambar 7 untuk metode *K-Means* dan *Gaussian Mixture Model* (GMM) dalam klasterisasi citra CT paru-paru berdasarkan nilai *Hounsfield Unit* (HU), diperoleh skor 0,70 untuk *K-Means* dan 0,7447 untuk GMM. *Silhouette Score* merupakan metrik yang menunjukkan seberapa baik data dikelompokkan, di mana nilai mendekati 1 menandakan pemisahan klaster yang baik dan konsistensi dalam setiap klaster. Skor 0,70 pada *K-Means* menunjukkan bahwa metode ini mampu mengelompokkan data dengan cukup baik, meskipun masih ada kemungkinan beberapa data yang kurang optimal dalam klasterisasi atau terjadi tumpang tindih antar klaster. Sementara itu, *Silhouette Score* pada GMM yang lebih tinggi, yaitu 0,7447, menunjukkan bahwa metode GMM memberikan hasil klasterisasi yang lebih baik dibandingkan *K-Means*. GMM, dengan pendekatan probabilitiknya, lebih fleksibel dalam menangani distribusi data yang kompleks, sehingga dapat menghasilkan klaster yang lebih terpisah dan kompak. Dengan demikian, skor yang lebih tinggi pada GMM mengindikasikan bahwa metode ini lebih efektif untuk mengklasterisasi nilai HU dalam citra CT paru-paru, sehingga GMM bisa menjadi pilihan yang lebih baik untuk analisis atau segmentasi pada data medis yang kompleks seperti ini.

KESIMPULAN

Kesimpulan dari penelitian ini yaitu klusterisasi nilai *Hounsfield Unit* (HU) dari citra CT toraks dengan diagnosis adenokarsinoma menggunakan algoritma *Gaussian Mixture Model* (GMM) dan *K-Means*, di mana GMM menunjukkan performa lebih unggul dalam mengelompokkan jaringan tubuh dengan distribusi HU yang kompleks. GMM mampu menghasilkan klaster yang lebih akurat sesuai karakteristik jaringan, seperti udara, lemak, otot, dan tulang, berkat fleksibilitasnya dalam menangani variasi densitas. Sebaliknya, *K-Means* menunjukkan keterbatasan dalam memisahkan jaringan dengan nilai HU yang berdekatan, seperti jaringan otot dan adenokarsinoma, yang meningkatkan potensi tumpang tindih dan salah klasifikasi pada area dengan perbedaan densitas tipis. Berdasarkan hasil *Silhouette Score*, GMM mendapatkan skor lebih tinggi (0,7447) dibandingkan *K-Means* (0,70), menunjukkan bahwa GMM menghasilkan pemisahan klaster yang lebih baik dan lebih kompak, sehingga lebih sesuai untuk analisis jaringan pada data medis yang kompleks seperti CT scan toraks dengan adenokarsinoma

DEKLARASI

Tim penulis menyatakan bahwa penelitian yang telah dilakukan merupakan hasil kerja dari tim penulis. Semua aspek yang terdapat dalam penelitian ini tidak memiliki konflik kepentingan yang berpengaruh terhadap hasil penelitian. Penelitian ini juga tidak mengandung plagiarisme karena penulis menghormati hak cipta dan privasi semua pihak yang bersangkutan

REFERENSI

- [1] S. Zheng, "Deep learning for lung cancer on computed tomography: early detection and prognostic prediction," 2021.
- [2] A. A. Thai, B. J. Solomon, L. V. Sequist, J. F. Gainor, and R. S. Heist, "Lung cancer," *Lancet*, vol. 398, no. 10299, pp. 535–554, 2021, doi: 10.1016/S0140-6736(21)00312-3.
- [3] A. Heidari, D. Javaheri, S. Toumaj, N. J. Navimipour, M. Rezaei, and M. Unal, "A new lung cancer detection method based on the chest CT images using Federated Learning and blockchain systems," *Artif. Intell. Med.*, vol. 141, p. 102572, 2023, doi: <https://doi.org/10.1016/j.artmed.2023.102572>.
- [4] I. Nazir, I. ul Haq, S. A. AlQahtani, M. M. Jadoon, and M. Dahshan, "Machine learning-Based Lung Cancer Detection Using Multiview Image Registration and Fusion," *J. Sensors*, 2023, doi: <https://doi.org/10.1155/2023/6683438>.
- [5] V. Rajasekar, M. P. Vaishnnave, S. Premkumar, V. Sarveshwaran, and V. Rangaraaj, "Lung cancer disease prediction with CT scan and histopathological images feature analysis using deep learning techniques," *Results Eng.*, vol. 18, no. August 2022, p. 101111, 2023, doi: 10.1016/j.rineng.2023.101111.
- [6] Y. Li, X. Wu, P. Yang, G. Jiang, and Y. Luo, "Machine learning for Lung Cancer Diagnosis, Treatment, and Prognosis," *Genomics, Proteomics Bioinforma.*, vol. 20, no. 5, pp. 850–866, 2022, doi: 10.1016/j.gpb.2022.11.003.
- [7] M. E. Widiatmoko and S. Ramadanti, "Nilai *Hounsfield Unit* (HU) CT-Scan pada Lesi Paru-Paru Pasien Suspek COVID-19," *J. Kesehat. Vokasional*, vol. 8, no. 3, p. 174, 2023, doi: 10.22146/jkesvo.78738.
- [8] S. Puspaningrum, N. Putu, R. Jeniyanthi, and I. M. P. Darmita, "Analisis nilai CT Number pada pemeriksaan CT scan Thorax pada Kasus Efusi Pleura di RS Bhayangkara Makassar," *Naut. J. Ilm. Multidisiplin*, vol. 1, no. 10, pp. 1212–1216, 2023.
- [9] E. Magdy, N. Zayed, and M. Fakhr, "Automatic Classification of Normal and Cancer Lung CT Images Using Multiscale AM-FM Features," *Int. J. Biomed. Imaging*, vol. 2015, 2015, doi: 10.1155/2015/230830.

- [10] M. Yudianto *et al.*, "Klasifikasi Kanker Paru Paru Menggunakan Naive Bayes Dengan Variasi Filter Dan Ekstraksi Ciri Gray Level Co-occurrence Matrix (GLCM)," *Indones. J. Appl. Phys.*, vol. 11, no. 2, pp. 256–267, 2021.
- [11] Sulistyowati, B. Eno Ketherin, A. Anjani Arifiyanti, and A. Sodik, "Analisa Segmentasi Konsumen Menggunakan Algoritma K-Means Clustering," *Sains Dan Teknol. Terap.*, pp. 51–58, 2018.
- [12] J. Riyono, S. D. Puspa, and C. E. Pujiastuti, "Simulasi Clustering Provinsi di Indonesia dalam Penyebaran Covid-19 Berdasarkan Indikator Kesehatan Masyarakat Menggunakan Algoritma Gaussian Mixture Model," *MAJAMATH J. Mat. dan Pendidik. Mat.*, vol. 5, no. 1, pp. 43–60, 2022.