# SISTEM PENGELOMPOKAN FOTO OTOMATIS DAN PENANGANAN KONDISI *FEW-SHOT* MENGGUNAKAN ALGORITMA HDBSCAN DAN AUGMENTASI DATA GENERATIF

**Proposal Skripsi**

Diajukan untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana
Komputer Jurusan Teknik Informatika pada Fakultas
Sains dan Teknologi UIN Alauddin Makassar

Oleh :

**ADE NURCHALISA**

NIM 60200122039

**FAKULTAS SAINS DAN TEKNOLOGI**

**UNIVERSITAS ISLAM NEGERI ALAUDDIN MAKASSAR**

**2026**

# BAB I

# PENDAHULUAN

## *A. Latar Belakang Masalah*

Dalam lingkungan organisasi kemahasiswaan dan institusi pendidikan, dokumentasi visual kegiatan seperti seminar, pelatihan, atau acara sosial merupakan aset penting yang terus terakumulasi dalam jumlah besar. Namun, seringkali ribuan foto ini tersimpan tanpa struktur pengarsipan yang memadai, sehingga proses pencarian dan pengelolaan foto spesifik menjadi tantangan signifikan yang berujung pada inefisiensi waktu. Kondisi ini menegaskan urgensi pengembangan sistem pengelompokan foto otomatis yang mampu mengidentifikasi kemiripan wajah secara mandiri. Bushey (2024) dan Malek dkk. (2024) menekankan bahwa peran sistem semacam ini krusial tidak hanya untuk efisiensi manajemen arsip digital, tetapi juga relevan dalam konteks keamanan dan dokumentasi kelembagaan.

Implementasi sistem ini menghadapi tantangan kompleksitas visual, seperti variasi pencahayaan, ekspresi, dan sudut pengambilan gambar. Keragaman ini menuntut metode *unsupervised learning* yang adaptif (Shin dkk., 2023). Algoritma tradisional seperti DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) yang membentuk klaster berdasarkan kepadatan titik data tanpa perlu menentukan jumlah klaster di awal (Ester dkk., 1996) seringkali gagal mengatasi variasi ini karena ketergantungannya pada parameter radius (*epsilon*) tunggal. Kalpavruksha dkk. (2025) menyoroti

bahwa hal ini menyebabkan DBSCAN kurang optimal dalam menangani *dataset* dengan kepadatan bervariasi, seperti memisahkan klaster yang sangat padat dan klaster yang renggang secara akurat.

Sebagai solusi, algoritma HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) dikembangkan dengan struktur hierarki yang menawarkan fleksibilitas lebih besar dan mengurangi sensitivitas parameter (Mcinnes & Healy, 2017). Algoritma ini terbukti lebih *robust* dalam menangani data kompleks dan berdimensi tinggi seperti *face embeddings* (Peña Asensio & Ferrari, 2025). Meski demikian, performanya masih bergantung pada metrik jarak. Metrik standar *Euclidean* seringkali kurang optimal menangkap struktur non-linear data dimensi tinggi (Du, 2023). Oleh karena itu, penelitian ini mengeksplorasi penggunaan metrik *Quantum Jensen-Shannon Divergence* (QJSD), yang terbukti lebih sensitif dalam mengukur perbedaan distribusi antar-sampel (Hoyos Osorio & Sanchez Giraldo, 2024).

Tantangan lain yang signifikan adalah kondisi *few-shot*, di mana jumlah sampel data sangat terbatas pada beberapa individu, sehingga menyulitkan pembentukan klaster yang stabil. Untuk mengatasinya, pendekatan augmentasi data generatif diusulkan (Fortın & Nishikawa, 2024). Berbeda dengan teknik klasik, pendekatan ini menggunakan model *deep learning* untuk menciptakan sampel sintetis yang memperkaya distribusi fitur (Che dkk., 2025). Dalam penelitian ini, diterapkan metode *Cluster-based Generative Augmentation* (CGA) yang bekerja di ruang fitur untuk

memperkaya klaster minoritas, sehingga batas antar-klaster dapat dikenali lebih baik (Haghbin dkk., 2025; Tan dkk., 2023).

Dalam perspektif Islam, prinsip keteraturan (*nizām*) dan efisiensi (*kifāyah*) bukan sekadar konsep teknis, melainkan refleksi dari nilai-nilai fundamental yang dianjurkan. Keteraturan alam semesta, meliputi peredaran matahari dan bulan untuk perhitungan waktu, menjadi inspirasi bagi manusia untuk menerapkan prinsip serupa dalam pengelolaan kehidupan, termasuk dalam pengembangan dan pemanfaatan teknologi. Sebagaimana digambarkan dalam firman Allah Swt. dalam Surah Yunus/10:5 sebagai berikut:

هُوَ الَّذِيْ جَعَلَ الشَّمْسَ ضِيَآءً وَّالْقَمَرَ نُوْرًا وَّقَدَّرَهٗ مَنَازِلَ لِتَعْلَمُوْا عَدَدَ السِّنِيْنَ وَالْحِسَابَۗ مَا خَلَقَ اللّٰهُ ذٰلِكَ اِلَّا بِالْحَقِّۗ يُفَصِّلُ الْاٰيٰتِ لِقَوْمٍ يَّعْلَمُوْنَ(٥)

**Terjemahnya :**

> "Dialah yang menjadikan matahari bersinar dan bulan bercahaya. Dialah pula yang menetapkan tempat-tempat orbitnya agar kamu mengetahui bilangan tahun dan perhitungan (waktu). Allah tidak menciptakan demikian itu, kecuali dengan benar. Dia menjelaskan tanda-tanda (kebesaran-Nya) kepada kaum yang mengetahui"(Kementerian Agama RI, 2022)

Berdasarkan Tafsir Kementerian Agama RI (2022), ayat ini menjelaskan bahwa Allah menciptakan sistem peredaran benda langit dengan keteraturan yang sangat presisi agar manusia dapat menata kehidupan duniawi secara terukur dan efisien. Quraish Shihab (2002b) dalam Tafsir Al-Misbah menegaskan bahwa ayat ini mengajarkan nilai *nizām* (keteraturan) dan *hikmah* (kebijaksanaan), bahwa segala ciptaan Allah mengandung pola yang dapat dijadikan teladan manusia dalam merancang dan mengelola

aktivitasnya secara sistematis. Keteraturan yang menjadi hukum alam ini merefleksikan prinsip efisiensi yang juga diupayakan dalam penelitian ini, yaitu mengoptimalkan sistem pengelompokan foto otomatis yang mampu bekerja secara teratur, efisien, dan adaptif. Sebagaimana keteraturan peredaran bulan dan matahari memudahkan manusia menghitung waktu, demikian pula keteraturan sistem cerdas dalam penelitian ini diharapkan mampu mempermudah pengelolaan data visual yang kompleks secara terstruktur.

Jika ayat pertama (Surah Yunus/10:5) menjadi landasan untuk meneladani keteraturan (*nizām*) ciptaan-Nya, maka prinsip ini diperkuat dengan larangan tegas untuk melakukan hal yang sebaliknya, yaitu berlebih-lebihan (*isrāf*). Islam secara tegas melarang perilaku *isrāf*, sebagaimana termaktub dalam Surah Al-A'raf/7:31 sebagai berikut:

۞ يٰبَنِيْٓ اٰدَمَ خُذُوْا زِيْنَتَكُمْ عِنْدَ كُلِّ مَسْجِدٍ وَّكُلُوْا وَاشْرَبُوْا وَلَا تُسْرِفُوْاۚ اِنَّهٗ لَا يُحِبُّ الْمُسْرِفِيْنَ ۝ (٣١)

**Terjemahnya :**

> "Wahai anak cucu Adam, pakailah pakaianmu yang indah pada setiap (memasuki) masjid dan makan serta minumlah, tetapi janganlah berlebihan. Sesungguhnya Dia tidak menyukai orang-orang yang berlebihan" (Kementerian Agama RI, 2022).

Menurut Tafsir Al-Misbah (Shihab, 2002a), larangan *isrāf* dalam ayat ini tidak hanya berkaitan dengan konsumsi makanan dan minuman, tetapi mencakup seluruh bentuk pemborosan terhadap sumber daya yang dianugerahkan Allah, termasuk waktu, tenaga, dan potensi akal. Tafsir

Kemenag (Kementerian Agama RI, 2022) menambahkan bahwa *isrāf* adalah penggunaan sesuatu di luar batas kebutuhan dan kemaslahatan. Dalam konteks modern, pemborosan dapat terjadi dalam bentuk penggunaan teknologi dan waktu yang tidak efisien. Penelitian ini sejalan dengan nilai tersebut karena berupaya mengurangi pemborosan waktu dalam proses pencarian dan pengelolaan foto melalui sistem otomatis berbasis kecerdasan buatan. Upaya optimalisasi sistem ini bukan semata efisiensi teknis, tetapi juga merupakan manifestasi nilai Islam tentang pentingnya menggunakan sumber daya secara bijak dan produktif demi kemaslahatan bersama.

Berdasarkan latar belakang permasalahan, potensi solusi, dan relevansinya dengan nilai-nilai keislaman tersebut, penelitian ini mengusulkan judul: **"Sistem Pengelompokan Foto Otomatis dan Penanganan Kondisi *Few-Shot* Menggunakan Algoritma HDBSCAN dan Augmentasi Data Generatif".**

## *B. Rumusan Masalah*

Berdasarkan uraian pada latar belakang di atas, dapat dirumuskan suatu permasalahan sebagai berikut:

1. Bagaimana merancang sistem pengelompokan otomatis yang efektif untuk mengatasi permasalahan data foto dokumentasi yang tidak terstruktur dan sulit dicari?
2. Bagaimana mengatasi permasalahan kategori foto dengan jumlah sampel sangat sedikit (*few-shot*) yang menyebabkan algoritma

clustering sulit membentuk cluster kecil atau cenderung menganggapnya sebagai noise?

## *C. Definisi Operasional Variabel dan Ruang Lingkup Penelitian*

### 1. Definisi Operasional Variabel

a. Variabel bebas (independen) dalam penelitian ini adalah:

1) Jenis metrik jarak yang digunakan dalam algoritma HDBSCAN, yaitu *Euclidean* dan *Quantum Jensen-Shannon Divergence* (QJSD).

a) *Euclidean Distance* yaitu metrik jarak standar yang menghitung garis lurus antara dua titik vektor $p$ dan $q$ (Salsabilah dkk., 2024). Dioperasionalkan menggunakan rumus:

$$D(i,j) = \sqrt{\left(x_{1i} - x_{1j}\right)^2 + \left(x_{2i} - x_{2j}\right)^2 + i \ldots + \left(x_{ki} - x_{kj}\right)^2} \quad (1.1)$$

Keterangan :

$D(i,j)$ : Jarak data i ke pusat klaster j

$x_{ki}$ : Data ke-i pada atribut data ke-k

$x_{kj}$ : Jarak antara titik pusat ke j pada atribut k

b) *Quantum Jensen-Shannon Divergence* (QJSD), dioperasionalkan menggunakan rumus (Virosztek, 2021):

$$J(A,B) = \frac{1}{2}\mathrm{Tr}\,\eta(A) + \frac{1}{2}\mathrm{Tr}\,\eta(B) - \mathrm{Tr}\,\eta\left(\frac{A+B}{2}\right) \quad (1.2)$$

Keterangan :

A, B : *Density matrices* (*positive definite, trace*=1)

Tr : *Trace operator* (*sum of diagonal elements*)

$\eta(x)$ : *Standard entropy function* = $x \log x$

(Catatan: Nilai -Tr $\eta(\rho)$ merepresentasikan *Von Neumann entropy*)

(A+B)/2 : *Arithmetic mean of density matrices*

2) Penerapan augmentasi data generatif menggunakan *Cluster-based Generative Augmentation* (CGA).

b. Variabel terikat (dependen) adalah:

1) Performa hasil klasterisasi yang diukur menggunakan dua metrik evaluasi utama, yaitu *Silhouette Score* dan *Davies–Bouldin Index* (DBI).

a) *Silhouette Score,* dioperasionalkan menggunakan rumus (Peña Asensio & Ferrari, 2025) :

$$S_i = \frac{b_i - a_i}{max\ (b_i, a_i)} \tag{1.3}$$

Keterangan :

$a_i$ = jarak rata-rata intra-klaster

$b_i$ = jarak rata-rata ke klaster terdekat

b) *Davies–Bouldin Index* (DBI), dioperasionalkan menggunakan rumus (Dinata dkk., 2020)

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \frac{SSW_i - SSW_{ij}}{SSB_{ij}} \tag{1.4}$$

Keterangan :

SSW = kohesi (sebaran data dalam klaster)

SSB = separasi (jarak antar pusat klaster)

c. Variabel kontrol meliputi:

1) Model ekstraksi fitur wajah yang digunakan, yaitu *InsightFace* untuk menghasilkan *face embeddings.*
2) Parameter *minimum cluster size* (*min_cluster_size*) pada HDBSCAN yang dijaga konstan selama eksperimen.
3) Jumlah data input yang dikontrol untuk membedakan kondisi dataset asli dan dataset hasil augmentasi.

Untuk menggambarkan secara sistematis bagaimana penelitian ini mengimplementasikan algoritma HDBSCAN dengan dua pendekatan baru yaitu pemilihan metrik jarak adaptif *(Euclidean vs QJSD)* dan augmentasi generatif berbasis klaster (CGA), berikut pseudocode rancangan eksperimen yang digunakan:

```
Algorithm Eksperimen_Utama_HDBSCAN(Dataset_Embeddings):
    // --- TAHAP 1: Manipulasi Metrik Jarak (Variabel Independen 1) ---
    // Jalankan HDBSCAN dengan dua level metrik berbeda
    Hasil_Euclidean = HDBSCAN(Dataset_Embeddings, metric='euclidean',
                              min_cluster_size=5)
    Hasil_QJSD      = HDBSCAN(Dataset_Embeddings,
                              metric=Fungsi_Kustom_QJSD,
                              min_cluster_size=5)

    // Evaluasi Kualitas Klaster (Variabel Dependen) Silhouette & DBI
    Skor_Euclidean  = Hitung_Evaluasi(Hasil_Euclidean)
    Skor_QJSD       = Hitung_Evaluasi(Hasil_QJSD)

    // Tentukan Metrik Terbaik sebagai Baseline
    IF Skor_QJSD > Skor_Euclidean THEN
        Metrik_Terbaik = Fungsi_Kustom_QJSD
    ELSE
        Metrik_Terbaik = 'euclidean'
    END IF
```

```
    // --- TAHAP 2: Manipulasi Augmentasi (Variabel Independen 2) ---
    // Terapkan CGA pada klaster minoritas untuk menghasilkan data
       sintetis
    Dataset_Augmentasi = Terapkan_CGA(Dataset_Embeddings,
                                      Hasil_Baseline)

    // Re-clustering pada dataset yang diperkaya (Kondisi: Dengan
       Augmentasi)
    Hasil_Akhir     = HDBSCAN(Dataset_Augmentasi, metric=Metrik_Terbaik,
                              min_cluster_size=5)
    Skor_Akhir      = Hitung_Evaluasi(Hasil_Akhir)

    RETURN Metrik_Terbaik, Skor_Akhir
End Algorithm
```

Pseudocode ini menggambarkan secara eksplisit dua tahap penting penelitian. Pertama, eksperimen perbandingan metrik jarak (*Euclidean* vs QJSD) yang menentukan konfigurasi HDBSCAN paling optimal. Kedua, penerapan *Cluster-based Generative Augmentation* (CGA) yang diterapkan setelah baseline *clustering* untuk memperkuat representasi data minoritas dan menguji pengaruhnya terhadap peningkatan kualitas hasil klasterisasi.

## 2. Ruang Lingkup Penelitian

Penelitian ini berfokus pada pengelompokan foto dokumentasi berbasis wajah, di mana input berupa citra wajah yang telah diekstraksi menjadi *embeddings* menggunakan *InsightFace*. Lingkup eksperimen dibatasi pada analisis perbandingan kinerja HDBSCAN dengan dua jenis metrik jarak (*Euclidean* dan QJSD), serta pengujian efek penerapan *Cluster-based Generative Augmentation* (CGA) terhadap hasil klasterisasi.

Penelitian tidak membahas aspek lain seperti deteksi wajah multi-objek, identifikasi wajah secara eksplisit, ataupun integrasi sistem berbasis *cloud* secara *end-to-end*. Fokus utama penelitian adalah evaluasi performa *clustering* pada ruang fitur (*embedding space*) dengan kondisi *few-shot*.

### *D. Kajian Pustaka/ Penelitian Terdahulu*

Berikut beberapa penelitian sebelumnya yang memiliki relevansi dan perbedaan dengan penelitian ini:

Penelitian pertama yang menjadi landasan utama adalah skripsi Rafiul Muiz.K (2025) berjudul *"Pengelompokan Foto Otomatis Berdasarkan Identifikasi Wajah Menggunakan Algoritma Density-Based Spatial Clustering of Applications with Noise (DBSCAN)"*. Studi ini berhasil mengembangkan sistem pengelompokan foto berbasis wajah menggunakan DBSCAN dan mencapai *Silhouette Score* 0,817. Namun, ditemukan bahwa kinerja DBSCAN menurun signifikan pada dataset besar dan sangat sensitif terhadap penentuan parameter *epsilon* . Persamaannya dengan penelitian ini terletak pada tujuan membangun sistem pengelompokan foto otomatis berbasis wajah. Perbedaannya, penelitian ini beralih menggunakan algoritma HDBSCAN untuk mengatasi kelemahan DBSCAN dalam menangani variasi kepadatan klaster dan meminimalkan sensitivitas parameter.

Selanjutnya, studi oleh Peña-Asensio dan Ferrari (2025) berjudul *"Meteoroid Stream Identification with HDBSCAN Unsupervised Clustering Algorithm"* memvalidasi kemampuan HDBSCAN. Meskipun diterapkan pada domain astronomi, studi ini relevan karena menghadapi karakteristik data

yang serupa dengan *face embeddings*: berdimensi tinggi, mengandung *noise*, dan memiliki kepadatan yang bervariasi. Hasilnya menunjukkan bahwa HDBSCAN menghasilkan klaster yang lebih stabil dan *robust* dibandingkan metode tradisional. Persamaannya adalah penggunaan algoritma HDBSCAN untuk meningkatkan stabilitas klaster pada data kompleks. Perbedaannya terletak pada objek data; penelitian ini menerapkan HDBSCAN secara spesifik pada domain pengelompokan wajah (*face clustering*).

Dalam aspek pengukuran jarak, penelitian Du (2023) berjudul *"A Robust and High-Dimensional Clustering Algorithm Based on Feature Weight and Entropy"* menyoroti kelemahan metrik *Euclidean* pada data berdimensi tinggi. Du menunjukkan bahwa penggunaan jarak non-Euclidean dapat meningkatkan stabilitas algoritma terhadap *noise* . Persamaannya adalah fokus pada peningkatan akurasi klasterisasi data berdimensi tinggi melalui modifikasi pendekatan jarak. Perbedaannya, Du berfokus pada pembobotan fitur dalam *fuzzy clustering*, sedangkan penelitian ini mengadopsi metrik alternatif *Quantum Jensen-Shannon Divergence* (QJSD) dalam kerangka HDBSCAN.

Lebih spesifik mengenai metrik QJSD, penelitian Hoyos-Osorio dan Sánchez-Giraldo (2024) berjudul *"The Representation Jensen–Shannon Divergence"* memperkenalkan metrik berbasis divergensi yang mampu menangkap struktur data non-linear dengan lebih baik daripada jarak *Euclidean* . Persamaannya adalah pemanfaatan metrik berbasis divergensi untuk meningkatkan akurasi pemetaan jarak antar-sampel. Perbedaannya

terletak pada implementasinya, di mana penelitian ini mengintegrasikan QJSD secara langsung sebagai fungsi jarak kustom di dalam algoritma HDBSCAN.

Terakhir, untuk menangani kondisi *few-shot*, penelitian Haghbin dkk. (2025) mengusulkan metode *Feature-level Cluster-based Generative Augmentation* (FICAug). Metode ini bekerja di ruang fitur untuk menghasilkan sampel sintetis guna memperkaya dataset terbatas dan mencegah *overfitting*. Persamaannya adalah penggunaan pendekatan augmentasi generatif di ruang fitur untuk mengatasi keterbatasan data. Perbedaannya terletak pada penerapannya: penelitian ini menggunakan prinsip tersebut dalam metode *Cluster-based Generative Augmentation* (CGA) yang difokuskan untuk memperkuat representasi klaster wajah yang renggang agar dapat dikenali dengan lebih baik oleh HDBSCAN.

Berdasarkan kajian tersebut, penelitian ini mengisi kesenjangan yang ada dengan menggabungkan keunggulan HDBSCAN (untuk variasi kepadatan), metrik QJSD (untuk akurasi dimensi tinggi), dan augmentasi CGA (untuk kondisi *few-shot*) dalam satu sistem yang terintegrasi.

## *E. Tujuan dan Kegunaan Penelitian*

### 1. Tujuan Penelitian

Penelitian ini bertujuan untuk :

a. Merancang dan membangun sistem pengelompokan foto otomatis menggunakan algoritma HDBSCAN yang dioptimalkan melalui pemilihan metrik jarak terbaik (antara *Euclidean* dan QJSD) untuk menyelesaikan masalah data foto yang tidak terstruktur.

b. Mengimplementasikan metode augmentasi data generatif (CGA) pada sistem untuk mengatasi kegagalan pembentukan klaster pada kategori foto dengan kondisi *few-shot*, sehingga *noise* dapat diminimalisir dan stabilitas klaster meningkat.

**2. Kegunaan Penelitian**

Adapun kegunaan dari penelitian ini, baik secara teoretis maupun praktis, adalah sebagai berikut:

**a. Kegunaan Teoritis**

Penelitian ini berkontribusi pada pengembangan ilmu *machine learning*, khususnya dalam domain *unsupervised clustering* pada data berdimensi tinggi. Secara spesifik, penelitian ini memperkaya literatur dengan mengkaji efektivitas integrasi metrik jarak alternatif *Quantum Jensen-Shannon Divergence* (QJSD) dan metode augmentasi *Cluster-based Generative Augmentation* (CGA) untuk mengoptimalkan performa algoritma HDBSCAN dalam menangani variasi kepadatan dan kondisi *few-shot*.

**b. Kegunaan Praktis**

Secara praktis, sistem yang dirancang memberikan solusi otomatisasi bagi organisasi untuk mengelola dan mencari arsip foto dokumentasi secara efisien, mengurangi beban kerja manual yang tidak produktif. Selain itu, hasil penelitian ini dapat menjadi acuan teknis bagi pengembang atau peneliti lain dalam menerapkan strategi augmentasi generatif untuk mengatasi masalah keterbatasan data pada sistem pengenalan pola visual.

# BAB II

# TINJAUAN TEORITIS

## *A. Pengelompokan Foto Otomatis*

Pengelompokan foto otomatis (*automatic photo clustering*) merupakan proses pengorganisasian kumpulan foto ke dalam kelompok-kelompok berdasarkan kemiripan karakteristik visual tanpa intervensi manual. Bushey (2024) menjelaskan bahwa adopsi kecerdasan buatan telah mengubah cara lembaga arsip mengelola data visual dari proses manual menjadi sistem otomatis yang mampu mengenali pola atau identitas . Malek dkk. (2024) menambahkan bahwa penerapan *machine learning*, khususnya algoritma *clustering*, meningkatkan efisiensi pengorganisasian foto secara signifikan tanpa memerlukan pelabelan awal (*unsupervised*), yang sangat relevan untuk menangani dokumentasi dalam skala besar.

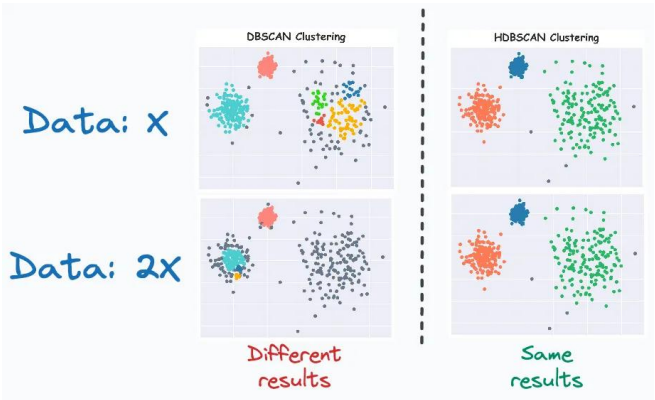
## *B. Pengenalan Wajah*

Pengenalan wajah (*face recognition*) merupakan salah satu teknik biometrik yang paling banyak digunakan dan diakui untuk mengenali individu berdasarkan keunikan fitur wajah mereka (Meena dkk., 2022). Dalam konteks manajemen arsip digital, Malek dkk. (2024) menekankan bahwa teknologi ini krusial untuk menstrukturkan data gambar melalui pengenalan pola visual secara otomatis. Proses intinya melibatkan transformasi citra menjadi *face embedding*. Bushey (2024) menjelaskan bahwa representasi *embedding* ini menjembatani data visual mentah dengan sistem analitik, memungkinkan

perbandingan antarwajah dilakukan melalui pengukuran jarak matematis di ruang fitur . Kualitas representasi ini sangat menentukan, merujuk pada McInnes dkk. (2017), keberhasilan algoritma seperti HDBSCAN dalam menemukan struktur klaster alami pada data berdimensi tinggi sangat bergantung pada fitur input yang stabil dan konsisten.

### *C. Hierarchical Density-Based Spatial Clustering of Applications with Noise*

HDBSCAN merupakan pengembangan dari algoritma DBSCAN yang diperkenalkan oleh McInnes, Healy, dan Astels (2017) untuk mengatasi kelemahan utama DBSCAN dalam menangani variasi kepadatan antar-klaster. Jika DBSCAN menggunakan satu nilai ambang batas radius (*epsilon*) yang kaku, HDBSCAN membangun struktur hierarki kepadatan yang memungkinkannya mengelompokkan data dengan tingkat densitas berbeda secara fleksibel dalam satu proses. Pendekatan ini membuat algoritma lebih stabil dan tidak memerlukan penentuan jumlah klaster secara eksplisit, berbeda dengan metode tradisional seperti *K-Means* (Mcinnes dkk., 2017).




**Gambar 2. 1**. Perbandingan Stabilitas DBSCAN vs HDBSCAN terhadap *Scaling* Dataset

Untuk mencapai stabilitas tersebut, HDBSCAN bekerja melalui transformasi ruang jarak untuk memisahkan *noise* dari klaster yang valid. Langkah fundamentalnya adalah menggantikan jarak *Euclidean* murni dengan jarak *mutual reachability*. Peña Asensio dan Ferrari (2025) menjelaskan bahwa jarak ini dihitung dengan mencari nilai maksimum antara jarak inti (*core distance*) dari dua titik yang dibandingkan dan jarak murni antara keduanya. Transformasi ini memastikan bahwa hubungan antar titik tidak hanya didasarkan pada kedekatan geometris semata, tetapi juga memperhitungkan kepadatan lokal di sekitarnya, sehingga titik-titik di area renggang akan "dijauhkan" secara matematis. Berikut rumus yang memastikan bahwa hubungan antar titik tidak hanya mempertimbangkan kedekatan geometris, tetapi juga kepadatan lokal di sekitar titik tersebut.

$$d_{\text{mreach}}(a, b) = \max\left(\text{core}_k(a), \text{core}_k(b), \text{dist}(a, b)\right)$$

### *D. Metrik Jarak dalam Clustering*

Pemilihan metrik jarak (*distance metric*) merupakan komponen fundamental dalam algoritma *clustering* karena menentukan bagaimana kemiripan antar data diukur secara kuantitatif. Jarak *Euclidean* adalah metrik yang paling umum digunakan, yang bekerja dengan mengukur panjang segmen garis lurus terpendek yang menghubungkan dua titik dalam ruang vektor. Dalam konteks analisis data, jarak ini merepresentasikan selisih geometris antar fitur objek (Dinata dkk., 2020). Namun, Du (2023) menyoroti bahwa efektivitas jarak Euclidean cenderung menurun pada data berdimensi tinggi (*face embeddings*) akibat fenomena "kutukan dimensi" (*curse of dimensionality*), di mana jarak antar titik

menjadi seragam dan gagal menangkap struktur non-linear yang kompleks, sehingga algoritma menjadi kurang stabil terhadap *noise*.

Sebagai alternatif untuk mengatasi keterbatasan tersebut, penelitian ini mengeksplorasi penggunaan *Quantum Jensen-Shannon Divergence* (QJSD). Berbeda dengan *Euclidean* yang mengukur jarak titik, QJSD adalah ukuran divergensi yang berakar pada teori informasi kuantum untuk mengukur perbedaan antara dua distribusi probabilitas atau matriks densitas. Virosztek (2021) membuktikan bahwa akar kuadrat dari QJSD memenuhi syarat matematis sebagai metrik sejati (*true metric*), sehingga valid digunakan sebagai pengganti jarak dalam algoritma geometris. Untuk penerapannya pada data vektor, digunakan konsep *Representation Jensen-Shannon Divergence* (RJSD). Hoyos Osorio & Sánchez Giraldo (2024) menjelaskan bahwa metode ini memetakan distribusi data ke dalam ruang fitur kernel (*Reproducing Kernel Hilbert Space*) dan merepresentasikannya sebagai operator kovarians. Pendekatan ini memungkinkan pengukuran perbedaan distribusi yang lebih sensitif dan mampu menangkap struktur geometris data yang lebih kaya dibandingkan pengukuran jarak lurus *Euclidean*.

### *E. Kondisi Few-Shot dalam Clustering*

Kondisi *few-shot* merujuk pada situasi di mana suatu klaster hanya memiliki sedikit sampel, menyebabkan representasi distribusi datanya menjadi tidak stabil. Peña-Asensio & Ferrari (2025) mencatat bahwa kondisi data yang tumpang tindih dan tidak seimbang dapat memicu kesalahan seperti *cluster merging* (penggabungan identitas berbeda) atau *fragmentation*. Meskipun HDBSCAN unggul dalam variasi kepadatan, kondisi *few-shot* tetap

berpotensi menurunkan kestabilan struktur hierarki. Haghbin dkk. (2025) menunjukkan bahwa kondisi ini melemahkan performa algoritma *unsupervised* karena kurangnya representasi distribusi fitur yang tegas. Oleh karena itu, penanganan khusus diperlukan untuk memperkaya titik data dalam klaster minoritas agar distribusi fitur menjadi lebih padat dan terdefinisi dengan baik.

### *F. Augmentasi Data Generatif*

Augmentasi data generatif adalah teknik untuk memperkaya *dataset* pelatihan, terutama pada kondisi *few-shot*. Haghbin dkk. (2025) mengusulkan metode *Feature-level Cluster-based Generative Augmentation* (FICAug) yang bekerja di ruang fitur untuk menghasilkan sampel sintetis guna mencegah *overfitting*. Berbeda dengan augmentasi klasik, pendekatan ini menggunakan model *deep learning* untuk menciptakan data baru yang realistis. Tan dkk. (2023) menunjukkan bahwa kemampuan ini memungkinkan pengayaan *dataset* secara fundamental, baik dalam jumlah maupun variasi fitur, sehingga menghasilkan representasi kategori yang lebih *robust* dan diskriminatif. Prinsip inilah yang diadopsi dalam penelitian ini melalui metode *Cluster-based Generative Augmentation* (CGA).

# BAB III

# METODOLOGI PENELITIAN

## *A. Jenis Penelitian*

Penelitian ini menggunakan jenis penelitian **kuantitatif**, yaitu pendekatan ilmiah yang berfokus pada pengumpulan dan analisis data numerik untuk menguji hubungan antar variabel secara objektif. Pendekatan ini menekankan aspek keterukuran, penggunaan metode statistik, serta pengujian hipotesis guna menarik kesimpulan yang terstruktur dan dapat digeneralisasi (PuTl, 2025).

Dalam konteks penelitian ini, jenis penelitian kuantitatif digunakan untuk mengukur kinerja dan efektivitas optimalisasi algoritma HDBSCAN dalam mengelompokkan foto. Data yang akan dianalisis berupa nilai kuantitatif dari metrik evaluasi *clustering*, yaitu *Silhouette Score* dan *Davies-Bouldin Index (DBI)*.

## *B. Pendekatan Penelitian*

Pendekatan yang digunakan dalam penelitian ini adalah eksperimental kuantitatif. Pendekatan ini dipilih karena penelitian akan melakukan serangkaian eksperimen terkontrol untuk menguji hipotesis mengenai pengaruh variabel independen terhadap variabel dependen. Secara spesifik, penelitian ini akan memanipulasi dua aspek utama, yaitu metrik jarak yang digunakan dalam algoritma HDBSCAN (*Euclidean* vs. QJSD) dan penerapan augmentasi data generatif (ada atau tidak adanya CGA).

Eksperimen akan dirancang untuk membandingkan hasil *clustering* (diukur secara kuantitatif menggunakan *Silhouette Score* dan DBI) antara kondisi *baseline* (HDBSCAN + *Euclidean*, tanpa augmentasi) dengan kondisi perlakuan (HDBSCAN + QJSD, dan HDBSCAN + metrik terbaik + CGA).

### *C. Sumber Data*

Sumber data yang digunakan dalam penelitian ini merupakan gabungan dari data sekunder dan data primer. Sebagian besar *dataset* diadopsi dari penelitian sebelumnya oleh Rafiul Muiz.K (2025), untuk menjaga kesinambungan basis pengujian. *Dataset* tersebut kemudian diperkaya dan diperbarui dengan data primer tambahan yang dikumpulkan secara mandiri oleh peneliti.

Secara spesifik, data mencakup kumpulan foto dokumentasi kegiatan dari dua organisasi kemahasiswaan, yaitu Inready Workgroup dan Google Developer Groups on Campus (GDGOC) UIN Alauddin Makassar. Koleksi foto tersebut disimpan dalam format digital (seperti JPG, JPEG, dan PNG) dan diakses secara terpusat melalui platform penyimpanan awan (*cloud storage*) Google Drive. Setiap foto dalam *dataset* gabungan ini merepresentasikan berbagai kegiatan organisasi dan memuat citra wajah dari anggota serta partisipan, yang menjadi objek utama untuk proses deteksi dan ekstraksi fitur (*face embeddings*).

### *D. Metode Pengumpulan dan Pengolahan Data*

Metode pengumpulan data dalam penelitian ini dilakukan dengan mengumpulkan tautan Google Drive yang berisi foto-foto dokumentasi

kegiatan dari *study club* Inready Workgroup. Proses ini melibatkan pengidentifikasian dan pemilihan foto yang mengandung wajah manusia, yang kemudian akan digunakan dalam tahap pengolahan data selanjutnya.

Proses pengolahan data dalam penelitian ini mengadopsi kerangka kerja SEMMA, yang merupakan standar proses data mining untuk memecahkan masalah melalui lima tahapan (Salsabilah dkk., 2024). Berikut adalah penerapan dari setiap tahap:

1. *Sample* (Pengumpulan Data)

Tahap ini adalah tahap awal pengumpulan data atau pengambilan sampel dari sumber data yang telah ditentukan. Data primer yang dikumpulkan adalah kumpulan foto dokumentasi kegiatan dari *study club* Inready Workgroup dan organisasi Google Developer Groups on Campus UIN Alauddin Makassar. Data ini diakses melalui tautan Google Drive yang berisi foto-foto dalam format digital (JPG, PNG, dll.).

2. *Explore* (Eksplorasi Data)

Pada tahap ini, *dataset* dieksplorasi untuk memahami karakteristik dan sebaran data, serta dilakukan pembersihan data seperti penanganan nilai yang hilang (*missing data*) atau duplikat. Dilakukan eksplorasi awal terhadap koleksi foto untuk memahami variasi visual (pencahayaan, pose, oklusi) dan mengidentifikasi potensi masalah seperti adanya individu yang hanya muncul di sedikit foto (*few-shot problem*), yang menjadi dasar untuk eksperimen augmentasi.

3. *Modify* (Modifikasi Data)

Tahap ini mencakup transformasi dan normalisasi data mentah agar data siap dan sesuai untuk digunakan dalam proses pemodelan. Persiapan data utama dilakukan di sini. Proses ini meliputi: (1) Deteksi wajah pada setiap foto menggunakan InsightFace; (2) Pemotongan (*crop*) area wajah; dan (3) Ekstraksi fitur dari setiap wajah menjadi *face embedding* berdimensi tinggi. Kumpulan *embeddings* inilah yang menjadi *dataset* final yang siap dimodelkan.

4. *Model* (Pemodelan Data)

Pada tahap ini, teknik atau algoritma data mining (seperti *clustering*) diterapkan pada data yang telah disiapkan untuk membangun model dan menemukan pola. Tahap pemodelan dalam penelitian ini adalah pelaksanaan dua skenario eksperimen. Eksperimen pertama menerapkan HDBSCAN dengan dua metrik jarak berbeda (*Euclidean* dan QJSD). Eksperimen kedua menerapkan teknik augmentasi data CGA pada klaster minoritas dan melakukan re-klasterisasi menggunakan HDBSCAN dengan metrik terbaik.

5. *Assess* (Evaluasi Data)

Tahap ini adalah penilaian kinerja model untuk mengevaluasi kualitas hasil yang diperoleh, seringkali menggunakan metrik kuantitatif seperti *Silhouette Score*. Evaluasi dilakukan secara kuantitatif dengan mengukur kualitas klaster yang dihasilkan pada setiap skenario pemodelan. Metrik yang digunakan adalah *Silhouette Score* dan *Davies-Bouldin Index* (DBI) untuk membandingkan performa antar-eksperimen dan menentukan konfigurasi metode yang paling optimal.

## *E. Instrumen Penelitian*

Instrumen penelitian yang digunakan dalam pelaksanaan studi ini meliputi perangkat keras dan perangkat lunak yang mendukung proses pengumpulan data, pengolahan, pemodelan, hingga evaluasi hasil.

### 1. Perangkat Keras

Perangkat keras utama yang digunakan untuk menjalankan eksperimen dan pengembangan sistem adalah satu unit laptop HP 14s dengan spesifikasi sebagai berikut:

a. *Processor 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz*
b. RAM 16 GB
c. SSD 512 GB

### 2. Perangkat Lunak

a. Bahasa Pemrograman Python
b. *Notebook* Google Colab
c. Sistem Operasi Windows

## *F. Teknik Pengujian dan Evaluasi Model*

### 1. Teknik Pengujian

Teknik pengujian yang diterapkan dalam penelitian ini difokuskan pada evaluasi kuantitatif terhadap efektivitas optimalisasi algoritma HDBSCAN. Pengujian dirancang untuk mengukur secara objektif bagaimana pemilihan metrik jarak (*Euclidean* vs. QJSD) dan penerapan augmentasi data generatif (CGA) mempengaruhi kualitas hasil pengelompokan *face embeddings*. Proses pengujian akan mencakup pengukuran performa

*clustering* dalam skenario perbandingan metrik jarak pada *dataset* asli dan pengukuran dampak augmentasi CGA pada kondisi *few-shot*. Evaluasi ini bertujuan untuk mengetahui sejauh mana metode yang diusulkan dapat meningkatkan akurasi dan stabilitas klasterisasi HDBSCAN, khususnya dalam menangani data berdimensi tinggi dengan kepadatan bervariasi dan keterbatasan jumlah sampel.

**2. Evaluasi Model**

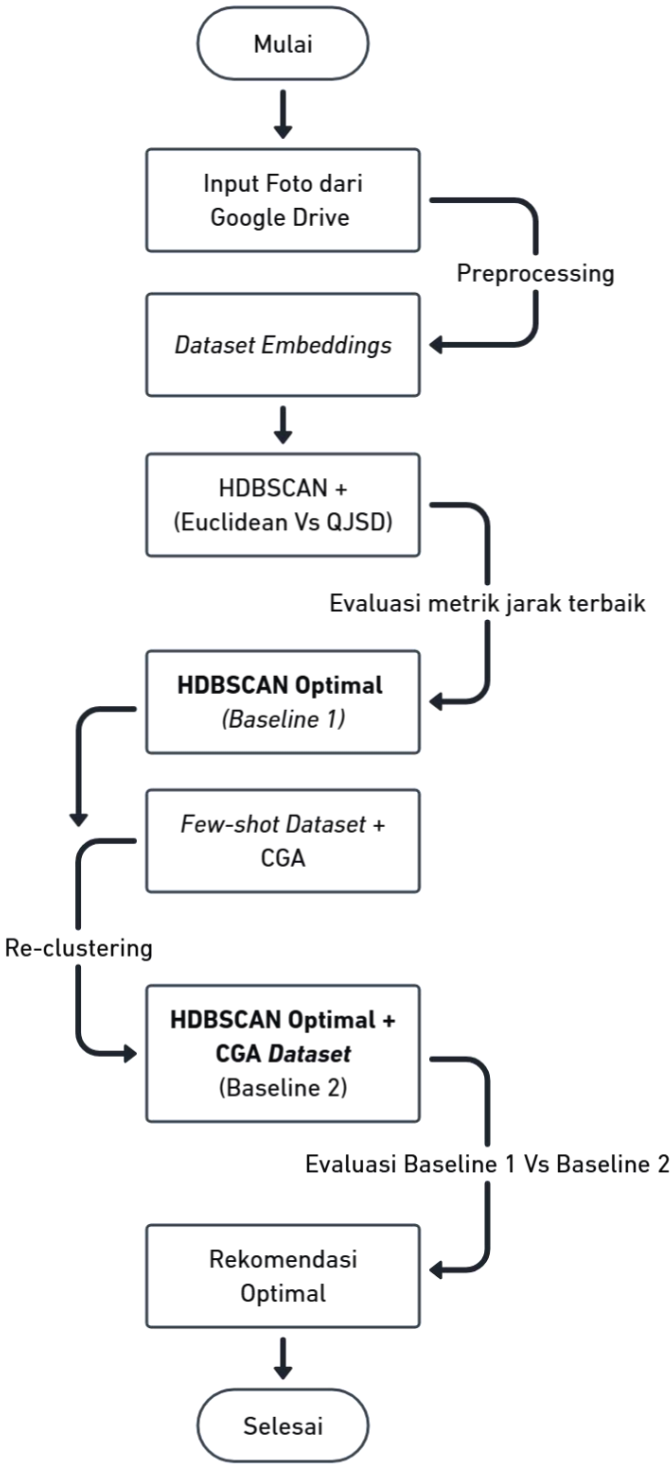
Evaluasi model dalam penelitian ini dilakukan dengan menggunakan metrik performa internal yang relevan untuk menilai kualitas struktur klaster yang dihasilkan oleh algoritma HDBSCAN pada setiap skenario eksperimen. Metrik utama yang digunakan adalah *Silhouette Score* dan *Davies-Bouldin Index (*DBI*)* .Penggunaan kedua metrik ini akan memberikan penilaian kuantitatif yang komprehensif terhadap efektivitas metode optimalisasi yang diuji.

## *G. Perancangan Sistem*

**1. Diagram Alur Sistem**

Alur kerja penelitian ini divisualisasikan pada Gambar 3.2. Proses diawali dengan input data berupa foto dari Google Drive, yang kemudian melalui tahap pra-pemrosesan untuk menghasilkan *dataset embeddings*. *Dataset* ini pertama-tama digunakan untuk mengevaluasi metrik jarak terbaik, dengan menjalankan HDBSCAN menggunakan *Euclidean* dan QJSD. Hasil evaluasi ini menentukan konfigurasi "HDBSCAN Optimal" yang menjadi Baseline 1. Selanjutnya, augmentasi CGA diterapkan pada *dataset* asli, dan hasilnya di-re-*clustering* menggunakan "HDBSCAN Optimal", menghasilkan

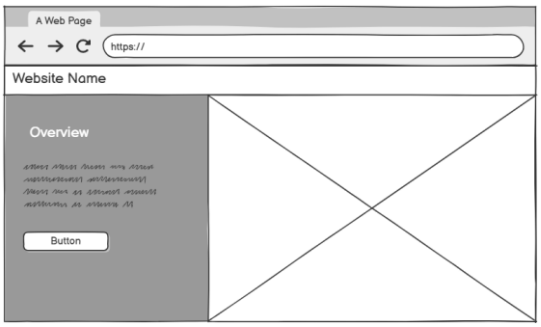
Baseline 2 (Hasil Setelah CGA). Tahap akhir adalah evaluasi perbandingan antara Baseline 1 dan Baseline 2 untuk menghasilkan rekomendasi optimal.



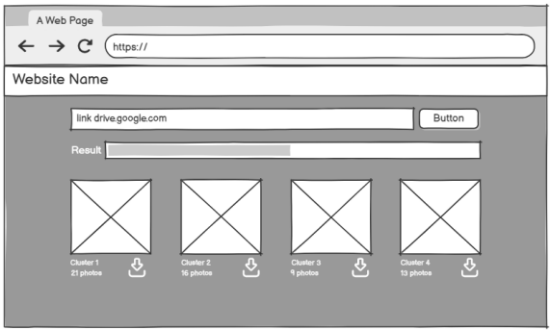

**Gambar 3. 1** *Diagram Alur* Sistem yang Diusulkan

### 2. *Wireframe*

Perancangan antarmuka sistem divisualisasikan melalui *wireframe* konseptual berbasis web untuk memberikan gambaran alur interaksi pengguna.




**Gambar 3. 2** *Wireframe* 1




**Gambar 3. 3** *Wireframe* 2

*Wireframe* pertama (Gambar 3.3) menggambarkan halaman utama atau *overview*. Halaman ini berfungsi sebagai titik masuk awal, menyajikan deskripsi singkat mengenai fungsi sistem dan menyediakan tombol "Mulai" (atau *Button*) untuk menginisiasi proses pengelompokan. *Wireframe* kedua (Gambar 3.4) menampilkan halaman interaksi utama. Di bagian atas, terdapat kolom input untuk memasukkan tautan Google Drive dan tombol untuk memulai pemrosesan.

## *H. Estimasi Waktu Penelitian*

Penelitian ini direncanakan akan dilaksanakan dalam kurun waktu kurang lebih enam bulan. Jadwal ini bersifat estimasi dan dapat disesuaikan kembali berdasarkan perkembangan dan dinamika yang terjadi selama proses penelitian.

**Tabel 3. 1** Estimasi Waktu Penelitian

| **No** | **Kegiatan** | **Bulan** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **5** | **6** |
| 1 | Pengumpulan Data & Studi Literatur | | | | | | |
| 2 | Analisa Awal & Persiapan Data | | | | | | |
| 3 | Perancangan & Pemodelan Eksperimen 1 | | | | | | |
| 4 | Pemodelan Eksperimen 2 (Augmentasi) | | | | | | |
| 5 | Pengujian & Evaluasi Hasil | | | | | | |
| 6 | Penyusunan Laporan & Revisi | | | | | | |

## Daftar Pustaka


Bushey, J. (2024). Envisioning Archival Images with Artificial Intelligence The growing volumes of images in archival institutions reflect our global shift. *Archeion*, *125*, 33–54. https://doi.org/10.4467/26581264ARC.24.007.20202

Che, Q. H., Le, D.-T., Pham, B. N., Lam, D.-K., & Nguyen, V. T. (2025). *Enhanced Generative Data Augmentation for Semantic Segmentation via Stronger Guidance*.

Dinata, R. K., Novriando, H., Hasdyna, N., & Retno, S. (2020). *Reduksi Atribut Menggunakan Information Gain untuk Optimasi Cluster Algoritma K-Means*. *6*(1), 48–53.

Du, X. (2023). A Robust and High-Dimensional Clustering Algorithm Based on Feature Weight and Entropy. *Entropy*, *25*(3). https://doi.org/10.3390/e25030510

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*.

Fortın, C. E. L., & Nishikawa, I. (2024). *Generative Data Augmentation for Few-Shot Domain Adaptation*. *Icpram*, 256–265. https://doi.org/10.5220/0012338000003654

Haghbin, Y., Moradi, H., & Hosseini, R. (2025). *Feature-to-Image Data Augmentation: Improving Model Feature Extraction with Cluster-Guided Synthetic Samples*. 1–10. http://arxiv.org/abs/2409.17685

Hoyos Osorio, J. K., & Sanchez Giraldo, L. G. (2024). *The Representation Jensen-Shannon Divergence*. http://arxiv.org/abs/2305.16446

Kalpavruksha, R., Kalpavruksha, R., Cha, T., & Cha, S.-H. (2025). Kernel Density Based Spatial Clustering of Applications with Noise. *The International FLAIRS Conference Proceedings*, *38*, 2–7. https://doi.org/https://doi.org/10.32473/flairs.38.1.138998

Kementerian Agama RI. (2022a). *Qur'an Kemenag*. https://quran.kemenag.go.id/quran/per-ayat/surah/10?from=5&to=5

Kementerian Agama RI. (2022b). *Qur'an Kemenag*. https://quran.kemenag.go.id/quran/per-ayat/surah/7?from=31&to=31

Malek, W. A., A. Jalil, S., Rahman, A., Kamaruddin, I., Abu, R., Ismail, S. A., Mansoor, M., Mokhtarudin, M., Norsuriati, N., Safuan, N., & Roslan, R. N. H. (2024). Artificial Intelligence and Archive Management on Malaysia's National Archive's Uncaptioned Photos Collection: Accuracy findings comparison based on clustering algorithms. *Environment-Behaviour Proceedings Journal*, *9*, 159–164.

https://doi.org/https://doi.org/10.21834/e-bpj.v9iSI18.5478

Mcinnes, L., & Healy, J. (2017). *Accelerated Hierarchical Density Clustering*. 1–32.

Mcinnes, L., Healy, J., & Astels, S. (2017). *hdbscan : Hierarchical density based clustering*. *2*(2017), 11–12. https://doi.org/10.21105/joss.00205

Meena, S., Vats, K., & Kumar, D. (2022). *Review of factors affecting facial recognition algorithms performance*. *6*(March), 9528–9541.

Muiz.K, R. (2025). *Pengelompokan Foto Otomatis Berdasarkan Identifikasi Wajah Menggunakan Algoritma Density Based Spatial Clustering of Application with Noise*.

Peña Asensio, E., & Ferrari, F. (2025). Meteoroid Stream Identification with HDBSCAN Unsupervised Clustering Algorithm. *The Astronomical Journal*, *170*(3), 140. https://doi.org/10.3847/1538-3881/adec8c

PuTl. (2025). *Penelitian Kuantitatif: Pengertian, Tujuan, Jenis dan Tahapanya*. https://dim.telkomuniversity.ac.id/penelitian-kuantitatif-pengertian-tujuan-jenis-dan-tahapanya/

Salsabilah, A. F., Zamzani, Z. M., & Gustrifa, W. (2024). Implwmwntasi Data Mining Menggunakan Metode K-Means Clustering Untuk Analisis Tingkat Pengangguran Terbuka di Jawa Timur. *JIFoSI*, *5*(3), 1–11.

Shihab, M. Q. (2002a). *Tafsir Al-Mishbah: Pesan, Kesan dan Keserasian Al-Qur'an (Vol. 5)*. Lentera Hati.

Shihab, M. Q. (2002b). *Tafsir Al-Mishbah: Pesan, Kesan dan Keserasian Al-Qur'an (Vol. 6)*. Lentera Hati.

Shin, J., Lee, H. J., Kim, H., Baek, J. H., Kim, D., & Jun, Y. (2023). Local Connectivity-Based Density Estimation for Face Clustering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2023-June*, 13621–13629. https://doi.org/10.1109/CVPR52729.2023.01309

Tan, W., Chen, S., & Yan, B. (2023). *DifFSS: Diffusion Model for Few-Shot Semantic Segmentation*. http://arxiv.org/abs/2307.00773

Virosztek, D. (2021). The metric property of the quantum Jensen-Shannon divergence. *Advances in Mathematics*, *380*(846294), 107595. https://doi.org/10.1016/j.aim.2021.107595