

**PENERAPAN *MACHINE LEARNING* UNTUK
PENGELOMPOKAN GALAKSI SERTA
PEMODELAN PROFIL GALAKSI *REDSHIFT*
TINGGI**

TESIS

Karya tulis sebagai salah satu syarat
untuk memperoleh gelar Magister dari
Institut Teknologi Bandung

Oleh

**MUHAMMAD NUR IHSAN EFFENDI
NIM 20324009**



**PROGRAM STUDI MAGISTER ASTRONOMI
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT TEKNOLOGI BANDUNG
2025**

PENGESAHAN

Penerapan *Machine Learning* untuk Pengelompokan Galaksi serta Pemodelan Profil Galaksi *Redshift* Tinggi

Oleh
Muhammad Nur Ihsan Effendi
NIM 20324009

Program Studi Magister Astronomi
Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Teknologi Bandung

Bandung, 13 Agustus 2025

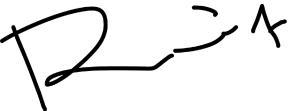
Menyetujui,

Pembimbing 1



Dr. Lucky Puspitarini
NIP 117110015

Pembimbing 2



Dr. Abdurrouf

Tim Pengaji:

1. Prof. Dra. Premana Wardayanti Premadi, Ph.D.
2. Dr. rer. nat. Mochamad Ikbal Arifyanto
3. Dr. M. Irfan Hakim

PEDOMAN PENGGUNAAN

BUKU TUGAS AKHIR

Buku Tugas Akhir Sarjana ini tidak dipublikasikan, namun terdaftar dan tersedia di Perpustakaan Institut Teknologi Bandung. Buku ini dapat diakses umum, dengan ketentuan bahwa penulis memiliki hak cipta dengan mengikuti aturan HaKI yang berlaku di Institut Teknologi Bandung. Referensi kepustakaan diperkenankan dicatat, tetapi pengutipan atau peringkasan hanya dapat dilakukan seizin penulis, dan harus disertai dengan kebiasaan ilmiah untuk menyebutkan sumbernya.

Memperbanyak atau menerbitkan sebagian atau seluruh buku Tugas Akhir harus atas izin Program Studi Sarjana Astronomi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung.

*This book was written in the midst of a genocide unfolding in Gaza
— let that not be forgotten.*

KATA PENGANTAR

Alhamdulillah, segala puji hanya bagi Allah SWT. karena atas izin-Nya penulis dapat menyelesaikan penelitian tesis ini untuk memperoleh gelar Magister Sains. Tiada daya dan upaya selain hanya dengan pertolongan Allah, penelitian berjudul *Penerapan Machine Learning Untuk Pengelompokan Galaksi Serta Pemodelan Profil Galaksi Redshift Tinggi* dapat terselesaikan tepat pada waktunya. *Shalawat* serta salam semoga selalu tercurahkan kepada Nabi Muhammad SAW. yang telah membawa umat manusia dari zaman kebodohan ke dalam cahaya iman.

Tesis ini disusun sebagai salah satu syarat kelulusan dari Program Studi Magister Astronomi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung. Semoga tesis ini dapat menjadi jalan tersampaikannya ilmu pengetahuan kepada para pembaca yang membutuhkannya. Sesungguhnya setiap kata yang dituliskan dalam buku ini berasal dari dan hanya dari Allah Yang Maha Pandai, melalui perantara guru-guru, dosen, dan banyak pihak lainnya. Oleh karena itu, dalam tulisan ini penulis ingin mengucapkan terima kasih dan apresiasi kepada:

1. Ibunda terkasih, Ibu Fillah, yang telah mendukung penulis secara moril dan materi.
2. Dosen pembimbing, Ibu Dr. Lucky Puspitarini, yang dengan sabar senantiasa memberi arahan kepada penulis dalam mengerjakan penelitian ini.
3. Dosen pembimbing kedua, Bapak Dr. Abdurrouf, yang senantiasa memberi saran dan masukan terhadap penelitian ini.
4. Prof. Premana Wardayanti Premadi, Ph.D., Dr. rer. nat. Mochamad Ikbal Arifyanto, dan Dr. Muhamad Irfan Hakim, selaku dosen penguji terhadap tesis ini. Penulis ucapkan terima kasih banyak atas ilmu pengetahuan, saran, serta masukan yang diberikan oleh Ibu dan Bapak.
5. Dr. Anton Timur Jaelani, Dr. Irham Taufik Andika, Dr. rer. nat. Hesti Retno Tri Wulandari, Dr. Itsna Khoirul Fitriana, Dr. Dian Puspita Triani, dan Dr. Ibnu Nurul Huda, selaku pembimbing dan mentor dalam tim riset yang diikuti oleh penulis dalam dua tahun terakhir. Terima kasih atas semua masukan dan arahan yang diberikan untuk menjadikan penelitian ini menjadi lebih baik.
6. Bapak ibu dosen dan seluruh staf Program Studi Astronomi ITB, atas semua ilmu pengetahuan yang telah diajarkan kepada penulis selama empat

tahun terakhir. Terima kasih kepada seluruh staf atas jasa administratif yang telah diberikan kepada penulis.

7. Penulis ucapkan terima kasih kepada keluarga, Bapak Ridwan Effendi, Hafidz Effendi-Utami Aria Puteri, Rindu Nurani-Andrian Nurahayu, Hifdhan Hasan Effendi-Riska Ayu Lestari, dan Febriani Amalina Shalihah, yang telah memberi inspirasi dan pelajaran kepada penulis selama ini.

8. Rekan-rekan mahasiswa tim riset penulis, Sultan Hadi Kusuma S.Si., Novan Saputra Haryana S.Si., dan Ryo Albert Sutanto S.Si., yang senantiasa bersama penulis selama penelitian ini.

9. Rekan-rekan PPSM angkatan 2020 + Akmal Husain S.Si., yang senantiasa bersama penulis di saat sulit dan di saat yang sangat sulit.

10. Sahabat-sahabat penulis, Jovidia Laviosa S.Kom yang telah memberi banyak masukan terkait metode *machine learning*, Widyana Khoerunnisa S.KG, Nadilla Z. Huwaida S.T., Karin Hanasyanila Salma S.Ars., dan seluruh teman-teman grup Hisahito, atas waktu dan kebersamaan di sela-sela kesibukan masing-masing.

11. Semua guru dan pengajar sejak TK hingga SMA yang telah membentuk penulis hingga berada di titik ini.

12. Pihak-pihak lainnya yang secara disadari maupun tidak disadari, telah mendukung penulisan tesis ini.

Semoga setiap ilmu pengetahuan, kebaikan, dan jasa yang diberikan oleh Bapak, Ibu, dan rekan-rekan semuanya tercatat sebagai amal kebaikan yang tidak pernah terputus.

Penulis menyadari akan ketidak sempurnaan dalam penulisan tesis ini. Oleh karena itu, penulis memohon maaf atas kesalahan yang mungkin ditemukan dalam buku ini, dan besar harapan agar pembaca sekalian dapat memberikan kritik dan saran terhadap tesis ini. Semoga keberkahan selalu dilimpahkan kepada penulis maupun pembaca sekalian.

Bandung, 29 Juli 2025

Muhammad Nur Ihsan Effendi
NIM 20324009

ABSTRAK

Pengamatan JWST membuka ruang baru dalam pengamatan galaksi *redshift* tinggi. Tinjauan terhadap aspek morfologi dan ukuran galaksi pada rentang *redshift* yang lebar dapat memberi pemahaman terkait proses evolusi galaksi. Di sisi lain, pembelajaran mesin dalam beberapa dekade terakhir terbukti lebih efisien dalam memproses data dengan jumlah yang sangat besar. Pada penelitian ini, akan dilakukan pengelompokan galaksi *redshift* tinggi dan dilakukan validasi berdasarkan informasi morfologi galaksi *redshift* dekat. Pada penelitian ini digunakan metode *Variational Autoencoder* (VAE) untuk mengekstrak fitur-fitur utama dari data galaksi, dan pengelompokan dilakukan menggunakan metode *Hierarchical* dan *K-Means clustering*. Hasil pengelompokan akan dibandingkan dengan beberapa parameter galaksi di dalam katalog *EAZY* dan dari hasil *fitting* parametrik galaksi dengan *multi Sérsic*. Dengan menggunakan metode tersebut, kami dapat membuat rekonstruksi data galaksi yang lebih halus. Kami menemukan bahwa faktor *noise* pada citra galaksi sangat memengaruhi proses pengelompokan galaksi berdasarkan morfologinya. Tetapi secara umum, arsitektur VAE yang dibangun pada penelitian ini dapat mengekstrak data galaksi hanya dalam 10 parameter laten, dan di dalamnya tersimpan informasi struktur fisis galaksi. Di sisi lain, hasil *fitting* parametrik galaksi memberikan informasi radius efektif galaksi, dan kami menemukan bahwa tren radius efektif terhadap massa memiliki *slope* yang lebih rendah untuk galaksi *star-forming* dibandingkan untuk galaksi *quiescent*. Radius efektif galaksi *star-forming* tampak semakin kecil pada *redshift* yang semakin tinggi. Hubungan radius efektif galaksi terhadap panjang gelombang menunjukkan penurunan seiring dengan panjang gelombang yang semakin panjang.

Kata kunci: morfologi galaksi, *machine learning*.

ABSTRACT

James Webb Space Telescope (JWST) has opened new frontiers in the observation of high-redshift galaxies. Investigations into the morphology and size of galaxies across a wide redshift range can provide valuable insights into the processes of galaxy evolution. On the other hand, machine learning has proven to be more efficient for processing big data. In this study, we perform clustering of high-redshift galaxies and validate the results using morphological information from low-redshift galaxies. We employ the Variational Autoencoder (VAE) method to extract the main features from galaxy photometric data, and clustering is conducted using Hierarchical and K-Means algorithms. The clustering results are then compared with several galaxy parameters available in the EAZY catalog and the outcomes of parametric galaxy fitting using multi-Sérsic models. Using these methods, we are able to produce smoother reconstructions of galaxy data. We find that noise in galaxy images significantly affects the clustering process. However, the VAE architecture developed in this study is capable of extracting galaxy data into just 10 latent parameters, which contain information about the physical structure of galaxies. Additionally, the parametric fitting results provide information on galaxy effective radii. We find that the trend of effective radius versus mass has a shallower slope for star-forming galaxies compared to quiescent galaxies. The effective radius of star-forming galaxies clearly decreases at higher redshifts. Furthermore, the relationship between effective radius and wavelength shows a decreasing trend with increasing wavelength.

Key words: galaxy morphologies, machine learning.

DAFTAR ISI

PENGESAHAN	i
PEDOMAN PENGGUNAAN BUKU TUGAS AKHIR	ii
DEDIKASI	iii
KATA PENGANTAR	iv
ABSTRAK	vi
ABSTRACT	vii
DAFTAR ISI	viii
DAFTAR TABEL	x
DAFTAR GAMBAR	xi
I PENDAHULUAN	1
I.1 Latar Belakang	1
I.2 Rumusan dan Batasan Masalah	3
I.3 Tujuan	4
I.4 Metodologi	4
I.5 Sistematika Penulisan	5
II TEORI DASAR	6
II.1 Galaksi	6
II.1.1 Morfologi Galaksi	6
II.1.2 Teori Pembentukan Galaksi	9
II.1.3 Klasifikasi Galaksi Berdasarkan Pembentukan Bintang di Dalamnya	12
II.2 <i>Machine Learning</i>	14
II.2.1 <i>Convolutional Neural Network</i>	16
II.2.2 <i>Variational Autoencoder</i>	19
II.2.3 <i>Clustering</i>	21
II.2.4 <i>Dimensionality Reduction</i>	23
III DATA DAN METODE PENELITIAN	25
III.1 Data	28
III.1.1 Data Utama	28
III.1.2 <i>Field Pengamatan</i>	29
III.1.3 Katalog <i>EAZY</i>	33

III.1.4	Katalog Galaksi Zoo	37
III.2	Seleksi Data JWST	37
III.2.1	Seleksi Data Secara Umum	37
III.2.2	Seleksi Data Berdasarkan Pembentukan Bintang di Da- lam Galaksi	41
III.3	Pengolahan Data JWST	43
III.3.1	Pembuatan <i>Cutout Images</i>	44
III.3.2	Pemodelan Parametrik	44
III.3.3	<i>Pre-Processing Data</i>	45
III.3.4	Proses Pengkodean (<i>Encoding</i>)	49
III.3.5	Pengelompokan Morfologi Galaksi Menggunakan Meto- de <i>Unsupervised Learning</i>	55
IV HASIL DAN ANALISIS		57
IV.1	Hasil Pemodelan Parametrik	57
IV.1.1	Hubungan Massa Terhadap Radius Efektif	59
IV.1.2	Hubungan Radius Efektif Terhadap <i>Redshift</i>	61
IV.1.3	Hubungan Radius Efektif Terhadap Panjang Gelombang	62
IV.1.4	Rapat Jumlah Galaksi Berdasarkan Radius Efektif . . .	63
IV.2	Pengelompokan Galaksi Dekat	66
IV.2.1	Hasil Autoencoding	66
IV.2.2	<i>Hierarchical Clustering</i>	70
IV.2.3	<i>K-Means Clustering</i>	74
IV.2.4	PCA dan UMAP	78
IV.3	Pengelompokan Galaksi Jauh	80
IV.3.1	Pengelompokan Galaksi Jauh Secara Global	81
IV.3.2	Pengelompokan Galaksi Jauh pada Berbagai Rentang <i>Redshift</i>	109
V SIMPULAN DAN SARAN		131
V.1	Simpulan	131
V.2	Saran	133
DAFTAR PUSTAKA		135
LAMPIRAN		138
A Visualisasi Parameter Laten dari Galaksi Dekat		139
A.1	Visualisasi 100 Parameter Laten dari Galaksi Dekat	139

DAFTAR TABEL

III.1 Jumlah data setelah dilakukan seleksi <i>redshift</i> dan magnitudo . . .	38
IV.1 Asumsi klaster-klaster yang berisi galaksi tipe <i>spheorid</i> dan <i>irregular</i>	127
IV.2 Nilai akurasi dan presisi galaksi hasil <i>clustering</i> dibandingkan morfologi dari inspeksi visual dari penelitian Effendi (2024) . . .	130

DAFTAR GAMBAR

II.1	Diagram <i>Hubble tuning-fork</i>	7
II.2	Skema klasifikasi galaksi de-Vaucouleurs	8
II.3	Contoh galaksi <i>irregular</i> pada <i>redshift</i> dekat	9
II.4	Skema dua teori pembentukan galaksi	10
II.5	Skema pembentukan galaksi elips dan galaksi spiral	11
II.6	Diagram UVJ untuk mengklasifikasikan galaksi <i>star forming</i> dan galaksi <i>quiescent</i>	12
II.7	Kurva laju pembentukan bintang terhadap massa untuk menyeleksi galaksi <i>star forming</i> dan <i>quiescent</i>	13
II.8	Diagram yang menunjukkan hubungan AI, ML, dan <i>deep learning</i> .	15
II.9	Sampel Galaksi-CEERS481	17
II.10	Visualisasi <i>output convolution layer</i>	17
II.11	Beberapa contoh fungsi aktivasi	18
II.12	Visualisasi <i>output activation layer</i>	18
II.13	Visualisasi <i>Output Max Pooling Layer</i>	19
II.14	Contoh penggunaan PCA	24
III.1	Diagram alur penelitian	26
III.2	Sebagian penjelasan diagram alur pada Gambar III.1	27
III.3	Sensitifitas instrumen JWST	28
III.4	Area survei CEERS	30
III.5	Area survei COSMOS-Web	31
III.6	Area survei FRESCO	32
III.7	Area survei PRIMER-UDS	33
III.8	Plot distribusi massa terhadap <i>redshift</i> untuk data dari keempat survei	34
III.9	Perbandingan nilai <i>redshift</i> dari pengukuran fotometri dan dari pengukuran spektroskopi	35
III.10	Distribusi magnitudo galaksi dari data katalog <i>EAZY</i>	36
III.11	Perbandingan data fotometri galaksi di sekitar magnitudo 27 . .	38
III.12	Sampel data bukan galaksi	39
III.13	Sampel data dengan hasil <i>fitting</i> Galftim yang kurang akurat . .	40

III.14 Sampel data galaksi yang tidak terdeteksi oleh Galclean	41
III.15 Seleksi data berdasarkan diagram warna pada penelitian ini	42
III.16 Seleksi data berdasarkan diagram hubungan SFR dan massa	43
III.17 Citra galaksi setelah melalui proses pembersihan menggunakan Galclean	46
III.18 Sampel galaksi yang tidak halus dibersihkan oleh aplikasi Galclean	47
III.19 Sampel galaksi sebelum dan setelah melalui proses rotasi ber- dasarkan nilai <i>Position Angle</i>	48
III.20 Perbandingan galaksi sebelum dan setelah proses <i>rescaling</i> dan pemangkasan	49
III.21 Skema proses <i>variational autoencoder</i>	50
III.22 Arsitektur <i>encoder</i> untuk galaksi jauh	51
III.23 Arsitektur <i>decoder</i> untuk galaksi jauh	52
III.24 Arsitektur <i>encoder</i> untuk galaksi dekat	54
III.25 Arsitektur <i>decoder</i> untuk galaksi dekat	55
IV.1 Contoh hasil <i>fitting multi Sérsic</i>	58
IV.2 Contoh hasil <i>fitting single Sérsic</i>	59
IV.3 Plot hubungan radius efektif galaksi terhadap massa	60
IV.4 Plot hubungan radius efektif galaksi terhadap <i>redshift</i>	61
IV.5 Plot hubungan radius efektif galaksi terhadap panjang gelombang	62
IV.6 Plot rapat jumlah galaksi pada berbagai radius efektif	65
IV.7 Plot nilai <i>loss</i> seiring bertambahnya <i>epoch</i> selama proses <i>auto- encoding</i> untuk galaksi dekat	67
IV.8 Total variansi data galaksi dekat untuk berbagai jumlah <i>prin- ciple components</i>	68
IV.9 Plot perbandingan <i>reconstructed images</i> terhadap data input dan residualnya, untuk galaksi dekat dengan 100 laten	69
IV.10 Plot perbandingan <i>reconstructed images</i> galaksi dekat dengan 10 laten	69
IV.11 Dendrogram hasil <i>hierarchical clustering</i> untuk galaksi dekat . .	70
IV.12 Hasil pengelompokan galaksi dekat dengan metode <i>hierarchical clustering</i> ke dalam 2 klaster	71
IV.13 Sebaran data galaksi berdasarkan morfologi katalog terhadap hasil <i>hierarchical clustering</i>	72
IV.14 Sampel galaksi dekat hasil pengelompokan dengan metode <i>hie- archical clustering</i> ke dalam 10 klaster	73

IV.15	<i>Elbow plot</i> untuk menentukan jumlah klaster pada pengelompokan galaksi dekat	74
IV.16	Hasil pengelompokan galaksi dekat dengan metode <i>K-Means clustering</i> ke dalam 2 klaster	75
IV.17	Sebaran data galaksi berdasarkan morfologi katalog terhadap hasil <i>K-Means clustering</i>	76
IV.18	<i>Confusion matrix</i> pengelompokan galaksi dekat dengan metode <i>k-means clustering</i>	77
IV.19	Proyeksi parameter laten dalam dua dimensi menggunakan PCA untuk galaksi dekat	78
IV.20	Proyeksi parameter laten dalam dua dimensi menggunakan UMAP untuk galaksi dekat	79
IV.21	Hasil klastering dengan metode <i>k-means</i> yang tampak dari proyeksi parameter laten untuk galaksi dekat	80
IV.22	Distribusi galaksi elips dan spiral dari representasi parameter laten menggunakan UMAP	80
IV.23	Plot nilai <i>loss</i> seiring bertambahnya <i>epoch</i> selama proses <i>auto-encoding</i> untuk galaksi jauh	82
IV.24	Plot perbandingan <i>reconstructed images</i> untuk 10 parameter laten terhadap data input dan residualnya, pada galaksi jauh	83
IV.25	Plot perbandingan <i>reconstructed images</i> untuk 100 parameter laten terhadap data input dan residualnya, pada galaksi jauh	84
IV.26	Perbandingan nilai <i>loss</i> setelah 50 <i>epoch</i> untuk beberapa nilai dimensi laten	85
IV.27	Nilai total variansi untuk berbagai jumlah <i>principle components</i>	86
IV.28	Visualisasi parameter laten	87
IV.29	Matriks korelasi parameter laten terhadap parameter galaksi	88
IV.30	Sampel galaksi kategori <i>disk</i> dengan galaksi lain yang memiliki <i>similarity cosine</i> > 0.9	89
IV.31	Sampel galaksi kategori <i>spheroid</i> dengan galaksi lain yang memiliki <i>similarity cosine</i> > 0.95	90
IV.32	Sampel galaksi kategori <i>irregular</i> dengan galaksi lain yang memiliki <i>similarity cosine</i> > 0.8	91
IV.33	Distribusi beberapa parameter galaksi terhadap tiga sampel kategori galaksi	92
IV.34	Perbandingan nilai rata-rata <i>cosine similarity</i> masing-masing klaster terhadap jumlah klaster	93

IV.35Dendrogram hasil <i>hierarchical clustering</i> untuk galaksi jauh	94
IV.36Sampel galaksi hasil pengelompokan dengan metode <i>hierarchical clustering</i> untuk klaster berwarna jingga	95
IV.37Sampel galaksi hasil pengelompokan dengan metode <i>hierarchical clustering</i> untuk klaster berwarna hijau	96
IV.38Sampel galaksi hasil pengelompokan dengan metode <i>hierarchical clustering</i> untuk klaster berwarna merah	96
IV.39Sampel galaksi hasil pengelompokan dengan metode <i>hierarchical clustering</i> untuk klaster berwarna ungu	97
IV.40Heatmap cosine similarity antar klaster	98
IV.41Elbow plot untuk menentukan jumlah klaster pada pengelompokan galaksi jauh	99
IV.42Hasil pengelompokan galaksi jauh dengan metode <i>k-means clustering</i> ke dalam 4 klaster	100
IV.43Heatmap cosine similarity antar klaster dari metode <i>K-Means Clustering</i>	101
IV.44Proyeksi parameter laten dalam dua dimensi menggunakan PCA	102
IV.45Proyeksi parameter laten dalam dua dimensi menggunakan UMAP	103
IV.46Hasil klastering dengan metode <i>K-Means</i> yang tampak dari proyeksi parameter laten	104
IV.47Hubungan parameter galaksi terhadap proyeksi parameter laten dalam dua dimensi dengan PCA	106
IV.48Hubungan parameter galaksi terhadap proyeksi parameter laten dalam dua dimensi dengan UMAP	108
IV.49Distribusi nilai rata-rata <i>signal-to-noise</i> dari data tiga filter	109
IV.50Distribusi <i>redshift</i> galaksi setelah dilakukan seleksi S/N> 20	110
IV.51Dendrogram galaksi pada rentang $z_{spec} \leq 2$	111
IV.52Sampel galaksi dari masing-masing klaster pada rentang $z_{spec} \leq 2$	112
IV.53Heatmap cosine similarity antarklaster untuk galaksi pada rentang $z_{spec} \leq 2$	113
IV.54Distribusi beberapa parameter galaksi terhadap klaster pada rentang $z_{spec} \leq 2$	114
IV.55Dendrogram galaksi pada rentang $2 < z_{spec} \leq 3$	115
IV.56Sampel galaksi dari masing-masing klaster pada rentang $2 < z_{spec} \leq 3$	116
IV.57Heatmap cosine similarity antarklaster untuk galaksi pada rentang $2 < z_{spec} \leq 3$	117

IV.58	Distribusi beberapa parameter galaksi terhadap klaster pada rentang $2 < z_{spec} \leq 3$	118
IV.59	Dendrogram galaksi pada rentang $3 < z_{spec} \leq 4$	119
IV.60	Sampel galaksi dari masing-masing klaster pada rentang $3 < z_{spec} \leq 4$	120
IV.61	<i>Heatmap cosine similarity</i> antarklaster untuk galaksi pada rentang $3 < z_{spec} \leq 4$	121
IV.62	Distribusi beberapa parameter galaksi terhadap klaster pada rentang $3 < z_{spec} \leq 4$	122
IV.63	Dendrogram galaksi pada rentang $z_{spec} > 4$	123
IV.64	Sampel galaksi dari masing-masing klaster pada rentang $z_{spec} > 4$	124
IV.65	<i>Heatmap cosine similarity</i> antarklaster untuk galaksi pada rentang $z_{spec} > 4$	125
IV.66	Distribusi beberapa parameter galaksi terhadap klaster pada rentang $z_{spec} > 4$	126
IV.67	Distribusi fraksi galaksi hasil pengelompokan yang dikategorikan secara visual pada berbagai <i>redshift</i>	128
IV.68	Distribusi parameter galaksi di berbagai <i>redshift</i>	129
A.1	Visualisasi parameter laten 1 hingga 10	139
A.2	Visualisasi parameter laten 11 hingga 20	140
A.3	Visualisasi parameter laten 21 hingga 30	140
A.4	Visualisasi parameter laten 31 hingga 40	141
A.5	Visualisasi parameter laten 41 hingga 50	141
A.6	Visualisasi parameter laten 51 hingga 60	142
A.7	Visualisasi parameter laten 61 hingga 70	142
A.8	Visualisasi parameter laten 71 hingga 80	143
A.9	Visualisasi parameter laten 81 hingga 90	143
A.10	Visualisasi parameter laten 91 hingga 100	144

BAB I

PENDAHULUAN

I.1 Latar Belakang

Galaksi merupakan *building blocks* alam semesta (Mo dkk., 2010). Hubble menemukan konsep alam semesta yang mengembang ketika mengamati gerak galaksi-galaksi yang menjauh. Dengan melihat kurva rotasi galaksi, kita mendapat petunjuk akan keberadaan materi gelap. Melalui survei galaksi, astronom dapat melihat jejak perturbasi densitas alam semesta dini melalui *baryon acoustic oscillation*. Dengan demikian, penelitian tentang galaksi menjadi hal yang krusial dalam memahami alam semesta.

Galaksi merupakan kumpulan bintang, gas, dan debu yang terikat secara gravitasi. Banyak aspek yang dapat dianalisis dari sebuah maupun sekumpulan galaksi. Berdasarkan pembentukan bintang di dalamnya, galaksi dapat diklasifikasikan ke dalam kategori galaksi *star forming* (galaksi yang aktif membentuk bintang) dan galaksi *quiescent* (galaksi yang telah berhenti membentuk bintang). Analisis ini lebih jauh mengarah pada proses evolusi galaksi. Beberapa penelitian yang spesifik membahas kategori galaksi ini diantaranya Speagle dkk. (2014) dan Valentino dkk. (2023). Sementara itu, galaksi juga dapat ditinjau dari bentuknya yang beragam. Klasifikasi morfologi galaksi ini telah dilakukan seperti pada Hubble (1926) dan de Vaucouleurs (1959). Analisis tentang morfologi galaksi juga menjadi bagian penting, meski bentuk galaksi yang diamati hanya menggambarkan salah satu fase dari proses evolusi yang dialami galaksi tersebut.

Pengembangan *machine learning* dan *artificial intelligence* beberapa dekade terakhir membuka metode baru dalam pengelompokan morfologi galaksi, seperti yang dilakukan oleh Dieleman dkk. (2015) menggunakan data Galaxy Zoo, dan juga Tohill dkk. (2024) menggunakan data pengamatan terbaru *James Webb Space Telescope* (JWST). Pengkategorian morfologi galaksi pada alam semesta lokal sudah lebih mapan dibandingkan morfologi galaksi jauh. Hal ini tentunya karena instrumen pengamatan yang belum memadai untuk melakukan pengamatan galaksi di *redshift* tinggi. Karena belum adanya pengelompokan yang jelas terhadap morfologi galaksi *redshift* tinggi, beberapa

penelitian melakukan pengelompokan menggunakan metode *unsupervised learning* seperti yang dilakukan Tohill dkk. (2024).

Selain bentuknya, beberapa penelitian meninjau galaksi dari sisi struktur fisinya dengan melakukan proses pemodelan parametrik dengan fungsi *Sérsic*. Pemodelan parametrik adalah metode pemodelan galaksi berdasarkan distribusi kecerlangan galaksi yang terproyeksi di bola langit. Kecerlangan galaksi akan didekati dengan fungsi analitik profil kecerlangan galaksi. Di sisi lain, dikenal juga metode pemodelan nonparametrik. Metode ini adalah metode pemodelan yang tidak bergantung pada model tertentu (*model-independent*), yang berarti tidak menggunakan asumsi fungsi analitik tertentu terhadap distribusi kecerlangan galaksi. Beberapa penelitian melakukan pemodelan dengan kedua metode ini, seperti dalam Tohill dkk. (2024) dan Kartaltepe dkk. (2023). Saat ini, telah ada banyak aplikasi pemodelan galaksi, seperti **Galfit** (Peng dkk., 2002), **MORFOMETRYKA** (Ferrari dkk., 2015), dan **statmorph** (Rodriguez-Gomez dkk., 2019).

Salah satu parameter struktur galaksi yang dapat diperoleh melalui pemodelan adalah ukuran galaksi. Penelitian terdahulu seperti dalam van der Wel dkk. (2014) membahas secara rinci distribusi ukuran galaksi pada rentang $0 < z < 3$. Penelitian ini menggunakan definisi radius efektif galaksi untuk menggambarkan ukurannya dan menganalisis hubungan massa terhadap radius efektif, serta hubungan radius efektif terhadap *redshift*.

James Webb Space Telescope yang diluncurkan pada tahun 2021 telah membuka jendela baru untuk pengamatan galaksi *redshift* tinggi. Instrumen JWST yang fokus utamanya melakukan pengamatan pada panjang gelombang inframerah membuat JWST dapat mengamati galaksi-galaksi *redshift* tinggi yang mengalami pergeseran merah akibat pengembangan alam semesta. Dengan ketersediaan data galaksi jauh dari pengamatan JWST, dimungkinkan untuk dilakukan analisis struktur dan morfologi galaksi sebagai petunjuk proses evolusi galaksi.

Dalam penelitian ini, penulis akan melakukan analisis morfologi dan struktur galaksi dengan menggunakan data galaksi pada *redshift* yang lebih tinggi dan rentang *redshift* yang lebih lebar. Penulis akan melakukan pengelompokan galaksi dengan menggunakan beberapa metode *unsupervised learning* dan membandingkan hasilnya dengan pekerjaan klasifikasi morfologi dengan inspeksi visual yang telah dilakukan sebelumnya. Selain itu, penulis akan melakukan pemodelan parametrik galaksi menggunakan aplikasi **Galfit** dengan *multi-sérsic fitting*. Hasil dari pengelompokan galaksi ini akan dibandingkan

dengan parameter galaksi dari katalog *EAZY* dan dari hasil pemodelan parametrik galaksi. Penelitian ini juga akan menganalisis distribusi galaksi berdasarkan pembentukan bintang didalamnya.

I.2 Rumusan dan Batasan Masalah

Pemahaman tentang proses evolusi galaksi salah satunya memerlukan informasi perubahan struktur galaksi dari waktu ke waktu. Namun, hal ini mustahil untuk dilakukan karena skala waktu evolusi galaksi yang sangat panjang. Oleh karena itu, untuk tetap dapat mempelajari proses evolusi galaksi, diperlukan data galaksi pada berbagai tahap evolusi. Dengan kata lain, dibutuhkan data galaksi pada berbagai era kosmik. Klasifikasi morfologi untuk galaksi lokal dianggap telah cukup baik, sehingga dalam penelitian ini digunakan data galaksi pada rentang *redshift* yang cukup lebar, yakni pada $z > 2$ dengan data JWST. Namun, untuk mengonfirmasi keberhasilan metode pengelompokan galaksi *redshift* tinggi, dalam penelitian ini akan dilakukan validasi dengan mengaplikasikan metode yang sama terhadap galaksi dekat yang telah dikelempokkan morfologinya. Data galaksi dekat yang digunakan berasal dari data pengamatan SDSS yang telah diklasifikasikan melalui program *Galaxy Zoo*.

Penelitian galaksi *redshift* tinggi berkembang pesat sejak diluncurkan JWST, seperti pada penelitian Kartaltepe dkk. (2023) yang menganalisis struktur galaksi hingga $z = 9$ dan Tohill dkk. (2024) yang menganalisis morfologi galaksi hingga $z > 7$. Teknologi pada teleskop ini membuka kesempatan untuk bisa mengamati galaksi yang lebih jauh, yang sebelumnya belum dapat dijangkau oleh teleskop-teleskop terdahulu seperti Hubble. Struktur galaksi *redshift* tinggi masih menjadi pertanyaan karena pengamatan menunjukkan galaksi-galaksi tersebut memiliki bentuk yang tidak beraturan dan membentuk *clump*. Tidak seperti galaksi-galaksi di alam semesta lokal yang tampak jelas fitur piringan atau lengan spiral di dalamnya, galaksi pada *redshift* yang sangat tinggi akan sangat redup dan sulit mendeskripsikan bentuknya. Oleh karena itu, untuk meminimalisir bias pengkategorian morfologi galaksi, penelitian ini akan menggunakan metode *unsupervised learning*, dimana mesin akan mempelajari fitur-fitur yang terdapat di setiap galaksi untuk selanjutnya mengelompokan galaksi berdasarkan kesamaan fitur yang dimiliki oleh setiap galaksi.

Semakin tinggi *redshift*-nya, galaksi-galaksi akan tampak semakin redup di panjang gelombang visual karena panjang gelombang galaksi tersebut akan semakin mengalami pemerahan. Sementara itu, pada penelitian ini penulis

menggunakan data pengamatan pada tiga filter, dengan panjang gelombang $1.15\mu m$, $2.77\mu m$, dan $4.4\mu m$. Penulis mengamati distribusi magnitudo galaksi dan mengambil limit magnitudo 27 sebagai batas sampel galaksi yang akan dianalisis. Pada magnitudo yang lebih tinggi dari limit tersebut, citra galaksi semakin sulit dianalisis.

I.3 Tujuan

Berdasarkan perumusan masalah yang telah dijelaskan pada bagian I.2, tujuan dari penelitian ini dapat dirumuskan dalam empat poin.

1. Merekonstruksi citra galaksi pada berbagai *redshift* menggunakan metode *Variational Autoencoder* (VAE) untuk pekerjaan *preliminary* pengelompokan morfologi galaksi.
2. Membandingkan hasil pemodelan parametrik *multi Sérsic* galaksi *redshift* tinggi terhadap parameter laten hasil ekstraksi data citra galaksi.
3. Membandingkan hasil pengelompokan galaksi dengan pendekatan *machine learning* dan inspeksi visual.
4. Melakukan analisis distribusi ukuran galaksi *redshift* tinggi hasil pemodelan *multi Sérsic*.

I.4 Metodologi

Penelitian ini diawali dengan melakukan studi pustaka terkait topik galaksi dan berbagai penelitian terbaru terkait struktur dan morfologinya. Penelitian ini dilanjutkan dengan melakukan pemodelan parametrik galaksi dengan menggunakan *multi Sérsic fitting*. Selanjutnya penulis melakukan ekstraksi parameter struktur galaksi dari data fotometri menggunakan metode *Variational Autoencoder* (VAE). *Output* dari proses VAE, yakni berupa parameter laten berisi fitur-fitur dalam data galaksi, akan menjadi input terhadap proses pengelompokan galaksi. Pengelompokan dilakukan dengan menggunakan metode *unsupervised learning*, *hierarchical clustering* dan *k-means clustering*. Kemudian akan dilakukan analisis parameter laten dengan melakukan reduksi dimensi menggunakan metode PCA dan UMAP. Parameter laten tersebut terakhir akan dibandingkan dengan beberapa parameter galaksi dari katalog, serta dibandingkan dengan hasil pemodelan parametrik galaksi.

I.5 Sistematika Penulisan

Setelah membahas pendahuluan di Bab I, pada Bab II akan dijelaskan teori yang mendasari penelitian ini. Pada Bab III akan dibahas data serta metode yang digunakan secara lebih rinci. Hasil awal dari penelitian ini akan dijelaskan pada Bab IV bersama dengan analisis penulis terhadap hasil yang didapat. Dalam Bab V penulis menyampaikan kesimpulan yang diperoleh dari penelitian ini, serta di bagian terakhir penulis menyampaikan saran untuk kelanjutan penelitian ini kedepannya.

BAB II

TEORI DASAR

Topik besar penelitian ini adalah mengenai pemanfaatan *machine learning* untuk melakukan pengelompokan galaksi di *redshift* tinggi. Oleh karena itu, pada bab ini akan dibahas teori astrofisika mengenai galaksi dan dilanjutkan dengan pembahasan tentang berbagai metode *machine learning* yang dapat digunakan untuk menganalisis struktur galaksi di *redshift* tinggi. Pada bab ini juga akan dibahas beberapa hasil penelitian terdahulu dengan bahasan topik yang serupa.

II.1 Galaksi

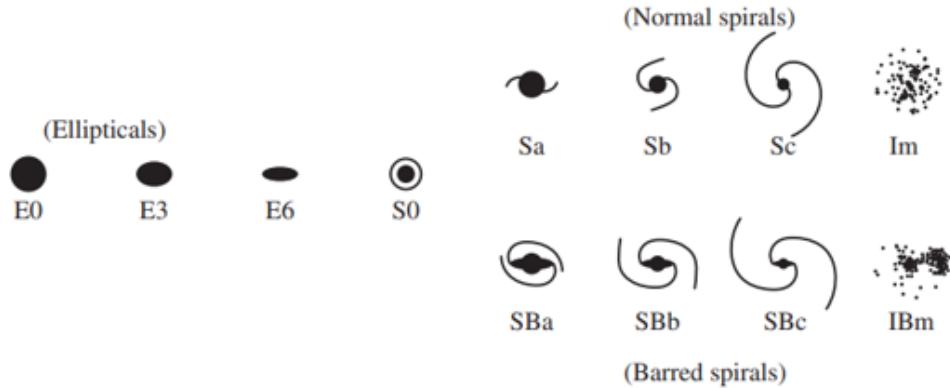
Galaksi merupakan struktur yang tersusun dari kumpulan bintang, gas, dan debu yang terikat secara gravitasi. Galaksi disebut sebagai *building blocks* alam semesta karena menjadi objek astrofisika yang membangun struktur skala besar. Galaksi juga sering disebut sebagai miniatur alam semesta karena didalamnya terdapat ekosistem yang kompleks.

Galaksi dianggap sebagai keseluruhan alam semesta hingga tahun 1920. Pada awalnya, manusia beranggapan bahwa setiap objek di langit berada di dalam Galaksi. Namun, pengamatan Hubble pada tahun 1929 menunjukkan bahwa jarak galaksi—yang disebut sebagai nebula oleh Hubble pada saat itu—lebih jauh dibandingkan jarak bintang-bintang di dalam Galaksi. Penemuan ini membuka pintu baru untuk pengamatan ekstragalaksi.

II.1.1 Morfologi Galaksi

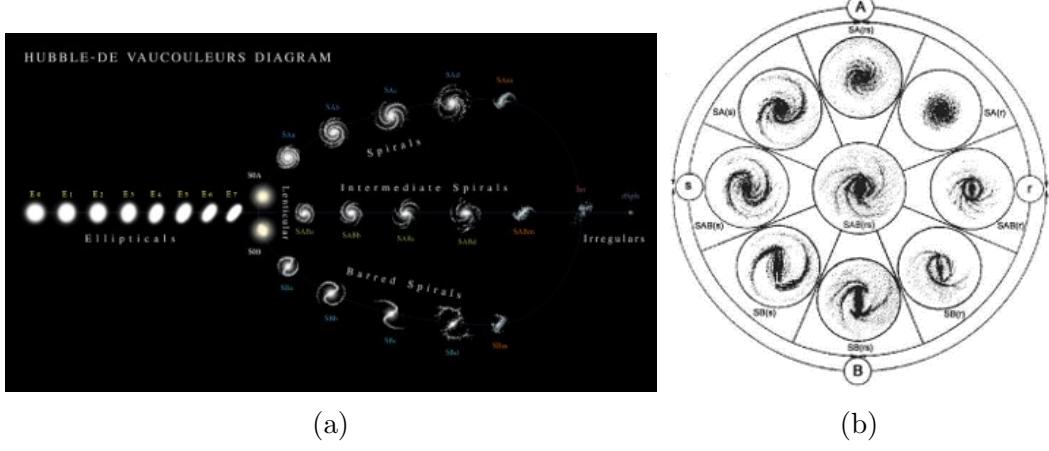
Pada tahun 1926, Hubble memperkenalkan sistem klasifikasi galaksi berdasarkan morfolohnya, yang kini dikenal sebagai diagram *Hubble tuning-fork*. Hubble mengkategorikan galaksi ke dalam dua kelas besar, yaitu galaksi elips dan galaksi spiral. Kedua kategori galaksi ini juga sering disebut sebagai kategori galaksi *early-types* (galaksi elips) dan *late-types* (galaksi spiral). Secara lebih rinci, Hubble mengklasifikasikan kedua kategori galaksi tersebut ke dalam kelas-kelas yang lebih kecil berdasarkan fitur-fitur spesifik yang dimiliki

pada setiap kategori. Selain itu, Hubble juga mengklasifikasikan galaksi dengan struktur yang tak beraturan ke dalam kelas morfologi *irregular*. Skema klasifikasi Hubble ini ditunjukkan pada Gambar II.1.



Gambar II.1: Diagram *Hubble tuning-fork* Sumber: Abraham (1998)

Selain skema klasifikasi Hubble, pada tahun 1959 seorang astronom Prancis bernama Gérard Henri de Vaucouleurs membuat skema klasifikasi galaksi de Vaucouleurs. Sistem klasifikasi yang ditunjukkan pada Gambar II.2 ini merupakan pengembangan dari sistem klasifikasi Hubble, namun dengan penambahan beberapa kategori baru. Kategori baru tersebut mendeskripsikan secara lebih rinci fitur-fitur yang terdapat dalam sebuah galaksi, seperti pemberian notasi *r* yang menunjukkan galaksi spiral dengan fitur cincin, dan galaksi dengan notasi *s* yang menunjukkan galaksi spiral tanpa fitur cincin.



Gambar II.2: Morfologi galaksi dalam skema Hubble-deVaucouleurs dikategorikan dalam kelas galaksi elips dan kelas galaksi spiral, dan setiap kategori dikategorikan secara lebih rinci (panel a). de Vaucouleurs mengkategorikan galaksi spiral berdasarkan keberadaan fitur batang dan cincin di sekelilingnya (panel b). Sumber: <https://www.cosmic-core.org/free/article-93-astronomy-redshift-the-non-expanding-universe/> dan Delaive (2013)

Galaksi elips memiliki karakteristik mengandung sedikit gas dan debu, karena sebagian besar gas dan debu didalamnya telah habis digunakan sebagai bahan pembentukan bintang. Bintang-bintang pada galaksi elips bergerak acak, dan galaksi elips biasanya tersusun atas bintang-bintang tua. Sementara itu, ciri khas galaksi spiral yaitu memiliki komponen *bulge* dan piringan. Bagian piringan galaksi spiral merupakan daerah pembentukan bintang sehingga masih terdapat cukup banyak gas di dalamnya. Sementara bagian *bulge* galaksi spiral tersusun atas bintang-bintang tua. Ditinjau dari proses evolusi yang terjadi, hal ini menunjukkan bahwa bagian *bulge* galaksi spiral terbentuk lebih awal dibandingkan bagian piringannya. Bintang-bintang pada galaksi spiral bergerak secara teratur dalam mengelilingi pusat galaksi. Galaksi *irregular* diyakini terbentuk dari interaksi antargalaksi. Galaksi ini biasanya ditemukan pada galaksi-galaksi yang sangat jauh, yang masih dalam proses pembentukan. Namun, galaksi *irregular* juga ditemukan pada galaksi-galaksi dekat, seperti NGC1427A yang ditunjukkan pada Gambar II.3.



Gambar II.3: NGC1427A sebagai salah satu galaksi *irregular* pada *redshift* dekat, yakni di $z = 0.0067$. Sumber: <https://science.nasa.gov/missions/hubble/the-impending-destruction-of-ngc-1427a/>

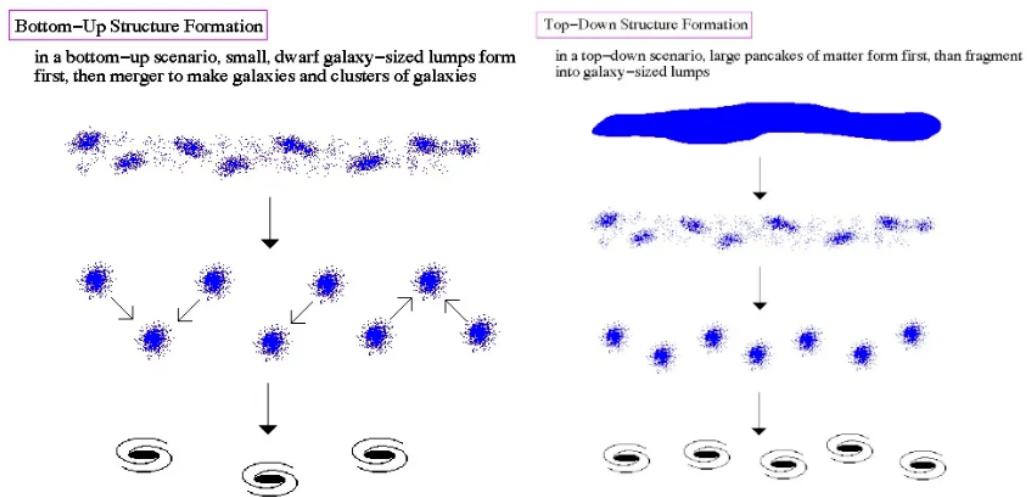
II.1.2 Teori Pembentukan Galaksi

Pada tahun 1962, Eggen, Lynden Bell, dan Sandage menjelaskan bahwa galaksi terbentuk dari keruntuhan awan gas akibat gravitasi. Namun, gaya gravitasi dari awan gas saja tidak cukup kuat agar dapat membentuk galaksi dan menciptakan struktur skala besar sebagaimana yang diamati saat ini. Struktur skala besar seperti yang diamati saat ini baru dapat terbentuk jika usia alam semesta lebih tua. Oleh karena itu, model kosmologi standar saat ini meyakini bahwa galaksi-galaksi terbentuk pada sumur potensial materi gelap.

Model kosmologi Λ CDM menjelaskan bahwa alam semesta tersusun atas $\sim 70\%$ *dark energy*, dan $\sim 20\%$ *dark matter*, serta sebagian kecil materi baryon. Dari hasil simulasi, model *cold dark matter* (CDM) merupakan model materi gelap yang mampu menghasilkan struktur skala besar yang sama seperti struktur yang diamati melalui pengamatan. Meski model ini cocok dalam menjelaskan struktur skala besar 13.8 miliar tahun setelah *Big Bang*, terdapat beberapa masalah skala kecil untuk model ini seperti *core-cusp problem*, *missing satellites problem*, dan *too big to fail problem*.

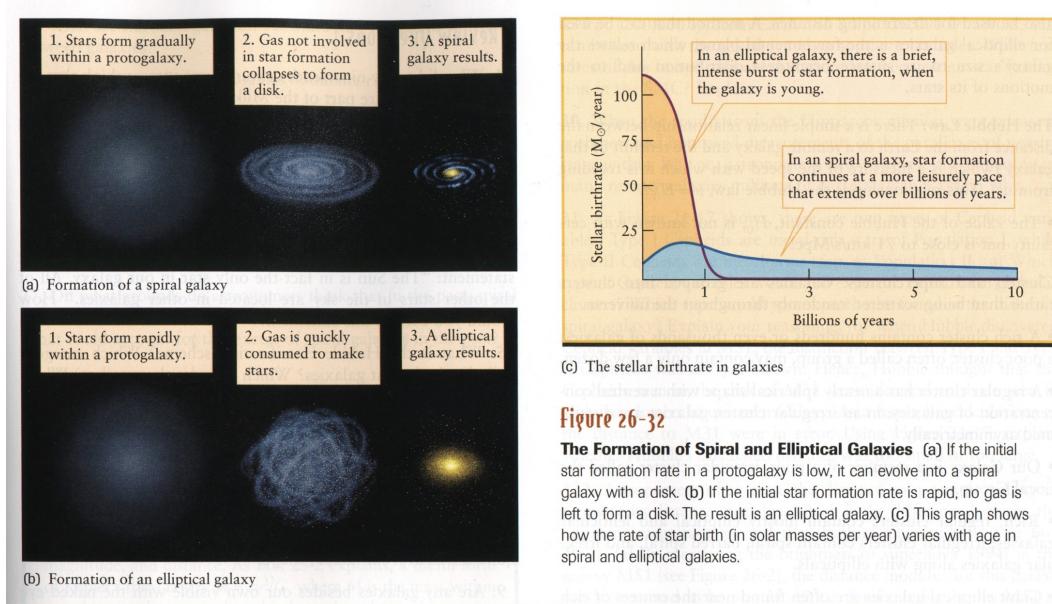
Model CDM merupakan model materi gelap yang bersifat non relativistik pada era pembentukan galaksi, sehingga model ini dapat menjadi rumah bagi galaksi terbentuk dan memiliki massa yang cukup besar untuk menyebabkan

keruntuhan gravitasi. Dari teori sejarah termal alam semesta diketahui bahwa materi gelap akan ter-*decouple* lebih awal dari foton, sehingga materi gelap dapat membuat sumur-sumur potensial lebih awal. Seiring dengan pengembangan alam semesta, kondisi alam semesta yang cukup renggang membuat materi baryon selanjutnya ter-*decouple* dari foton. Materi baryon tersebut kemudian masuk ke dalam sumur-sumur potensial yang telah dibangun oleh materi gelap sebelumnya.



Gambar II.4: Dua skema pembentukan galaksi yang diyakini saat ini. Sumber: <http://abyss.uoregon.edu/~js/ast123/lectures/lec24.html>

Terdapat dua skema pembentukan galaksi yang diyakini saat ini, yakni skema klasik atau *top-down* dan skema hierarki atau *bottom-up*. Kedua skema pembentukan galaksi ini ditunjukkan dalam Gambar II.4. Dalam skema pembentukan galaksi klasik, galaksi terbentuk dari satu awan gas besar yang runtuh akibat tarikan gravitasi. Sementara itu, dalam skema pembentukan galaksi hierarki, galaksi-galaksi kecil terbentuk lebih dulu dan membentuk galaksi besar melalui peristiwa *merging*. Proses *merging* dimulai dari bergabungnya *halo* materi gelap sebelum bergabungnya bintang-bintang dan gas.



(c) The stellar birthrate in galaxies

Figure 26-32

The Formation of Spiral and Elliptical Galaxies (a) If the initial star formation rate in a protogalaxy is low, it can evolve into a spiral galaxy with a disk. (b) If the initial star formation rate is rapid, no gas is left to form a disk. The result is an elliptical galaxy. (c) This graph shows how the rate of star birth (in solar masses per year) varies with age in spiral and elliptical galaxies.

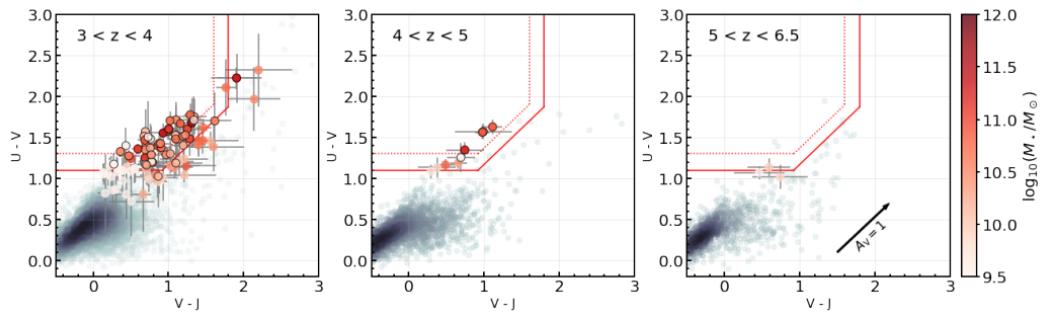
Gambar II.5: Mekanisme pembentukan galaksi elips dan spiral, serta kurva laju pembentukan bintang di kedua tipe galaksi tersebut. Sumber: <http://w3.phys.nthu.edu.tw/~hkchang/ga2/f2608-starbirth.JPG>

Klasifikasi galaksi ke dalam dua kategori besar, yakni galaksi elips dan galaksi spiral, memerlukan penjelasan terkait mekanisme pembentukan kedua tipe galaksi. Perbedaan mekanisme pembentukan galaksi ini terkait dengan laju pembentukan bintang di dalam galaksi. Gambar II.5 menunjukkan skema pembentukan galaksi elips dan spiral serta kurva laju pembentukan bintang terhadap waktu. Jika setelah terjadi keruntuhan gravitasi masih tersisa cukup banyak komponen gas, maka awan gas akan menyusut hingga keruntuhan gravitasi diimbangi oleh momentum sudut akibat gerak rotasi galaksi. Sistem ini akan menciptakan galaksi dengan fitur piringan atau galaksi spiral yang momentumnya diimbangi momentum rotasi galaksi (*rotation-supported*). Pada plot laju kelahiran bintang terhadap waktu, mekanisme ini menunjukkan laju kelahiran bintang yang relatif rendah, namun menurun secara gradual untuk rentang waktu yang panjang. Hal ini dapat menjelaskan keberadaan berbagai populasi bintang di dalam galaksi spiral. Sementara itu, jika sebagian besar gas langsung membentuk bintang ketika proses keruntuhan akibat gravitasi terjadi, maka akan terbentuk galaksi elips. Mekanisme ini akan menciptakan sistem yang *pressure-supported*, dengan tekanan gravitasi akan diimbangi oleh tekanan akibat gerak acak bintang-bintang. Pada plot laju pembentukan bintang untuk galaksi elips, tampak laju pembentukan bintang yang cukup tinggi sejak awal dan menurun secara tajam dalam waktu yang relatif singkat.

Hal ini dapat menjelaskan fakta pengamatan yang menunjukkan galaksi elips terdiri dari bintang-bintang pada usia yang relatif sama.

II.1.3 Klasifikasi Galaksi Berdasarkan Pembentukan Bintang di Dalamnya

Selain diklasifikasikan berdasarkan morfologinya seperti yang telah dijelaskan pada bagian II.1.1, galaksi juga dapat dibedakan berdasarkan pembentukan bintang di dalamnya. Galaksi-galaksi yang masih aktif membentuk bintang dikategorikan sebagai galaksi *star forming*, sementara galaksi-galaksi yang telah berhenti membentuk bintang dikategorikan sebagai galaksi *quiescent*. Salah satu cara untuk mengklasifikasikan kedua tipe galaksi ini adalah dengan diagram UVJ (Williams dkk., 2009). Gambar II.6 menunjukkan contoh penyeleksian galaksi *quiescent* dan *star forming* menggunakan diagram UVJ. Galaksi *quiescent* memiliki nilai $U - V$ yang tinggi, dan posisinya berada di pojok kiri atas diagram tersebut. Sementara galaksi *star forming* berada di luar populasi galaksi *quiescent*. Garis yang membatasi kedua tipe galaksi ini pada diagram UVJ diberikan pada persamaan II.1 (Williams dkk., 2009). Selain itu, Valentino dkk. (2023) memberikan toleransi 0.2 magnitudo untuk seleksi galaksi ini, sehingga seleksi galaksi *star forming* dan galaksi *quiescent* dibatasi dengan garis yang mengikuti persamaan II.2.

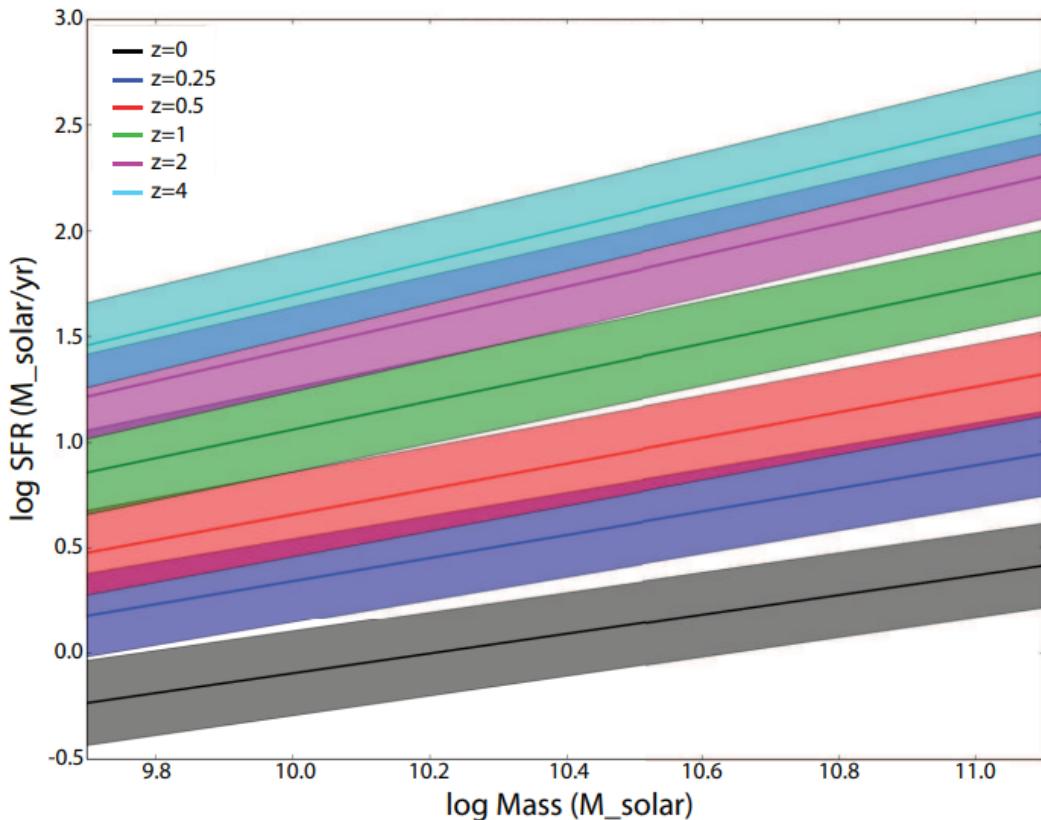


Gambar II.6: Diagram UVJ untuk mengklasifikasi galaksi *star forming* dan galaksi *quiescent*. Sumber: Valentino dkk. (2023)

$$\begin{aligned}
 (U - V) &> 0.88(V - J) + 0.49, \\
 U - V &> 1.3, \\
 V - J &< 1.6,
 \end{aligned} \tag{II.1}$$

$$\begin{aligned}
(U - V) &> 0.88(V - J) + 0.29, \\
U - V &> 1.1, \\
V - J &< 1.8,
\end{aligned} \tag{II.2}$$

Seleksi galaksi *star forming* dan *quiescent* juga dapat dilakukan dengan melintasi kurva laju pembentukan bintang di galaksi terhadap massanya, yang dinamakan *star forming main sequence* (Speagle dkk., 2014), seperti yang ditunjukkan dalam Gambar II.7. Galaksi yang berada di atas kurva yang diberikan dalam persamaan II.3 dan II.4 merupakan galaksi *star forming*, sementara galaksi-galaksi yang berada dibawahnya merupakan galaksi *quiescent*. Persamaan II.3 digunakan untuk menyeleksi galaksi-galaksi pada $z \leq 5$, sementara persamaan II.4 digunakan untuk menyeleksi galaksi-galaksi pada $z > 5$.



Gambar II.7: Kurva laju pembentukan bintang terhadap massa galaksi untuk menyeleksi galaksi *star forming* dan *quiescent*. Sumber: Speagle dkk. (2014)

$$\log \psi(M_*, t) = (0.84 \pm 0.02 - (0.026 \pm 0.003)t) \log M_* - (6.51 \pm 0.24 - (0.11 \pm 0.03)t)$$

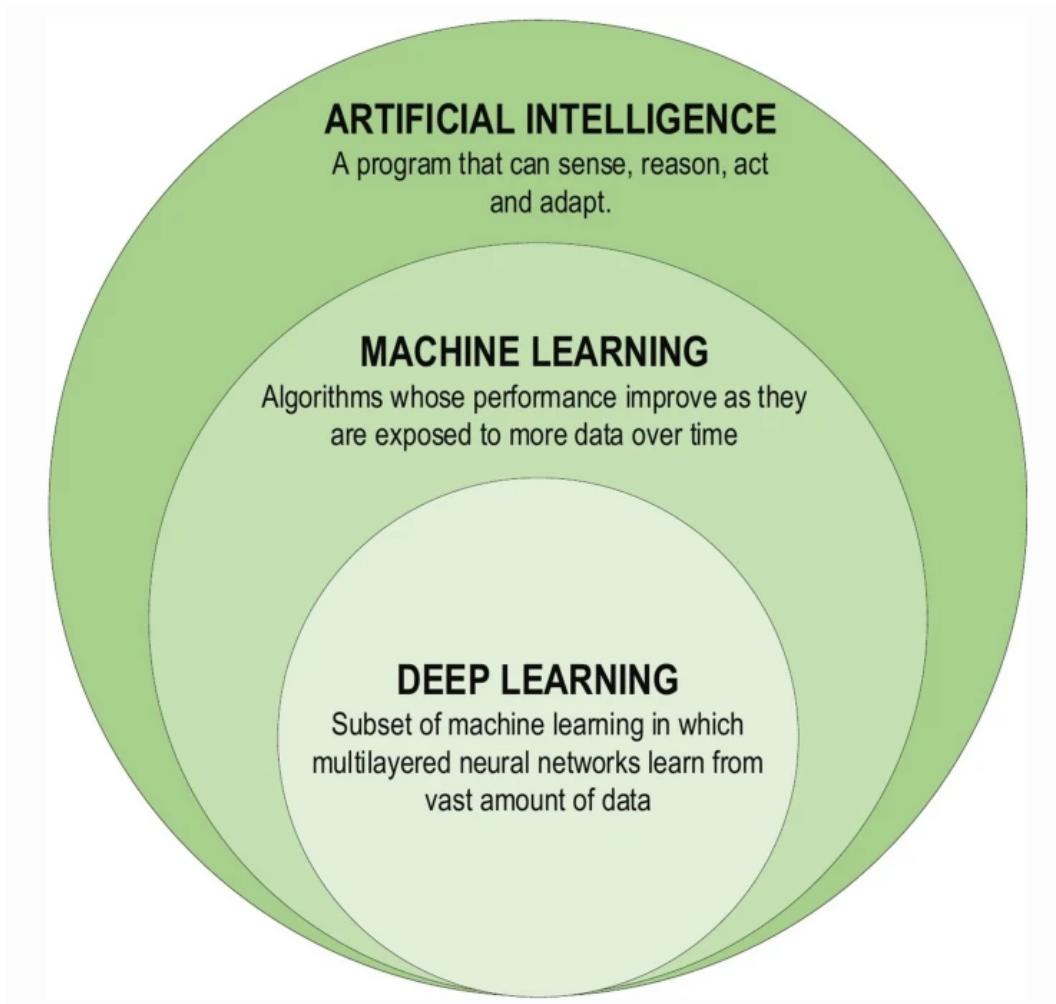
(II.3)

$$\log \psi(M_*, t) = (0.80 \pm 0.02 - (0.022 \pm 0.003)t) \log M_* - (6.09 \pm 0.23 - (0.07 \pm 0.03)t)$$

(II.4)

II.2 *Machine Learning*

Machine learning (ML) atau pembelajaran mesin adalah bagian dari *artificial intelligence* untuk menjalankan sistem yang dapat belajar secara otomatis dan meningkatkan performa serta akurasi berdasarkan pengalaman yang dipelajari. *Artificial intelligence* sendiri merupakan salah satu cabang sains yang didedikasikan untuk perancangan sistem yang dapat bekerja layaknya manusia. Sementara itu, dikenal istilah *deep learning* yang merupakan algoritma untuk menopang sistem kerja *machine learning*. Gambar II.8 menunjukkan keterkaitan antara AI, ML, dan *deep learning*. Selain ketiga istilah tersebut, dikenal istilah *artificial neural network* (ANN) yang merupakan jaringan saraf tiruan yang digunakan dalam *deep learning*, dan menjadi inti dari *deep learning* itu sendiri. Dengan demikian, keempat hal tersebut memiliki keterkaitan antar satu sama lain.



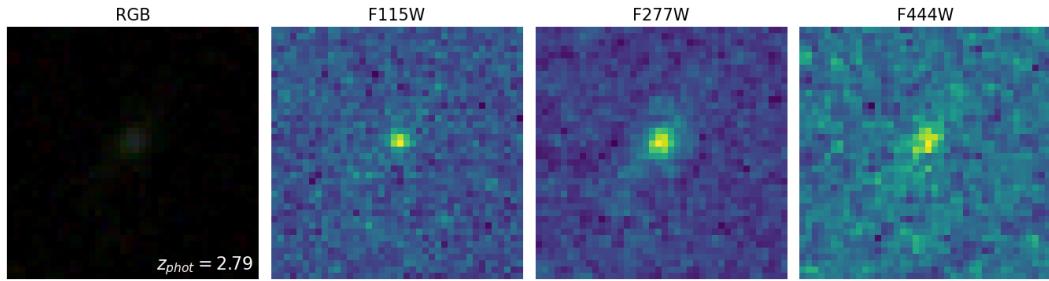
Gambar II.8: Diagram yang menunjukkan hubungan *artificial intelligence* (AI), *machine learning* (ML), dan *deep learning*. Sumber: Alzubaidi dkk. (2021)

ML telah digunakan dalam berbagai bidang, baik di dunia industri maupun akademik. Di dunia industri, ML banyak digunakan di dunia medis seperti untuk mendeteksi suatu penyakit berdasarkan kondisi pasien. Dalam penelitian Hossain dkk. (2023) misalnya, mereka membuat model untuk memprediksi kelainan jantung berdasarkan data hasil tes kesehatan pasien dan beberapa pertanyaan yang diajukan secara langsung kepada pasien. Dalam penelitian tersebut, mereka menggunakan 7 metode *supervised learning* dan membandingkan performa untuk setiap model. ML juga digunakan dalam analisis sentimen pada beberapa perusahaan yang bergerak di bidang monitoring media sosial. Untuk keperluan riset sains dasar, metode ML juga banyak dilakukan, misalnya untuk melakukan pengelompokan morfologi galaksi seperti yang dilakukan Tohill dkk. (2024).

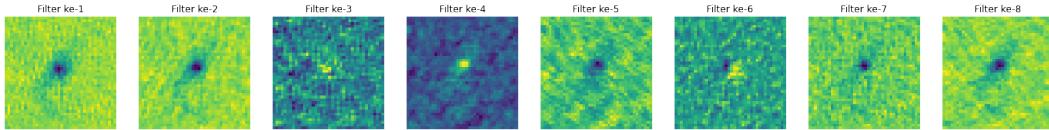
Terdapat tiga metode ML, yakni *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. *Supervised learning* dan *unsupervised learning* dipraktikkan bagi tipe data yang berbeda. *Supervised learning* digunakan untuk data yang memiliki informasi input dan output yang saling berkorespondensi, atau tipe data berlabel. Oleh karena itu, tujuan algoritma *supervised learning* adalah menganalisis kesamaan pola pada data dan membuat model untuk memetakan input terhadap *output*/label. Salah satu implementasi *supervised learning* adalah pada klasifikasi galaksi *lensing* seperti yang dilakukan Jaelani dkk. (2024). Sementara itu, *unsupervised learning* digunakan untuk menganalisis data yang tidak diketahui outputnya secara historis, atau tipe data tak berlabel. Setiap data tidak berkorespondensi dengan label atau *output* tertentu, sehingga algoritma ini akan mengidentifikasi kesamaan pola data input untuk dapat dikelompokkan ke dalam sejumlah kelas atau kategori. *Autoencoder* merupakan salah satu algoritma *unsupervised learning* yang digunakan untuk mengekstrak fitur-fitur penting dalam sebuah data, kemudian merekonstruksi data berdasarkan fitur-fitur tersebut. *Autoencoder* bermanfaat dalam mengurangi bias data dan meminimalisir waktu komputasi dalam proses analisis data. *Autoencoder* diaplikasikan dalam Tohill dkk. (2024) untuk melakukan pengelompokan morfologi galaksi. *Autoencoder* lebih rinci akan dijelaskan pada Subbab II.2.2. Metode ML yang ketiga yaitu *reinforcement learning*. Metode ini membangun mesin yang dapat melakukan pengambilan keputusan dengan mempertimbangkan keuntungan dari beberapa opsi jalan yang tersedia. Algoritma ini bekerja dengan memahami lingkungan sekitarnya dan mengambil keputusan terbaik berdasarkan skema *trial and error*.

II.2.1 *Convolutional Neural Network*

Convolutional Neural Network (CNN) merupakan salah satu metode *machine learning* yang menggunakan jaringan-jaringan neuron layaknya ANN, namun dengan memberi lapisan tambahan berupa lapisan konvolusi. Lapisan konvolusi adalah lapisan berisi filter yang digunakan untuk mengenali pola atau fitur-fitur pada data. CNN merupakan salah satu metode *feed-forward neural network*, yaitu *neural network* yang mengirimkan informasi dalam satu arah dan tanpa melalui tahap *feedback*. Metode CNN banyak digunakan untuk menangani data berupa gambar, misalnya mengekstrak fitur dalam gambar untuk mengklasifikasikan data ke dalam kategori tertentu.

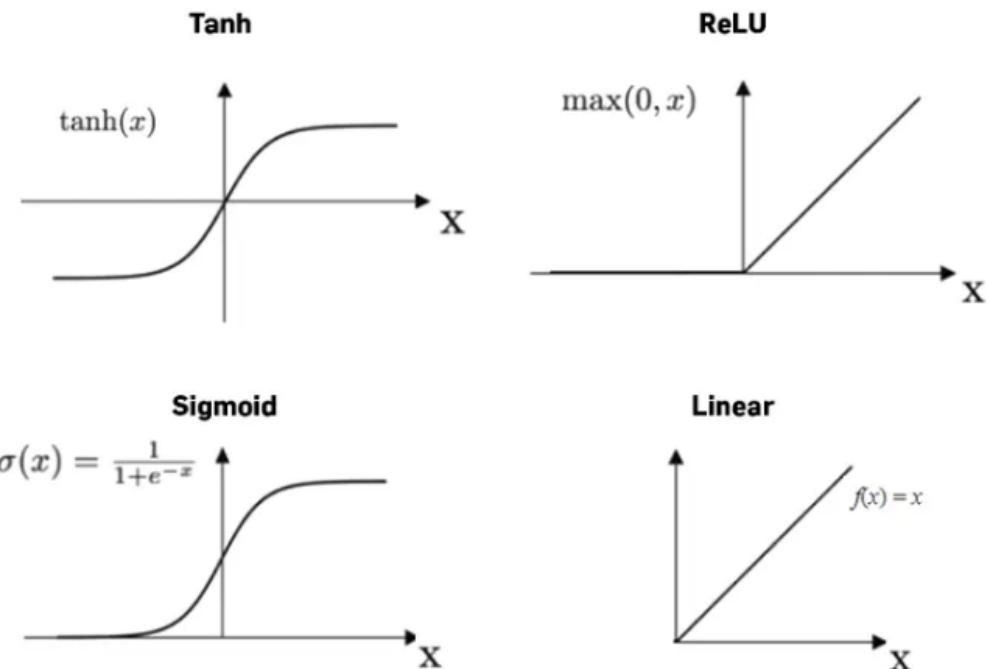


Gambar II.9: Sampel data galaksi (CEERS 481) dari tiga filter dan gabungan ketiga filter (*RGB Image*).



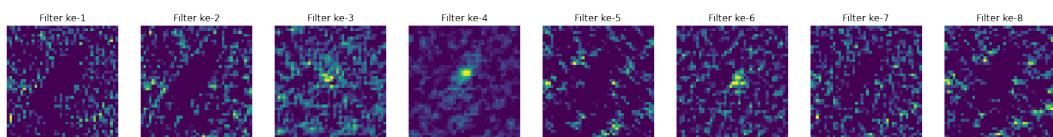
Gambar II.10: Visualisasi *output* dari *convolution layer* dengan input berupa citra RGB dari Gambar II.9.

Gambar II.9 menunjukkan salah satu sampel galaksi sebelum melalui lapisan-lapisan dalam model CNN. Gambar II.10 menunjukkan data RGB galaksi yang sama setelah melalui lapisan konvolusi dengan 8 filter, menggunakan kernel berukuran 3x3. Data akan dikalikan melalui operasi *dot product* menggunakan kernel yang akan diterapkan pada seluruh area data input. Karena terdapat 8 filter dalam lapisan ini, maka akan dihasilkan 8 *feature map*. Selain lapisan konvolusi (*convolution layer*), beberapa jenis lapisan lainnya yang sering ditemui dalam jaringan saraf tiruan diantaranya adalah *activation layer* dan *pooling layer*. *Activation layer* adalah lapisan yang berfungsi untuk menambahkan non-linearitas ke dalam model CNN. Dengan menambahkan non-linearitas, mesin dapat mempelajari pola yang lebih kompleks. Dengan kata lain, hasil *dot product* sebelumnya akan dibuat sedemikian sehingga mengikuti sebuah fungsi aktivasi. Beberapa fungsi aktivasi yang biasa digunakan diantaranya adalah ReLU, Sigmoid, dan Tanh. Gambar II.11 menunjukkan plot beberapa macam fungsi aktivasi.

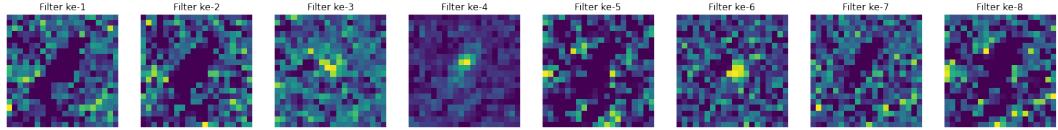


Gambar II.11: Beberapa contoh fungsi aktivasi. Sumber: <https://machine-learning.paperspace.com/wiki/activation-function>.

Setelah melewati lapisan konvolusi dan menghasilkan *output* pada Gambar II.10, selanjutnya *output* tersebut akan tampak seperti pada Gambar II.12 setelah melewati lapisan aktivasi. Sementara itu, lapisan *pooling* merupakan lapisan untuk mengurangi dimensi data yang fungsinya mengurangi waktu komputasi dan mencegah *overfitting*. Terdapat dua jenis lapisan *pooling* yang sering digunakan, yaitu *max pooling* dan *average pooling*. Sesuai dengan namanya, *max pooling* akan mengambil nilai tertinggi dari setiap blok data dengan ukuran tertentu, sementara *average pooling* akan mengambil nilai rata-rata dari setiap blok data tersebut. Gambar II.13 menunjukkan visualisasi *output* dari lapisan *max pooling* yang dilewati oleh Gambar II.12. Apabila input data sebelum memasuki lapisan *pooling* adalah 40×40 , dan ukuran *pooling* adalah 2×2 , maka secara *default* *output* dari *layer* ini adalah *feature map* berukuran 20×20 .



Gambar II.12: Visualisasi *output* dari *activation layer* dengan input dari Gambar II.10.



Gambar II.13: Visualisasi *output* dari *max pooling layer* dengan input dari Gambar II.12.

Lapisan lainnya yang sering digunakan yaitu *flatten layer* dan *dense layer*. *Flatten layer* merupakan lapisan yang berfungsi mengurangi dimensi *feature maps* menjadi hanya satu dimensi. Misalnya data yang semula berukuran $10 \times 10 \times 16$, setelah melewati lapisan *flatten* akan berukuran 1×1600 . Semenara itu, *dense layer* atau sering disebut sebagai *fully-connected layer* adalah lapisan akhir yang berfungsi dalam proses pembuatan keputusan, prediksi, regresi, maupun klasifikasi. Lapisan ini sesuai dengan namanya merupakan lapisan dimana setiap neuron terhubung dengan setiap neuron di lapisan sebelumnya dan lapisan selanjutnya. Dalam CNN, lapisan ini akan menafsirkan *feature map* n-dimensi ke dalam beberapa kelompok.

II.2.2 *Variational Autoencoder*

Autoencoder merupakan salah satu jenis jaringan saraf tiruan yang bertujuan untuk membuat representasi data dalam bentuk yang lebih ringkas, lalu membangun kembali data berdasarkan representasi tersebut. *Autoencoder* yang baik mampu menghasilkan data buatan yang mirip dengan data aslinya. *Autoencoder* diharapkan mampu menghasilkan data buatan yang hanya berisi fitur-fitur penting dalam data asli. *Autoencoders* terdiri dari tiga bagian, yaitu *encoder*, *decoder*, dan penghubung yang dinamakan *bottleneck*.

Encoder merupakan bagian yang berperan dalam mengekstrak fitur dari data asli. Oleh karena itu, algoritma *encoder* pada dasarnya memiliki konsep yang mirip dengan metode CNN dalam proses ekstraksi fitur. *Output* dari *encoder* adalah *feature map* dalam satu dimensi yang selanjutnya akan dipetakan ke dalam sebuah ruang laten.

Ruang laten adalah sebuah ruang n-dimensi yang menyimpan parameter-parameter penting dalam data yang telah diekstrak dari bagian *encoder*. Dengan kata lain, ruang laten merupakan representasi dari data dalam bentuk vektor yang berdimensi lebih kecil dari data aslinya, dan mengandung informasi penting yang menggambarkan data asli. Ruang laten merupakan penghubung antara *encoder* dan *decoder*, sehingga dapat disebut sebagai lapisan *bottleneck*.

Decoder merupakan proses untuk merekonstruksi kembali data berdasarkan parameter dalam ruang laten. *Decoder* dibangun dengan menggunakan lapisan-lapisan yang digunakan dalam proses *encoder*, namun dengan urutan yang berkebalikan dengannya. *Output* dari proses *decoder* merupakan data buatan yang diharapkan cukup baik merepresentasikan data asli, karena dibangun hanya dari fitur-fitur utama dari data asli.

Terdapat beberapa jenis *autoencoder* berdasarkan prosesnya, salah satunya adalah *variational autoencoder* (VAE). Apabila *autoencoder* biasa merepresentasikan data dalam sebuah parameter yang deterministik, VAE merepresentasikan data dalam bentuk parameter laten yang probabilistik. Secara sederhana, *feature map* satu dimensi dari *encoder* akan dipetakan ke dalam n-parameter laten sebagai nilai $\text{mean}(\mu)$ dan variansi(σ^2). Sementara itu, parameter laten (z) yang juga berupa vektor n-dimensi dihitung dengan *sampling* mengikuti persamaan II.5 (Kingma dan Welling , 2019).

$$z = \mu + \sigma \odot \epsilon \quad (\text{II.5})$$

Dengan ϵ merupakan distribusi normal standar ($\epsilon \sim \mathcal{N}(0, 1)$) yang dibuat agar nilai parameter laten z memiliki probabilitas yang mengikuti distribusi Gaussian. Sementara itu, deviasi standar dan variansi dapat dituliskan sebagai $\sigma = e^{0.5 \log(\sigma^2)}$, sehingga persamaan II.5 dapat dituliskan dalam bentuk persamaan II.6. Nilai parameter laten dan distribusinya akan berubah seiring dengan belajarnya mesin dalam setiap iterasi. Perubahan nilai parameter laten berubah seiring dengan memminimumkan nilai *loss*.

$$z = \mu + \exp(0.5 \log(\sigma^2)) \cdot \epsilon \quad (\text{II.6})$$

Dalam proses VAE, terdapat dua definisi *loss* yang digunakan, yaitu *reconstruction loss* dan *regularization loss*. *Reconstruction loss* didefinisikan sebagai besarnya perbedaan data asli dengan data yang direkonstruksi oleh *decoder*. *Reconstruction loss* dapat dihitung dengan beberapa metode, diantaranya dengan nilai *mean squared error* (MSE), *mean absolute error* (MAE) atau *binary cross-entropy* (BCE). Di sisi lain, *regularization loss* menghitung seberapa dekat distribusi parameter laten dengan distribusi normal. Hal ini bertujuan untuk mengontrol representasi parameter laten agar tetap teratur. Pada umumnya, *regularization loss* dihitung dengan metode *Kullback–Leibler*

(KL) *divergence loss*. Perhitungan *KL-loss* dirumuskan menurut persamaan II.7.

$$KL_{loss}(\mu, \sigma) = \frac{1}{2}\beta(-\log(\sigma^2) - 1 + \sigma^2 + \mu^2) \quad (\text{II.7})$$

Selain menggunakan *KL-loss*, *regularization loss* juga dapat dihitung dengan definisi *maximum mean discrepancy* (MMD) *loss*. Apabila *KL-loss* membandingkan distribusi secara keseluruhan, *MMD-loss* membandingkan rata-rata sampel kedua distribusi. *MMD-loss* dirumuskan dalam persamaan II.8, dengan $k(z, z')$ dapat berupa fungsi yang universal, dan biasanya menggunakan kernel Gaussian.

$$MMD_{loss}(\mu, \sigma) = \mathbb{E}_{p(z), p(z')}[k(z, z')] + \mathbb{E}_{q(z), q(z')}[k(z, z')] - 2\mathbb{E}_{p(z), q(z')}[k(z, z')] \quad (\text{II.8})$$

II.2.3 Clustering

Tujuan dari penelitian ini adalah untuk dapat mengetahui berbagai tipe morfologi galaksi di *redshift* tinggi. Beberapa penelitian mencoba mengkategorikan galaksi ke dalam tiga kategori, yaitu *disk galaxies*, *spheroidal galaxies*, dan *irregular galaxies*. Tetapi pengkategorian ini berpotensi menutup kemungkinan adanya kategori galaksi lainnya dengan bentuk selain ketiga bentuk tersebut. Agar pengkategorian galaksi tidak dibatasi oleh asumsi subjektif bahwa galaksi hanya terdiri dari tiga kategori, dalam penelitian ini dilakukan metode *clustering* sebagai pendekatan yang lebih objektif untuk mengelompokan galaksi. Sehingga jumlah kategori yang dihasilkan akan berdasarkan kesamaan pola pada data yang dipelajari oleh mesin.

Hierarchical Clustering

Hierarchical clustering adalah metode pengelompokan data dimana data dibentuk dalam struktur *dendrogram* berdasarkan kemiripan antardata. Terdapat dua pendekatan dalam metode pengelompokan ini, yaitu *agglomerative* atau *bottom-up*, dan *divisive* atau *top-down*. Pada pendekatan *agglomerative*, setiap titik data awalnya dianggap sebagai satu klaster, lalu bergabung berdasarkan kemiripan antarklaster. Masing-masing klaster akan terus bergabung hingga pada akhirnya semua titik data dianggap sebagai satu klaster. Semen-

tara itu, pendekatan *divisive* mengasumsikan seluruh data awalnya sebagai satu klaster besar, lalu berpisah menjadi klaster yang lebih kecil.

Agglomerative clustering relatif lebih sering digunakan dalam metode *hierarchical clustering* karena lebih sederhana dalam implementasinya. Setiap titik data yang dianggap sebagai masing-masing klaster akan dikelompokkan berdasarkan jarak kedua titik. Jarak antartitik data ini dapat dihitung dengan beberapa cara, diantaranya dengan definisi jarak *euclidean*, jarak *manhattan*, dan jarak maksimum. Setelah terbentuk beberapa klaster, selanjutnya akan dihitung jarak antarklaster. Beberapa metode yang biasa digunakan diantaranya adalah *single linkage*, *complete linkage*, *average linkage*, dan *ward linkage*. *Single linkage* memperhitungkan jarak terdekat dari dua klaster. *Complete linkage* memperhitungkan jarak terjauh dari dua klaster. *Average linkage* memperhitungkan jarak rata-rata antara setiap pasangan titik dalam dua klaster. Sementara itu, *ward linkage* menghitung jarak dengan meminimumkan variansi klaster apabila kedua klaster digabungkan.

Di sisi lain, *divisive clustering* akan membagi data yang awalnya dianggap sebagai satu klaster besar menjadi klaster-klaster yang lebih kecil. Proses membagi data di dalam klaster ini menggunakan metode klustering lain, salah satunya dapat menggunakan metode *k-means clustering*. Dari satu klaster besar, klaster akan dibagi ke dalam dua klaster menggunakan metode *k-means*, lalu masing-masing klaster kembali dibagi ke dalam dua klaster menggunakan metode yang sama, hingga pada akhirnya setiap titik data dianggap sebagai satu klaster.

K-Means Clustering

Tidak seperti metode *hierarchical clustering* yang tidak perlu menginisiasiikan jumlah klaster, metode *k-means clustering* adalah metode pengelompokan dengan mengasumsikan sejumlah titik *centroid* sesuai jumlah klaster yang diinginkan. Pada metode ini akan dihitung jarak setiap titik data terhadap setiap *centroid* yang diinisiasi. Setiap titik data akan dikelompokkan ke dalam klaster dengan jarak *centroid* terdekat. Posisi *centroid* akan berubah seiring iterasi untuk meminimumkan nilai jarak rata-rata setiap titik data dalam klaster terhadap *centroid* atau hingga nilainya konvergen.

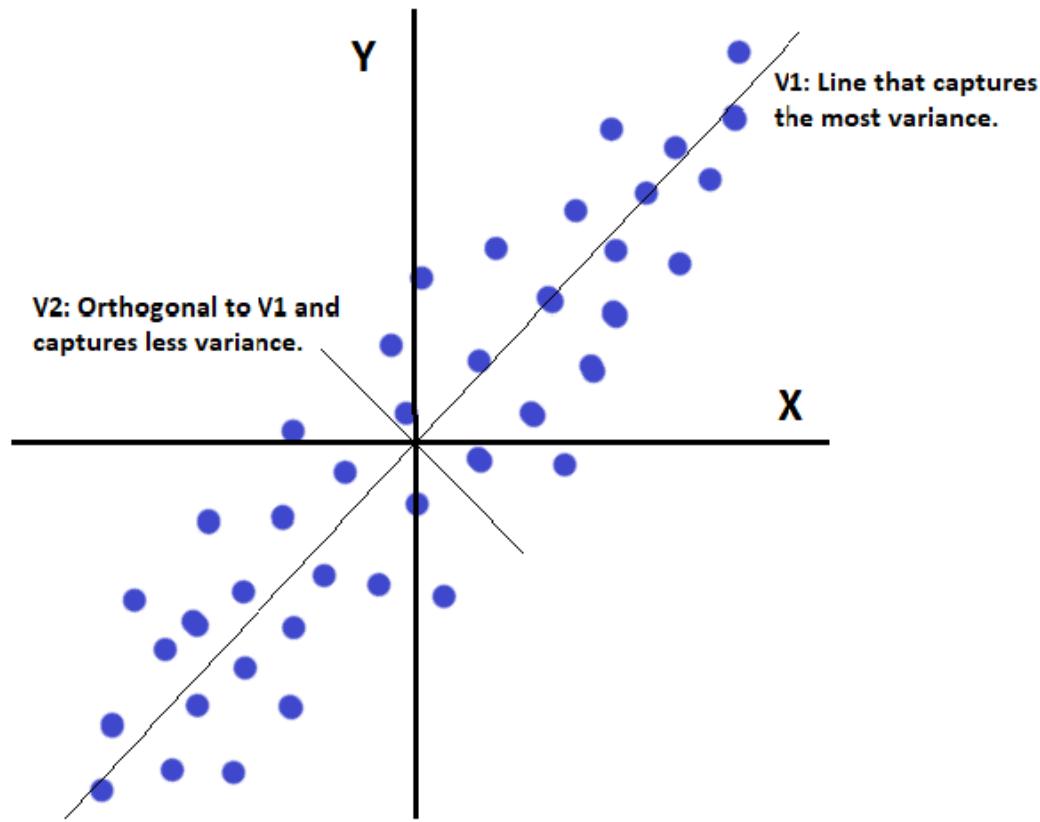
Pada metode ini, menentukan jumlah klaster di awal menjadi hal yang penting untuk dipertimbangkan. Salah satu metode untuk menentukan jumlah klaster adalah dengan menggunakan metode siku (*elbow method*). Metode ini menghitung nilai *Within Cluster Sum of Squares* (WCSS) atau inersia,

yang didefinisikan sebagai jumlah kuadrat jarak setiap titik data ke *centroid* masing-masing klaster. Nilai WCSS awalnya sangat besar, lalu perlahan mengecil seiring dengan jumlah klaster yang semakin banyak. Jika dibuat plot antara nilai WCSS terhadap jumlah klaster, seringkali terlihat penurunan yang signifikan menyerupai tekukan siku, lalu setelah itu nilai WCSS menurun dengan landai. Hal ini yang menjadi acuan untuk memutuskan bahwa pada titik siku tersebut peningkatan jumlah klaster tidak lagi menurunkan nilai WCSS dengan signifikan.

II.2.4 *Dimensionality Reduction*

Dalam proses analisis data, seringkali dijumpai sejumlah besar variabel yang membuat proses komputasi memakan waktu lebih banyak, atau apabila variabel tersebut berupa *noise* akan membuat analisis menjadi kurang akurat. Oleh karena itu, salah satu metode yang dapat dilakukan untuk meminimalisir variabel yang tidak relevan adalah dengan mereduksi dimensi data (*dimensionality reduction*). Proses ini akan membuat bentuk dataset menjadi lebih ringkas namun tetap menyimpan informasi yang penting di dalam data. Salah satu metode reduksi dimensi yang biasa digunakan adalah metode *Principle Component Analysis* (PCA).

PCA adalah metode untuk mereduksi dimensi data dengan melakukan transformasi linear terhadap data. Transformasi linear ini dilakukan dengan menghitung nilai vektor *eigen* dan nilai *eigen* dari matriks kovarians. Matriks kovarians menyimpan informasi hubungan linear antar variabel di dalam data. Secara sederhana, vektor *eigen* memberikan informasi arah variansi, sementara nilai *eigen* memberi informasi besarnya variansi pada arah tersebut. Jadi, PCA akan memilih arah dengan variansi terbesar sebagai *principle component* pertama (PC1), dan *principle component* kedua (PC2) menjadi arah yang tegak lurus terhadap PC1. Gambar II.14 menunjukkan contoh bagaimana metode PCA dalam menentukan PC1 dan PC2.



Gambar II.14: Contoh penggunaan *principle component analysis* (PCA). Sumber: <https://statisticsbyjim.com/basics/principal-component-analysis/>

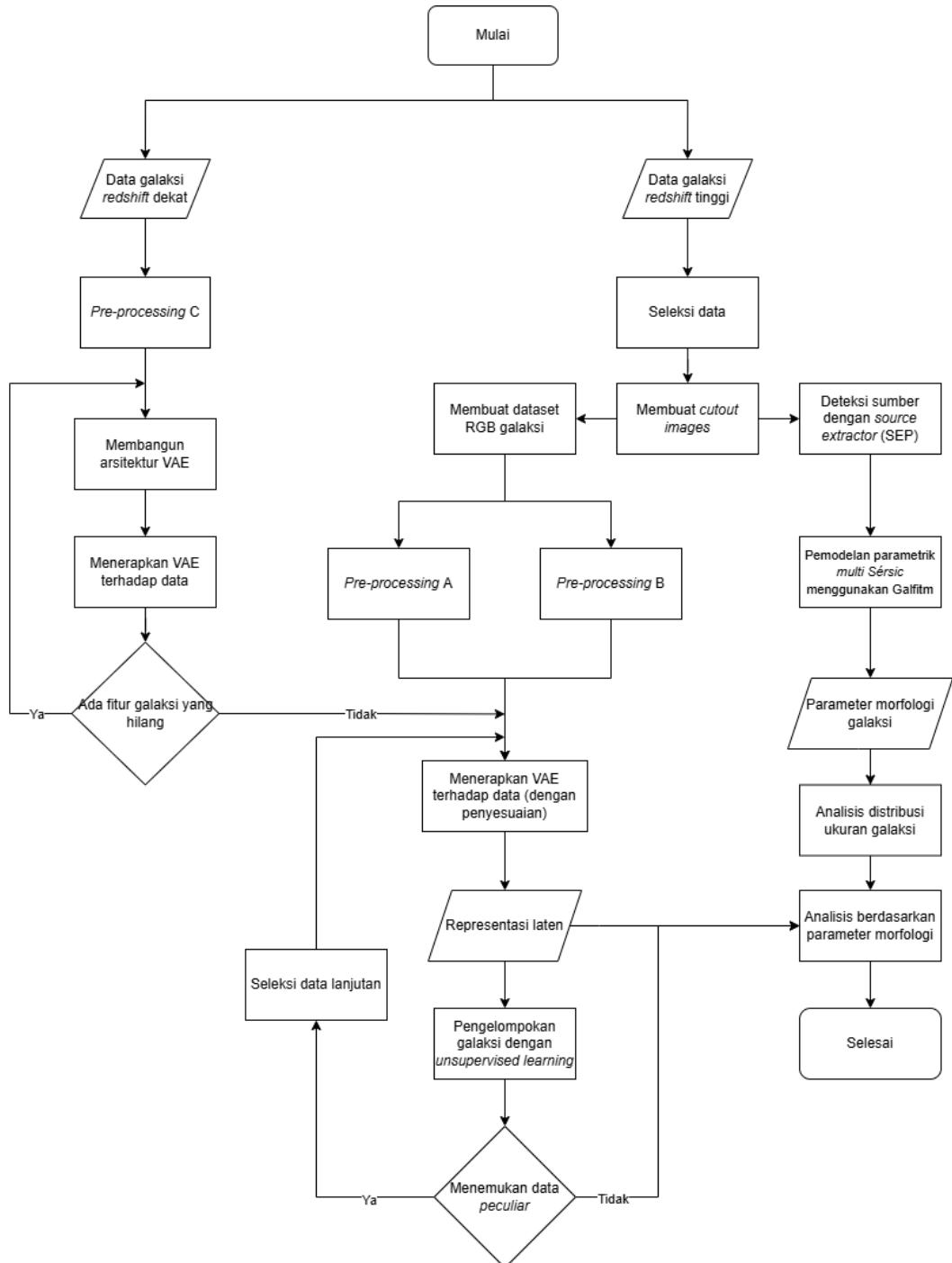
Selain PCA, metode reduksi dimensi lainnya yang dapat digunakan adalah *Uniform Manifold Approximation and Projection* (UMAP) (McInnes dkk., 2018). Jika PCA melakukan reduksi dimensi data dengan melakukan transformasi linear, UMAP melakukan reduksi dimensi data dengan transformasi non-linear. UMAP memperhitungkan jarak antara setiap titik data dan menghitung nilai kemiripannya, namun dengan mempertimbangkan jumlah *nearest neighbor* dari setiap titik data.

BAB III

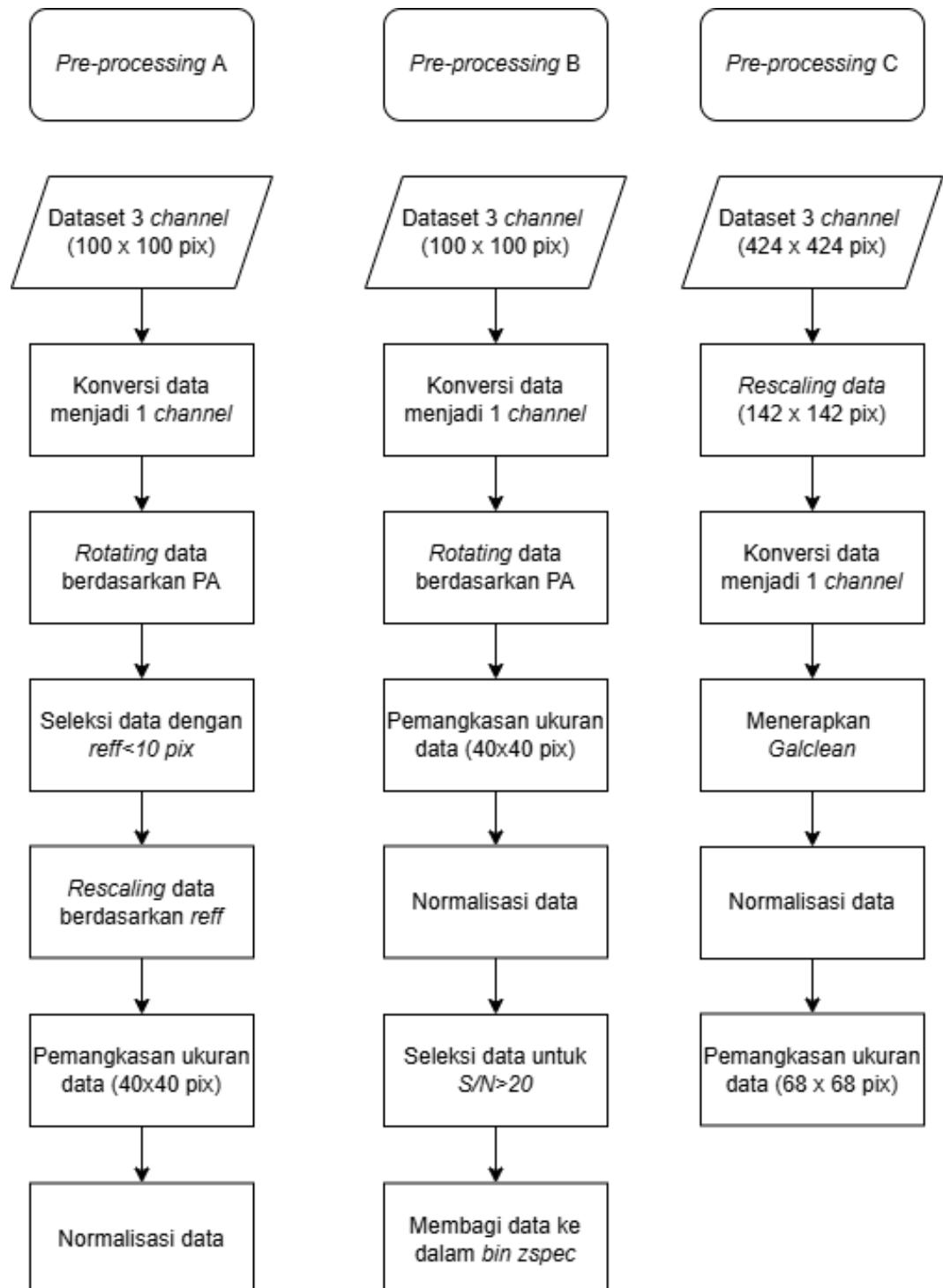
DATA DAN METODE PENELITIAN

Pada bab ini akan dibahas data yang digunakan dalam penelitian ini, dilanjutkan dengan proses seleksi dan pengolahan data. Pada bagian selanjutnya akan dibahas mengenai pemodelan parametrik, serta proses ekstraksi fitur dengan menggunakan metode VAE. Terakhir, akan dibahas metode *unsupervised learning* yang akan digunakan untuk melakukan pengelompokan galaksi.

Secara garis besar, alur berjalannya penelitian ini ditunjukkan dalam *flowchart* pada Gambar III.1. Dalam *flowchart* tersebut terdapat beberapa tahap *pre-processing* yang berbeda. Beragam tahap *pre-processing* ini dilakukan sebagai bentuk uji coba untuk mengevaluasi hasil pengelompokan galaksi. Penjelasan tentang masing-masing tahap *pre-processing* ditunjukkan dalam Gambar III.2.



Gambar III.1: Diagram alur penelitian ini.



Gambar III.2: Sebagian penjelasan diagram alur pada Gambar III.1

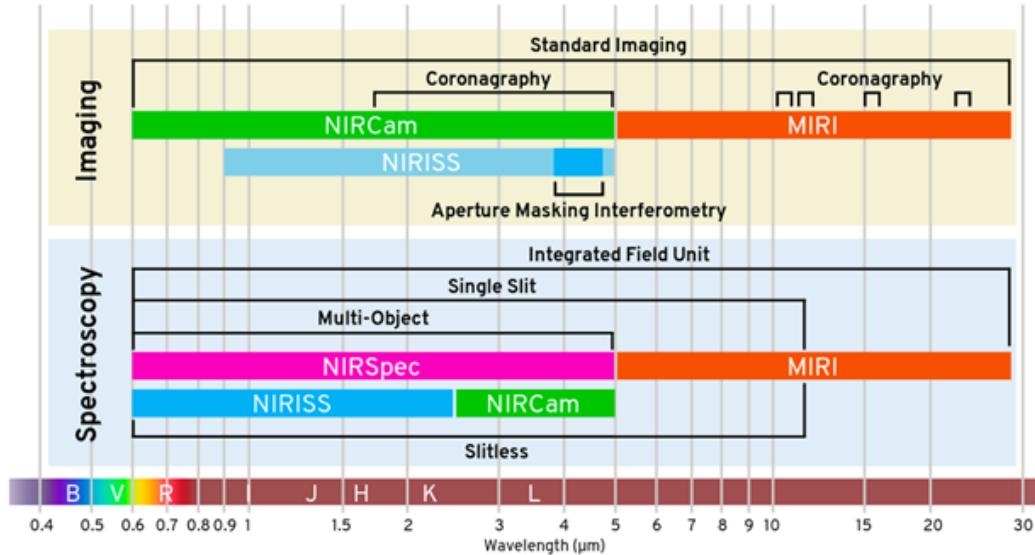
III.1 Data

Pada penelitian ini akan digunakan data galaksi *redshift* tinggi sebagai data utama, serta data galaksi *redshift* dekat untuk menjadi pembanding dalam melakukan validasi metode.

III.1.1 Data Utama

Data yang digunakan dalam penelitian ini merupakan data dari pengamatan *James Webb Space Telescope* (JWST). Teleskop luar angkasa yang diluncurkan pada tahun 2021 ini memiliki empat misi sains utama, yaitu mempelajari alam semesta dini, evolusi galaksi, siklus hidup bintang, dan deteksi eksoplanet serta mempelajari pembentukan dan evolusi sistem keplanetan.

Pada mulanya, teleskop JWST dirancang untuk melakukan pengamatan objek-objek hingga $z \sim 10$. Tetapi JWST rupanya mampu mengamati objek pada jarak yang melebihi target tersebut. Saat ini, galaksi terjauh yang telah dikonfirmasi oleh teleksop JWST berada pada $z \sim 14$. JWST telah mendeteksi 9 kandidat galaksi pada $z > 11$, namun hanya dua galaksi yang baru berhasil dikonfirmasi hingga saat ini. Dengan demikian, data galaksi pada $z > 10$ dari pengamatan JWST masih belum lengkap (*undersampled*).



Gambar III.3: Sensitifitas instrumen JWST. Sumber: <https://science.gsfc.nasa.gov/691/JWSTSSS/index.html>

JWST memiliki beberapa instrumen seperti ditunjukkan pada Gambar

III.3. Instrumen JWST, baik untuk pengamatan fotometri maupun spektroskopii, keduanya melakukan pengamatan di panjang gelombang inframerah. Dalam penelitian ini digunakan data dari pengamatan menggunakan instrumen *NIRCam*.

Data yang digunakan dalam penelitian ini merupakan data citra (*imaging*) yang telah direduksi menggunakan *pipeline grizli* (Brammer, 2023). Selain itu, digunakan data katalog *EAZY* yang merupakan hasil *fitting Spectral Energy Distribution* (SED) galaksi menggunakan program *EAZY* (Brammer dkk., 2008). Sehingga, pada penelitian ini penulis tidak menggunakan data mentah JWST, melainkan data yang telah direduksi dan diarsipkan dalam bentuk katalog *EAZY*.

Selain data *imaging*, pengamatan JWST juga menyediakan data citra PSF pada setiap filter pengamatan. Citra PSF ini menggambarkan seberapa menyebar sebuah objek titik (*point source*) ketika diamati oleh instrumen JWST. Citra PSF ini nantinya akan digunakan dalam proses *fitting* parametrik galaksi.

Selain data JWST sebagai sumber data utama, pada penelitian ini juga akan digunakan data pengamatan *Sloan Digital Sky Survey* (SDSS) untuk proses validasi metode untuk data fotometri galaksi di *redshift* dekat. Pengamatan SDSS dipilih karena data ini digunakan dalam projek *Galaxy Zoo* untuk mengelompokan morfologi galaksi dekat.

III.1.2 *Field* Pengamatan

JWST melakukan pengamatan untuk berbagai survei, dengan misi yang berbeda-beda untuk setiap surveinya. Setiap survei mengamati area yang berbeda di langit, dengan luas area yang juga berbeda. Untuk mengamati objek-objek jauh, biasanya akan dipilih area yang 'kosong', yaitu area yang jauh dari objek-objek terang seperti bintang atau galaksi dekat. Diantara berbagai survei yang dilakukan oleh JWST, empat diantaranya menjadi sumber data penelitian ini. Keempat survei tersebut yaitu *Cosmic Evolution Early Release Science* (CEERS), *Cosmic Evolution Survey* (COSMOS), *The First Reionization Epoch Spectroscopic COmplete Survey* (FRESCO), dan *Public Release IMaging for Extragalactic Research – Ultra Deep Survey* (PRIMER-UDS).

CEERS

Survei CEERS mengamati area langit pada daerah yang dinamakan *Extended Groth Strip* (EGS). Area pengamatan CEERS ini berada di antara rasi bintang

Boötes dan *Ursa Major*. Daerah langit tersebut merupakan area langit yang banyak diamati untuk misi pengamatan *deep object*. Dalam koordinat ekuatorial, survei CEERS ini berada di lokasi $RA : 14^h19^m39.59^s$ $Dec : 52^\circ52'17.84''$ (J2000), dengan luas area 23.11×7.70 menit busur.

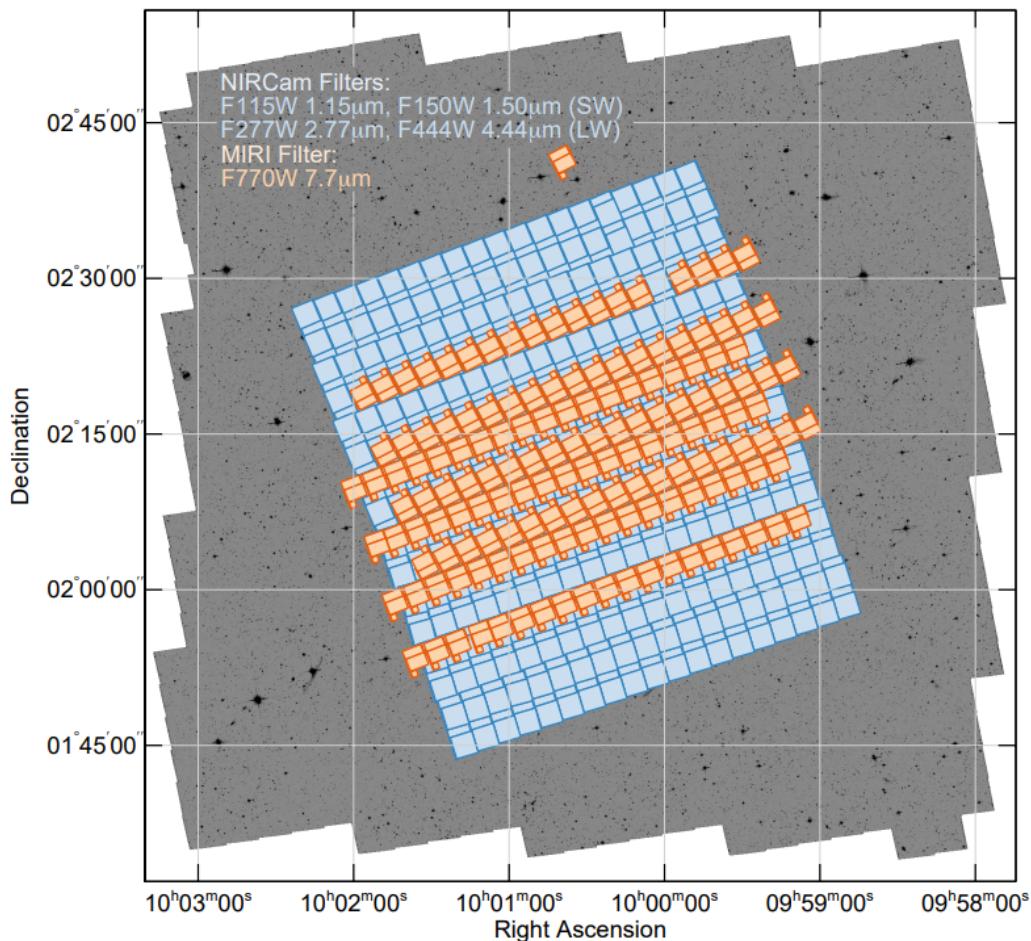


Gambar III.4: Daerah langit yang diamati oleh survei CEERS. Sumber: <https://esawebb.org/images/CEERS1/>

Survei CEERS melakukan pengamatan dengan instrumen NIRCam pada tujuh filter yang berbeda, yaitu F115W, F150W, F200W, F277W, F356W, F410M dan F444W. Total galaksi yang diamati pada *field* ini adalah sebanyak 70514 galaksi. Ukuran piksel data *imaging* yang digunakan dalam penelitian ini adalah sebesar $0.04''$ untuk seluruh filter pengamatan NIRCam dari survei CEERS.

COSMOS-Web

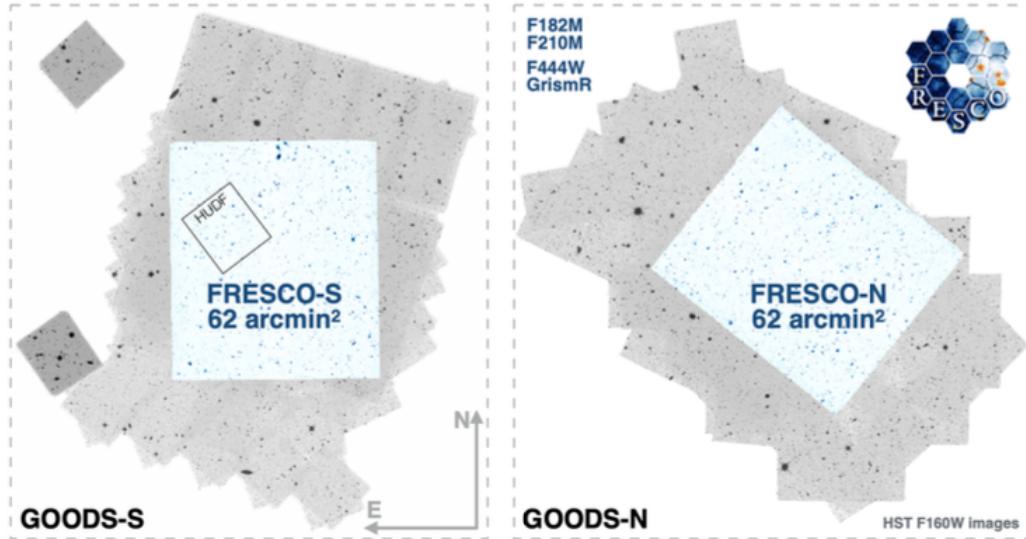
COSMOS-Web merupakan survei *imaging* dengan instrumen NIRCam pada empat filter yang berbeda, yaitu F155W, F150W, F277W, dan F444W. Pengamatan menggunakan instrumen ini dilakukan pada area seluas $0.54deg^2$, dengan pusat daerah langit yang diamati yakni pada $RA : 10^h00^m27.92^s$ $Dec : 02^\circ12'03.50''$ (J2000). Selain menggunakan instrumen NIRCam, survei COSMOS-Web juga melakukan pengamatan pada filter F770W dengan menggunakan instrumen MIRI. Dengan menggunakan instrumen MIRI, survei COSMOS-Web mengamati area sebesar $0.19deg^2$, namun pada area yang paralel seperti ditunjukkan pada Gambar III.5. Total galaksi yang diamati pada *field* COSMOS-Web ini adalah sebanyak 29358 galaksi. Ukuran piksel data *imaging* dari semua filter pengamatan NIRCam pada survei COSMOS-Web adalah sebesar $0.04''$.



Gambar III.5: Daerah langit yang diamati oleh survei COSMOS-Web. Sumber: Casey dkk. (2023)

FRESCO

FRESCO merupakan survei pengamatan pada area langit yang diamati oleh *The Great Observatories Origins Deep Survey* (GOODS). GOODS merupakan survei yang dilakukan Hubble dan teleskop luar angkasa lainnya pada misi pengamatan objek *deep sky*. Survei FRESCO melakukan pengamatan pada area *GOODS-South* dan *GOODS-North*. Namun, pada penelitian ini data yang digunakan hanya berasal dari pengamatan pada area *GOODS-South*.

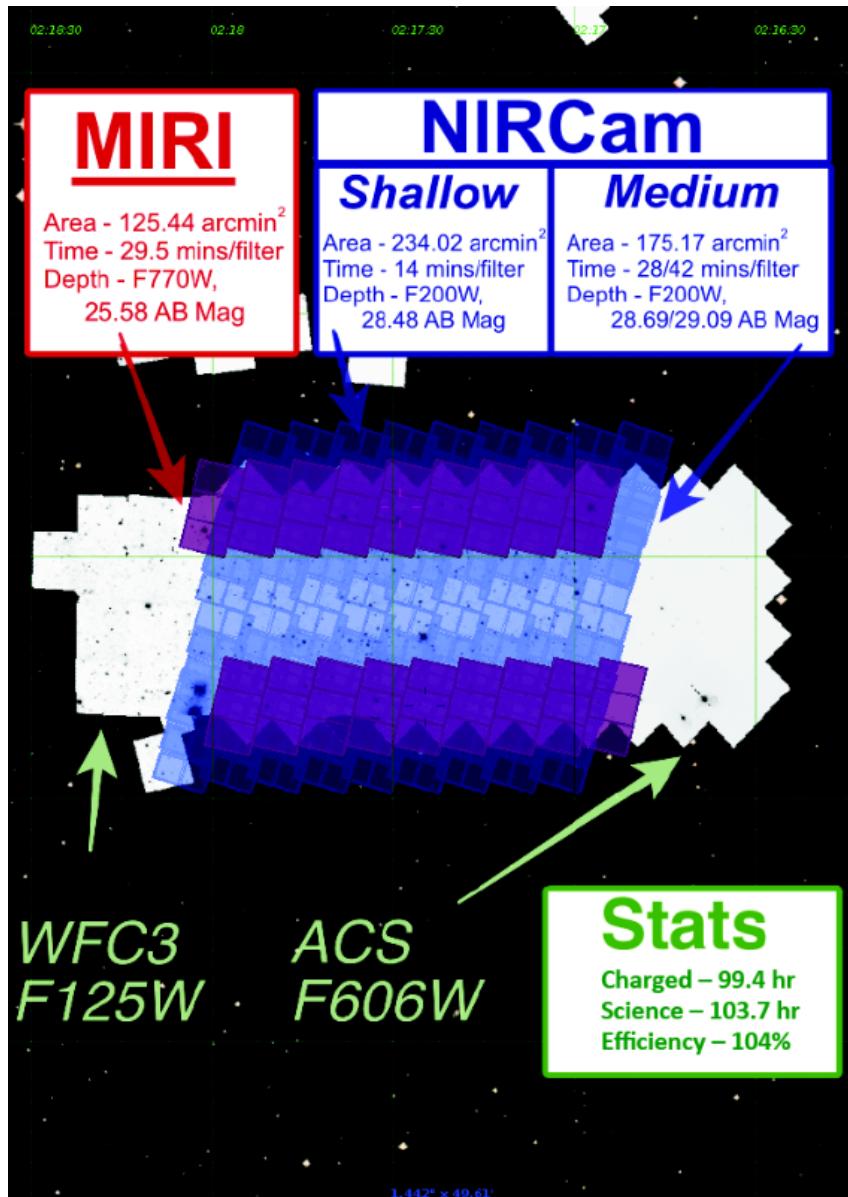


Gambar III.6: Daerah langit yang diamati oleh survei FRESCO. Sumber: <https://archive.stsci.edu/hlsp/fresco>

Area langit *GOODS-South* berada pada lokasi $RA : 3^h32^m31^s$ $Dec : -27^\circ48'04''$ (J2000) (koordinat di bagian tengah *field* pengamatan). Survei pada area ini memiliki luas area pengamatan 15.90×9.71 menit busur. Dalam survei ini, instrumen NIRCam melakukan pengamatan pada 14 filter berbeda, yaitu F090W, F115W, F150W, F182M, F200W, F210M, F277W, F335M, F356W, F410M, F444W, F460M, F480M. Ukuran piksel data *imaging* yang digunakan dalam penelitian ini adalah sebesar $0.02''$ untuk filter panjang gelombang pendek (F090W, F115W, F150W, F182M, F200W, F210M) dan $0.04''$ untuk filter panjang gelombang panjang (F277W, F335M, F356W, F410M, F444W, F460M, F480M). Survei FRESCO pada area *GOODS-South* mendeteksi sejumlah 50820 galaksi.

PRIMER

Survei PRIMER melakukan pengamatan dengan instrumen NIRCam dan MIRI pada daerah pengamatan COSMOS dan UDS (*Ultra Deep Survey*). Namun, pada penelitian ini data survei PRIMER yang digunakan hanya data yang berasal dari *field* UDS dan dengan instrumen NIRCam. *Field* UDS ini berada pada lokasi $RA : 2^h17^m37.69^s$ $Dec : -5^\circ13'13.41''$ (J2000). Pengamatan NIRCam dalam survei ini dilakukan pada delapan filter, yaitu F090W, F115W, F150W, F200W, F277W, F356W, F410M, F444W. Data pada setiap filter memiliki ukuran $0.04''$ per piksel.

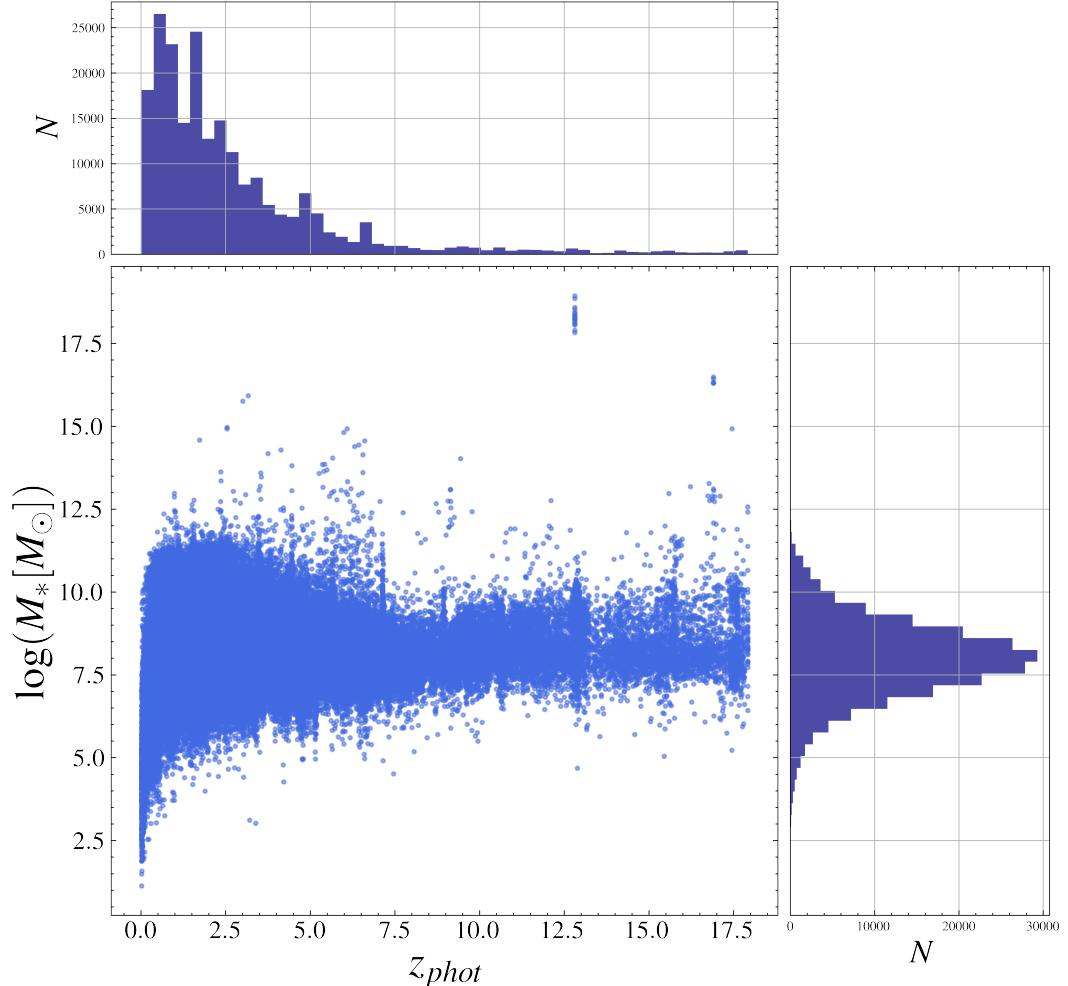


Gambar III.7: Daerah langit pada *field* UDS yang diamati oleh survei PRIMER. Sumber: <https://primer-jwst.github.io/observations.html>

III.1.3 Katalog *EAZY*

EAZY (*Easy and Accurate Redshifts from Yale*) merupakan program pengukuran *redshift* dengan metode fotometri (Brammer dkk., 2008). Secara sederhana, *EAZY* akan melakukan *fitting SED* dengan membandingkan data fotometrik objek terhadap data fotometrik buatan. Program *EAZY* digunakan ketika tidak ada data pengukuran *redshift* menggunakan metode spektroskopi, atau ketersediaan data *redshift* spektroskopi yang sedikit dibandingkan seluruh

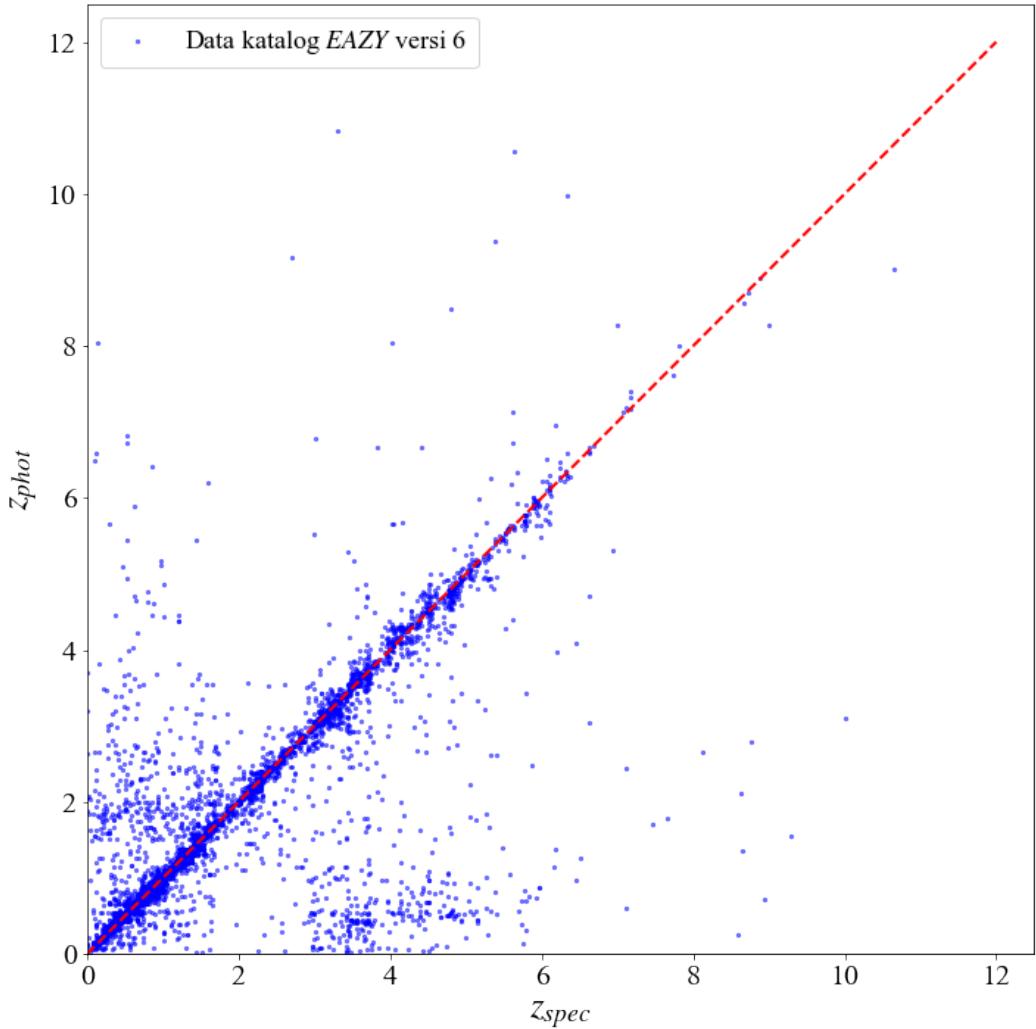
data objek yang dimiliki. Selain *EAZY*, terdapat beberapa program lainnya yang telah dikembangkan untuk melakukan *fitting SED*, seperti *piXedfit* (Abdurro'uf dkk., 2021) dan *dense basis* (Iyer dkk., 2019).



Gambar III.8: Plot distribusi massa terhadap *redshift* untuk data dari keempat survei.

Dalam penelitian ini, digunakan data pengamatan dari keempat survei yang telah dilakukan *fitting* dan diarsipkan sebagai sebuah katalog galaksi. Katalog ini menyimpan informasi berbagai parameter untuk setiap galaksi, diantaranya posisi setiap galaksi, massa, *redshift*, *star formation rate* (SFR), dan banyak parameter lainnya. Gambar III.8 menunjukkan sebaran massa terhadap *redshift* dari katalog *EAZY* untuk keempat survei. Pada Gambar III.8 terlihat distribusi massa galaksi-galaksi dari pengamatan keempat survei membentuk distribusi Gaussian, dengan sebagian besar galaksi memiliki massa $\sim 10^9 M_{\odot}$. Selain itu, dari gambar tersebut juga terlihat sebagian besar galaksi berada pada $z \lesssim 3$, dan pada *redshift* yang semakin tinggi jumlahnya semakin sedikit.

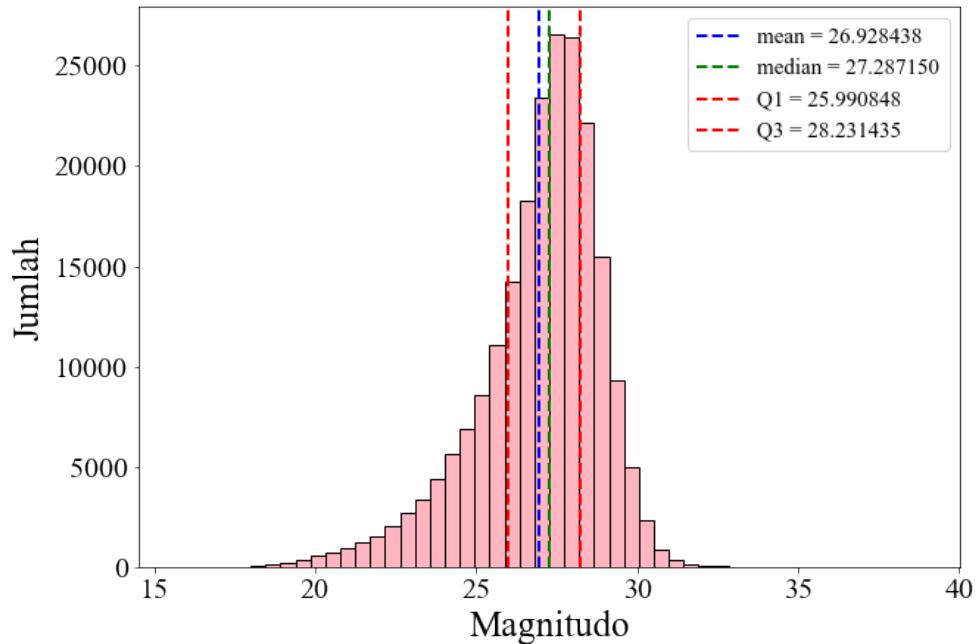
Perlu diperhatikan bahwa sebagaimana disebutkan sebelumnya, JWST baru dapat mengonfirmasi galaksi terjauh berada pada $z \sim 14$. Oleh karena itu, nilai *redshift* fotometri dari katalog *EAZY* tidak sepenuhnya akurat, terutama untuk galaksi-galaksi di *redshift* tinggi. Data galaksi-galaksi pada $z_{phot} \gtrsim 10$ seharusnya masih *undersampled* dengan pengamatan teleskop JWST yang memiliki limit pengamatan pada $z \sim 10$.



Gambar III.9: Plot yang menunjukkan perbandingan nilai *redshift* dari pengukuran fotometri dan dari pengukuran spektroskopi

Pengukuran *redshift* menggunakan metode spektroskopi memberikan nilai *redshift* yang lebih akurat dibandingkan dengan metode *fitting* fotometri. Namun, pengukuran *redshift* dengan metode spektroskopi membutuhkan *exposure time* yang lebih panjang dibandingkan pengamatan fotometri, sehingga tidak banyak galaksi yang memiliki informasi *redshift* spektroskopi. Pada Gam-

bar III.9 ditunjukkan perbandingan *redshift* fotometri dan spektroskopi untuk ~ 6000 galaksi yang memiliki data *redshift* spektroskopi. Pada Gambar III.9 terlihat nilai *redshift* dari pengukuran fotometri dan pengukuran spektroskopi cukup sebanding, meski masih terdapat sebagian *outliers*.



Gambar III.10: Plot distribusi magnitudo galaksi dari data katalog *EAZY*. Garis berwarna biru dan hijau menunjukkan nilai rata-rata dan median dari distribusi tersebut, sedangkan garis berwarna merah menunjukkan quartil 1 dan quartil 3 dari distribusi magnitudo.

Parameter lainnya yang juga terdapat di dalam katalog *EAZY* adalah magnitudo galaksi, tepatnya magnitudo absolut galaksi (*AB Magnitude*). Magnitudo absolut adalah sistem magnitudo monokromatik, yang diukur berdasarkan pengukuran densitas fluks. Magnitudo absolut mengukur kecerlangan galaksi yang konsisten di seluruh spektrum elektromagnetik, karena berbasis pada fluks energi per satuan frekuensi. Sistem magnitudo ini banyak digunakan pada data galaksi yang diamati dalam berbagai panjang gelombang atau filter. Gambar III.10 menunjukkan histogram data magnitudo galaksi yang tercatat pada katalog *EAZY*. Sebagian besar galaksi di katalog ini tercatat memiliki magnitudo $\sim 25 - 29$, dengan rata-rata magnitudo galaksi yang tidak jauh berbeda dengan nilai mediannya, yakni di sekitar magnitudo ~ 27 .

III.1.4 Katalog Galaksi Zoo

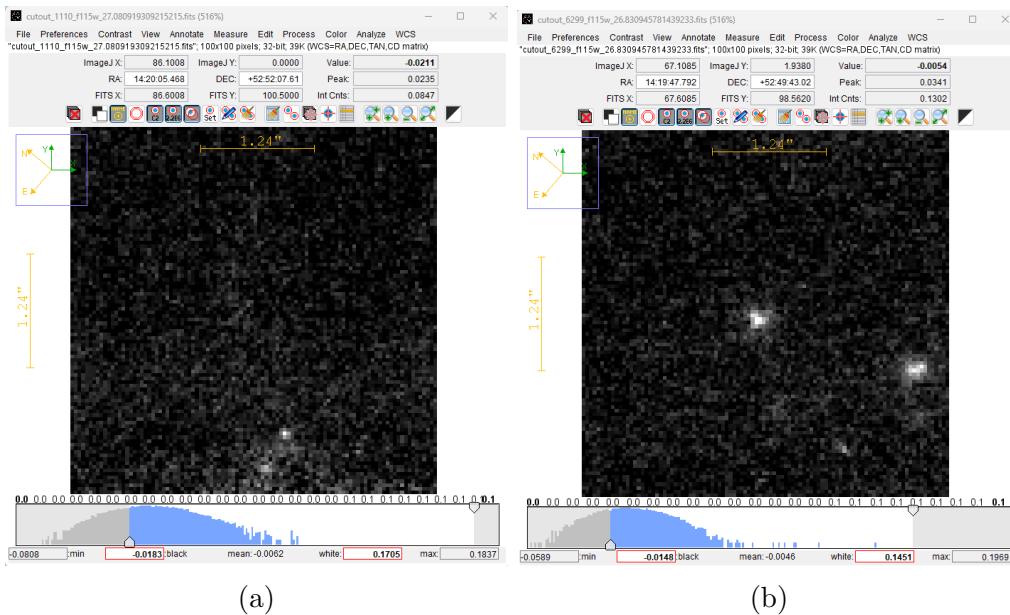
Pada penelitian ini digunakan katalog morfologi galaksi dari proyek *Galaxy Zoo 2* (Willett dkk., 2013; Hart dkk., 2016) menggunakan citra galaksi dari SDSS. Terdapat sebanyak 14060 gambar galaksi dalam format RGB yang telah diklasifikasikan ke dalam sejumlah kategori morfologi. Katalog morfologi galaksi ini digunakan sebagai acuan seberapa baik metode pengelompokan galaksi dapat dilakukan pada penelitian ini. Galaksi dalam katalog ini berada pada rentang $0.0005 < z < 0.25$, dan secara visual memiliki struktur yang lebih kompak dibandingkan galaksi jauh yang menjadi objek utama penelitian ini.

III.2 Seleksi Data JWST

Data yang telah dijelaskan dalam Subbab III.1 akan melewati serangkaian tahap seleksi berdasarkan beberapa kriteria. Selain seleksi untuk menyisihkan data karena beberapa faktor, pada Subbab ini juga akan dijelaskan seleksi data ke dalam kategori galaksi *star forming* dan *quiescent*.

III.2.1 Seleksi Data Secara Umum

Dari 216098 galaksi dari keempat survei, akan dilakukan seleksi data berdasarkan beberapa kriteria. Kriteria pertama adalah dipilih galaksi-galaksi pada $z > 2$. Tujuan penelitian ini adalah untuk mempelajari struktur galaksi pada *redshift* tinggi, maka dipilih batas *redshift* yang lebih tinggi dibandingkan penelitian-penelitian terdahulu. Kriteria yang kedua untuk menyeleksi data adalah dipilih galaksi-galaksi dengan magnitudo < 27 . Batas ini dipilih setelah melihat sampel beberapa galaksi di sekitar nilai rata-rata dan median distribusi magnitudo dalam katalog *EAZY* yang ditunjukkan pada Gambar III.10. Gambar III.11 menunjukkan contoh galaksi di sekitar magnitudo 27. Selain kedua kriteria tersebut, pada penelitian ini penulis hanya mengambil galaksi yang memiliki data pengamatan pada tiga filter, yaitu F115W, F277W, dan F444W. Pemilihan filter ini bertujuan untuk menyamakan data yang digunakan dari keempat survei. Dari ketiga kriteria tersebut, total galaksi yang menjadi objek penelitian ini berjumlah 26.133 galaksi, dengan jumlah galaksi di setiap *field* ditunjukkan pada Tabel III.1.

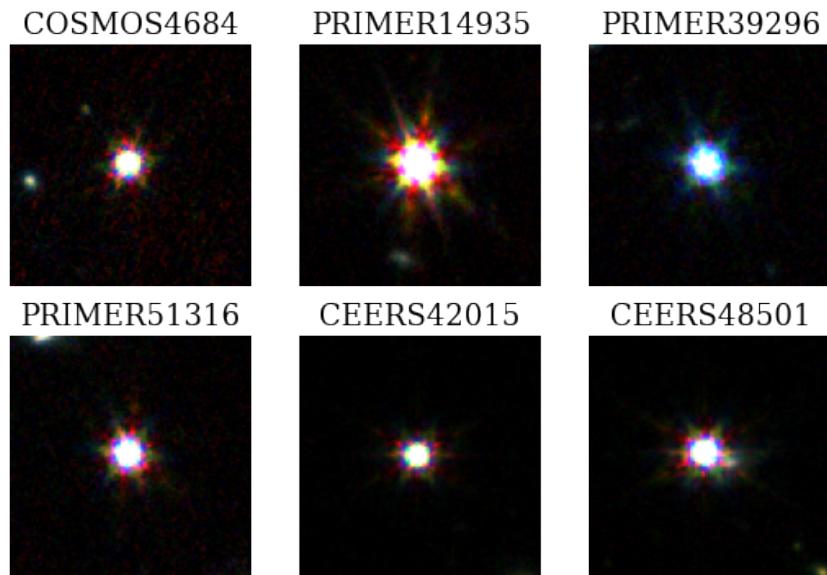


Gambar III.11: Contoh sebuah galaksi dengan magnitudo 27.089 (panel a) dan contoh sebuah galaksi dengan magnitudo 26.83 (b).

Tabel III.1: Jumlah galaksi dari keempat survei setelah dilakukan seleksi *redshift* dan magnitudo.

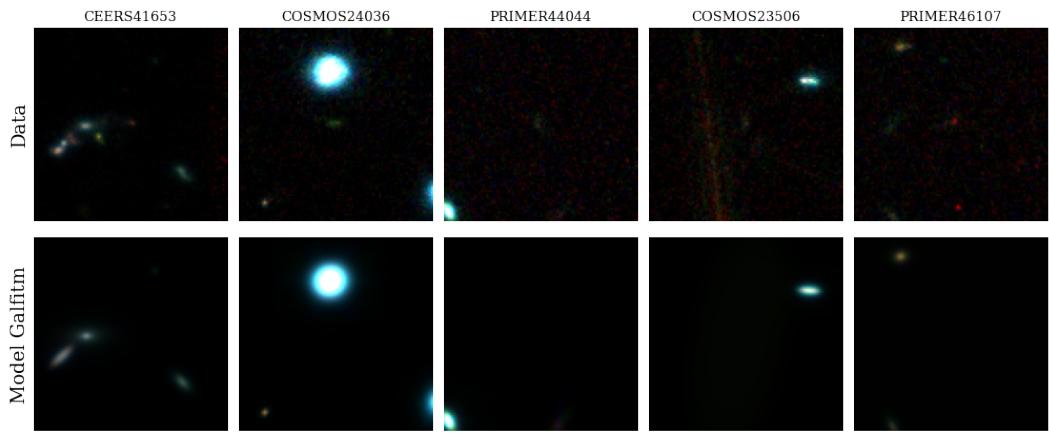
Field Pengamatan	Jumlah Galaksi
CEERS	8030
COSMOS-Web	4901
FRESCO	2460
PRIMER	10742

Selain seleksi berdasarkan beberapa kriteria diatas, seleksi data tambahan juga akan dilakukan salah satunya untuk menyisihkan data objek bukan galaksi. Karena jumlah data yang sangat banyak, menyisihkan data objek bukan galaksi secara visual memakan waktu yang sangat panjang, sehingga seleksi dilakukan dengan memanfaatkan arsitektur VAE serta metode klastering yang sama dengan yang akan digunakan untuk mencapai tujuan utama penelitian ini. Gambar III.12 menunjukkan beberapa sampel data objek bukan galaksi yang didapat dari proses klastering awal.



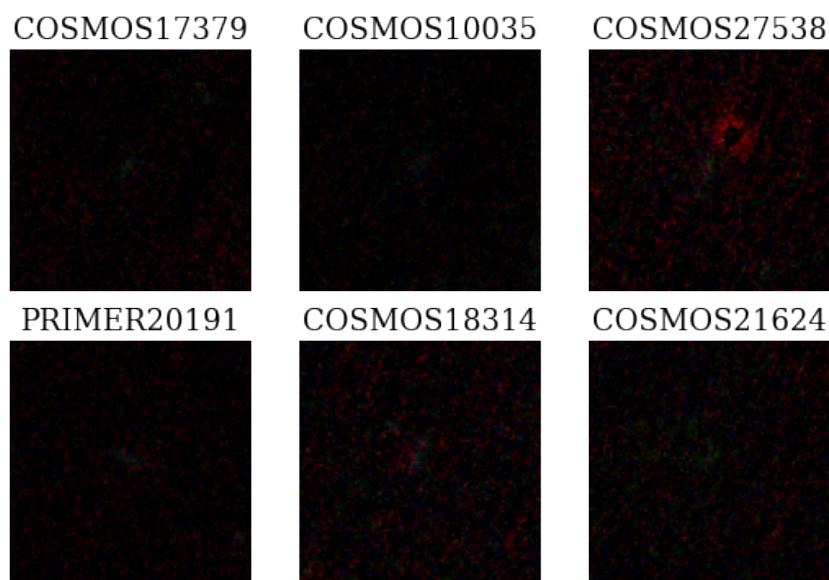
Gambar III.12: Beberapa sampel objek bukan galaksi yang didapat dari proses klastering awal.

Selain itu, seleksi data juga dilakukan untuk menyisihkan galaksi dengan hasil *fitting* yang tidak akurat. Kriteria tidak akurat disini adalah objek target, yaitu objek yang berada di posisi tengah citra, tidak berhasil dideteksi oleh **source extractor** dengan *threshold* yang diberikan. Kasus seperti ini terjadi ketika objek terlalu redup, sementara terdapat objek yang jauh lebih terang didekatnya. Kasus seperti ini juga ditemukan pada data dengan nilai S/N yang rendah. Seleksi ini kembali dilakukan dengan memanfaatkan arsitektur VAE dan proses klastering. Gambar III.13 menunjukkan contoh beberapa galaksi target yang tidak berhasil dideteksi, sehingga tidak dimodelkan oleh **Galfitm**. Seleksi ini dilakukan karena salah satu tahap *pre-processing data* yang dapat dilakukan adalah *rescaling* berdasarkan informasi radius efektif galaksi dari proses *fitting*. Maka jika radius efektif dari proses *fitting* kurang tepat, proses *rescaling* nantinya juga akan keliru.



Gambar III.13: Beberapa sampel galaksi yang gagal dimodelkan oleh Galfitm karena objek target terlalu redup sehingga tidak dimodelkan.

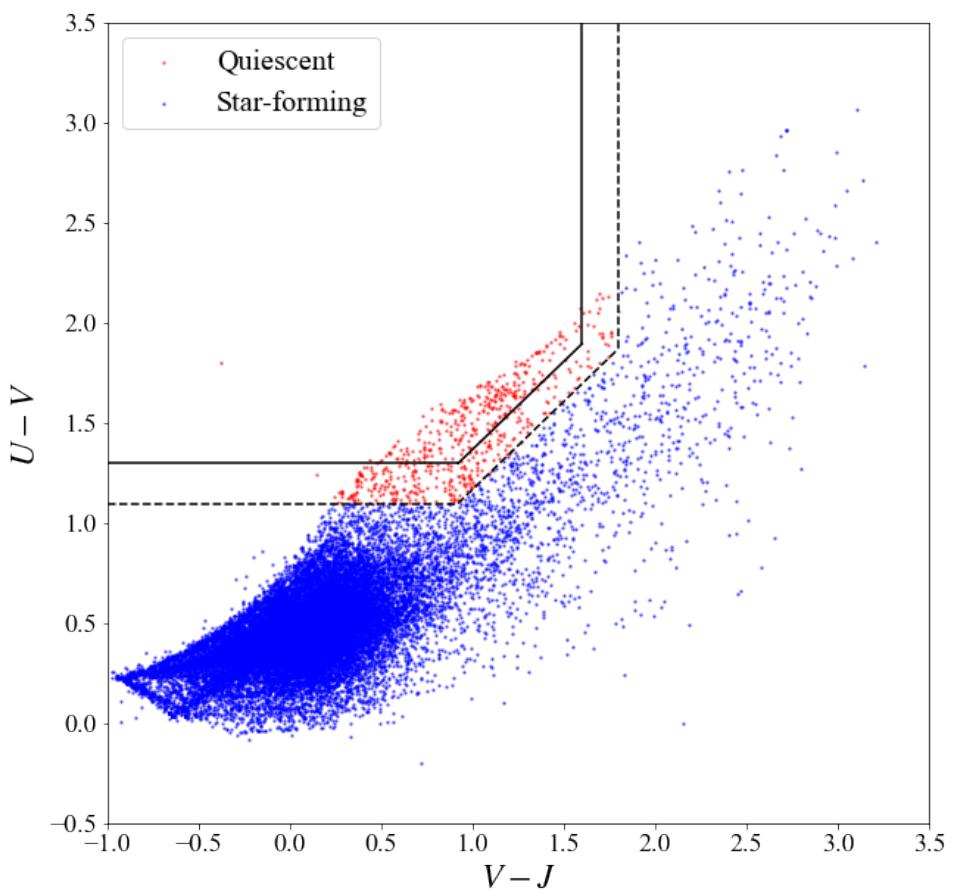
Seleksi terakhir juga dilakukan untuk menyisihkan data yang gagal diberesihkan dengan menggunakan aplikasi `Galclean` (Ferreira, 2018). Metode pembersihan ini lebih lanjut akan dijelaskan pada Subbab III.3.3. Kasus seperti ini dijumpai pada galaksi-galaksi yang terlalu redup atau memiliki *noise* yang cukup besar. Gambar III.14 menunjukkan beberapa galaksi yang tidak terdeteksi oleh `Galclean` karena terlalu redup. `Galclean` adalah aplikasi yang dikembangkan untuk menghilangkan objek terang yang berada di dekat galaksi target, dengan menggunakan paket `PhotUtils` dari `Astropy`.



Gambar III.14: Beberapa sampel galaksi yang tidak terdeteksi oleh Galclean karena terlalu redup.

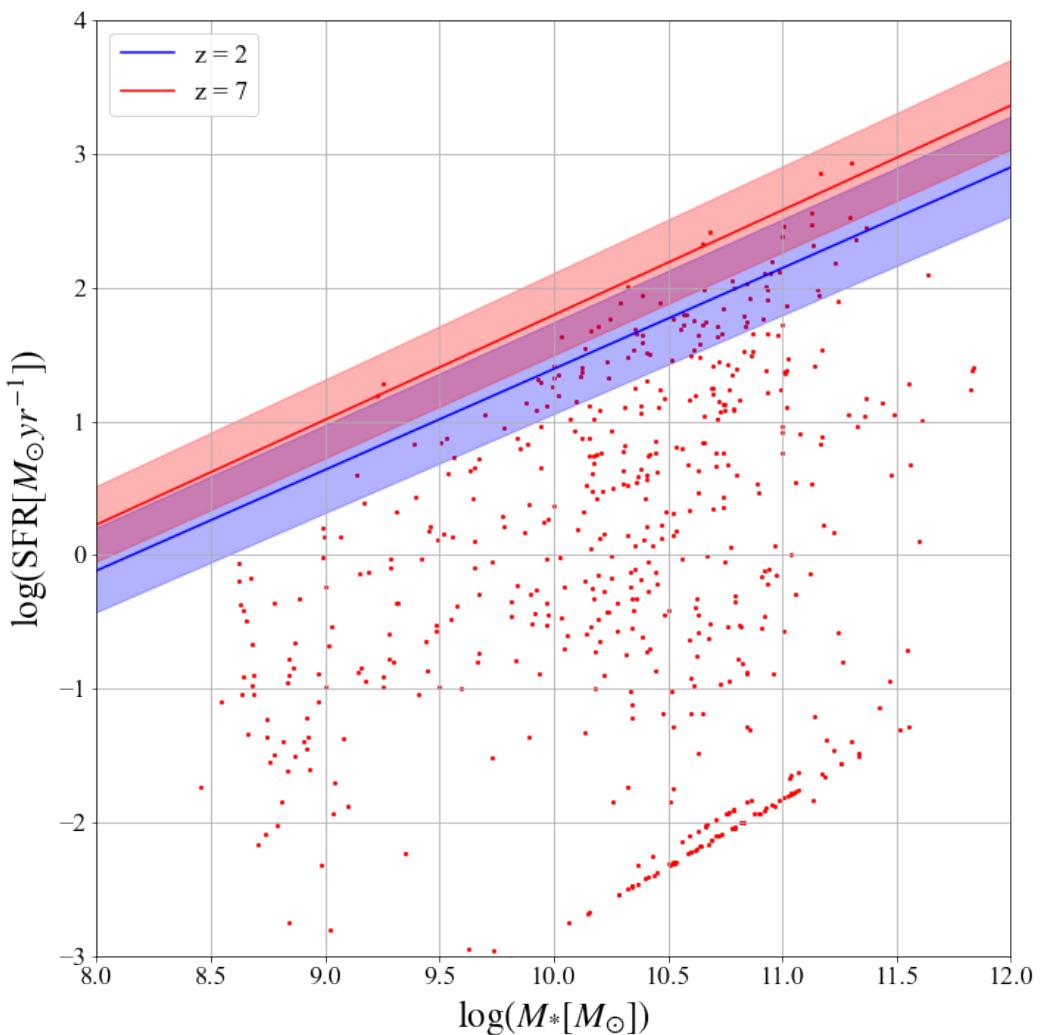
III.2.2 Seleksi Data Berdasarkan Pembentukan Bintang di Dalam Galaksi

Seleksi data lainnya yang akan dilakukan adalah mengkategorikan galaksi *quiescent* dan galaksi *star forming* dari sampel data yang sebelumnya telah diseleksi. Tidak seperti seleksi sebelumnya yang bertujuan untuk menyisihkan data, seleksi kali ini bertujuan sebatas membagi data ke dalam kedua kategori galaksi berdasarkan pembentukan bintang di dalamnya. Seleksi data ini dilakukan dengan metode yang sama seperti yang dijelaskan pada subbab II.1.3, yakni dengan menggunakan diagram UVJ dan kurva laju pembentukan bintang. Gambar III.16 menunjukkan seleksi data pada penelitian ini berdasarkan diagram warna. Dari hasil seleksi ini, terdapat 586 galaksi *quiescent* dan 25547 galaksi *star forming*.



Gambar III.15: Seleksi data berdasarkan diagram warna pada penelitian ini

Setelah melakukan seleksi berdasarkan diagram UVJ, data galaksi *quiescent* selanjutnya melalui seleksi berdasarkan diagram SFR terhadap massa, ditunjukkan pada Gambar III.16. Berdasarkan seleksi ini, seluruh data tampak berada di bawah rentang ketidakpastian kurva, sehingga seleksi data tetap memberikan hasil akhir 586 galaksi *quiescent* dan 25547 galaksi *star forming*.



Gambar III.16: Seleksi data berdasarkan diagram hubungan SFR dan massa

III.3 Pengolahan Data JWST

Pada bagian ini akan dibahas proses pembuatan *cutout images* dari data area survei galaksi di keempat *field*. Kemudian dilanjutkan dengan penjelasan mengenai pemodelan parametrik yang dilakukan untuk memperoleh parameter struktur galaksi. Selanjutnya akan dibahas mengenai beberapa tahap *pre-processing data* yang dapat dilakukan terhadap data yang sebelumnya telah diseleksi. Terakhir akan dijelaskan mengenai metode VAE yang akan diaplikasikan terhadap data serta pengelompokan galaksi berdasarkan fitur-fitur yang dipelajari mesin melalui tahap VAE.

III.3.1 Pembuatan *Cutout Images*

Dari data *imaging* yang telah direduksi *grizli* dan data dari katalog *EAZY*, selanjutnya akan dibuat *cutout images* untuk setiap galaksi. Proses pembuatan citra ini dilakukan dengan mengambil potongan kecil area yang berpusat di lokasi setiap galaksi berdasarkan RA dan deklinasinya. *Cutout* galaksi dari data *imaging* yang memiliki ukuran piksel 0.04 akan dibuat dalam ukuran 100×100 piksel, sementara *cutout* dari data *imaging* dengan ukuran piksel 0.02 akan dibuat dalam ukuran 200×200 piksel.

Data *imaging* yang tersedia dalam penelitian ini terdiri dari *science image* dan *weight image*. *Science image* adalah data pengamatan yang berisi hasil pengukuran parameter-parameter sains yang relevan, sementara *weight image* adalah data yang berisi bobot relatif setiap piksel. *Science image* akan digunakan dalam pemodelan parametrik dan digunakan sebagai *input* dalam proses VAE, sementara *weight image* hanya akan digunakan dalam proses pemodelan parametrik. Namun, *weight image* yang akan digunakan dalam proses pemodelan merupakan data yang dikonversi menjadi nilai variansi (*sigma*), dan dibuat *cutout* sebagaimana yang dilakukan dalam pembuatan *cutout image* dari *science image*. Dengan demikian, data *weight image* yang telah dibuat *cutout* dan dikonversi sebagai nilai variansi selanjutnya disebut sebagai *sigma image*.

III.3.2 Pemodelan Parametrik

Parameter fisis galaksi dapat diperkirakan salah satunya melalui pencocokan data pengamatan galaksi terhadap suatu model kecerlangan. Model galaksi yang paling sederhana yang dapat dibuat adalah model galaksi berbentuk elipsoid. Dalam penelitian ini, akan dilakukan proses pemodelan parametrik galaksi menggunakan aplikasi **Galfitm** (Vika dkk., 2013). **Galfitm** adalah aplikasi pemodelan parametrik galaksi yang dikembangkan dari aplikasi **Galfit** (Peng dkk., 2002, 2010). Hal yang membedakan antara **Galfit** dan **Galfitm** adalah **Galfitm** dapat melakukan pemodelan secara simultan terhadap data galaksi di beberapa filter. Dengan demikian, **Galfitm** cocok digunakan dalam pemodelan yang melibatkan data pengamatan multi panjang gelombang. Penelitian Effendi (2024) menunjukkan bahwa hasil pemodelan parametrik menggunakan **Galfitm** menghasilkan model yang lebih cocok dengan pengamatan karena pemodelan secara sinkron dapat mengikat model di berbagai filter. Pada penelitian ini akan dilakukan pemodelan dengan *multi Sérsic*

untuk mengatasi kasus dimana terdapat lebih dari satu galaksi dalam satu *cutout images*.

Salah satu *input* yang dibutuhkan untuk melakukan *fitting* menggunakan **Galfitm** adalah posisi galaksi pada citra dalam koordinat kartesian (x,y). Pada dasarnya, posisi galaksi target seharusnya berada di tengah citra, yang berarti pada posisi (50,50) untuk data dengan ukuran 100×100 piksel. Namun nyatanya, posisi galaksi tidak selalu tepat di tengah, melainkan sedikit bergeser dari posisi (50,50). Selain itu, untuk melakukan *fitting multi Sérsic* diperlukan informasi posisi setiap galaksi tetangga dalam setiap citra. Untuk mengetahui posisi setiap galaksi tetangga pada setiap citra, akan dilakukan deteksi sumber cahaya menggunakan paket **Source Extractor** (Bertin dan Arnouts , 1996). **Source Extractor** adalah aplikasi untuk mendeteksi dan mengukur objek langit seperti bintang dan galaksi dari data pengamatan area langit. Aplikasi ini dapat mengidentifikasi sumber cahaya melalui pendekripsi piksel-piksel dengan fluks yang melebihi ambang batas tertentu. Oleh karenanya, aplikasi ini dapat memberikan informasi posisi setiap galaksi dalam setiap *cutout images*. Pada penelitian ini, digunakan **Source Extractor** yang dikembangkan dalam bentuk paket **python** bernama **SEP** (Barbary, 2016).

III.3.3 *Pre-Processing Data*

Sebelum mulai melakukan pengkodean, akan dilakukan tahap persiapan data untuk meminimalisir bias dan *noise* dari data. Mulai dari mempersiapkan dataset dari *cutout images* yang telah dibuat, melakukan *rotating data*, memangkas ukuran data hingga melakukan normalisasi terhadap dataset.

Dari *cutout images* berukuran 100×100 piksel pada masing-masing filter, pertama akan dibuat dataset RGB untuk memudahkan proses VAE dan pengelompokan selanjutnya. Dari keempat *field* yang digunakan, *field FRESCO* memiliki ukuran piksel yang berbeda pada filter F115W, sehingga tidak memungkinkan untuk membangun citra RGB untuk *field FRESCO*. Selain itu, dari hasil *fitting multi Sérsic* didapatkan bahwa tidak semua galaksi dapat dimodelkan oleh aplikasi **Galfitm**. Pada sebagian galaksi, aplikasi tidak bisa membangun model kecerlangan yang konvergen. Dengan demikian, dataset yang dibuat hanya memasukkan data galaksi yang berhasil di-*fitting* oleh **Galfitm** dan hanya dari tiga *field* pengamatan. Sehingga dataset pada akhirnya hanya berukuran $15130 \times 100 \times 100 \times 3$, yang berarti dalam dataset tersebut terdapat data fotometri 15130 galaksi berukuran 100×100 untuk tiga *channel* warna, secara berurutan dimulai dari data pada filter F115W, F277W, dan F444W.

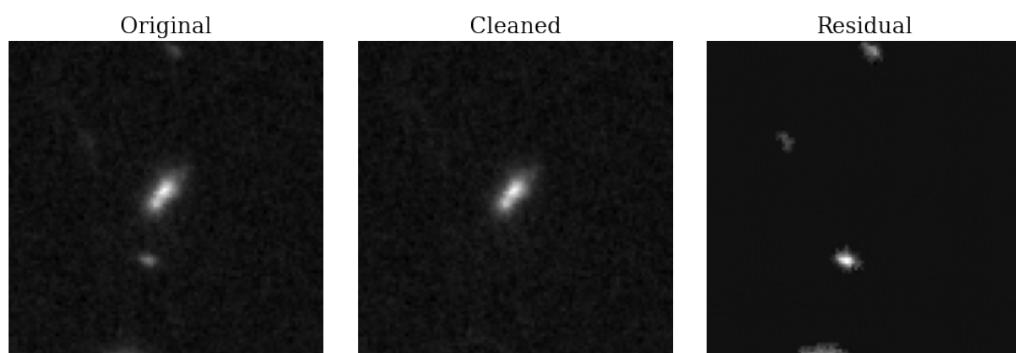
Dataset ini sudah dapat digunakan dalam proses VAE, namun beberapa tahap *pre-processing* dapat diaplikasikan terhadap dataset sebelum dijadikan input dalam proses VAE.

Perlu ditekankan bahwa beberapa tahap *pre-processing* yang dijelaskan pada Subbab ini merupakan beberapa alternatif proses yang bisa dilakukan untuk mencapai tujuan pengelompokan galaksi. Lebih lanjut mengenai tahapan *pre-processing* yang benar-benar diaplikasikan akan dijelaskan pada Bab IV.

Pembersihan Data

Salah satu masalah yang seringkali dijumpai dalam melakukan pengelompokan morfologi galaksi adalah ketika dalam satu citra terdapat lebih dari satu objek. Kita mengharapkan mesin hanya mempelajari bentuk dari satu galaksi dalam citra, sehingga apabila terdapat galaksi tetangga dikhawatirkan mesin mengelompokkan galaksi berdasarkan jumlah objek di dalam satu citra. Untuk meminimalisir hal ini, salah satu cara yang dapat dilakukan adalah pemangkasan ukuran citra semaksimal mungkin, sehingga dalam setiap citra hanya terdapat satu objek yang menjadi target utama kita.

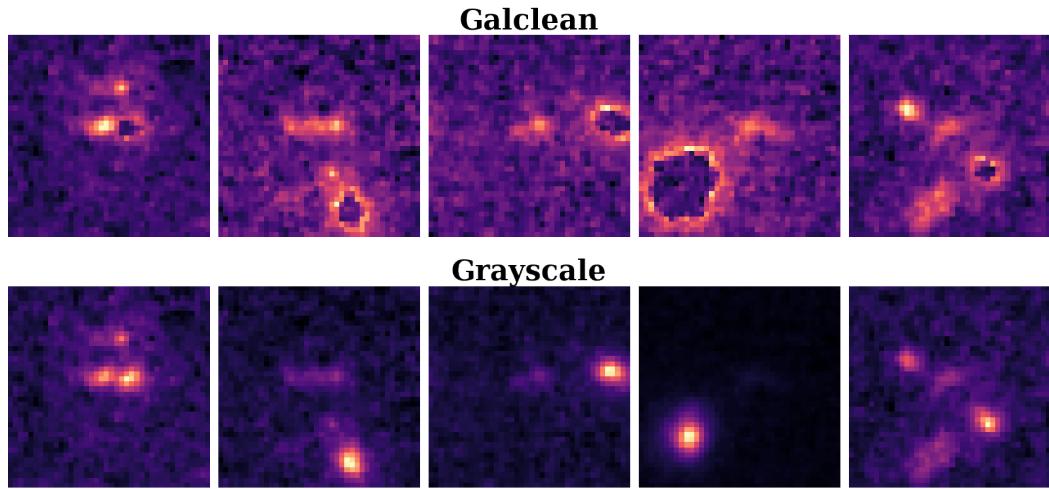
Selain dengan melakukan pemangkasan ukuran citra, opsi pembersihan data juga dapat dilakukan dengan menggunakan **Galclean** (Ferreira, 2018). Aplikasi ini dapat diaplikasikan terhadap data 1 *channel* atau *grayscale*, sehingga dataset RGB yang telah dibuat sebelumnya perlu dikonversi lebih dulu ke dalam format *grayscale*. Gambar III.17 menunjukkan perbandingan citra galaksi sebelum dan setelah melalui proses pembersihan menggunakan Galclean.



Gambar III.17: Perbandingan citra galaksi dalam format *grayscale* dengan citra galaksi setelah melalui proses pembersihan menggunakan **Galclean**

Namun, dalam beberapa galaksi ditemukan bahwa **Galclean** tidak membersihkan objek terang di dekat galaksi dengan halus, dan tampak 'merusak'

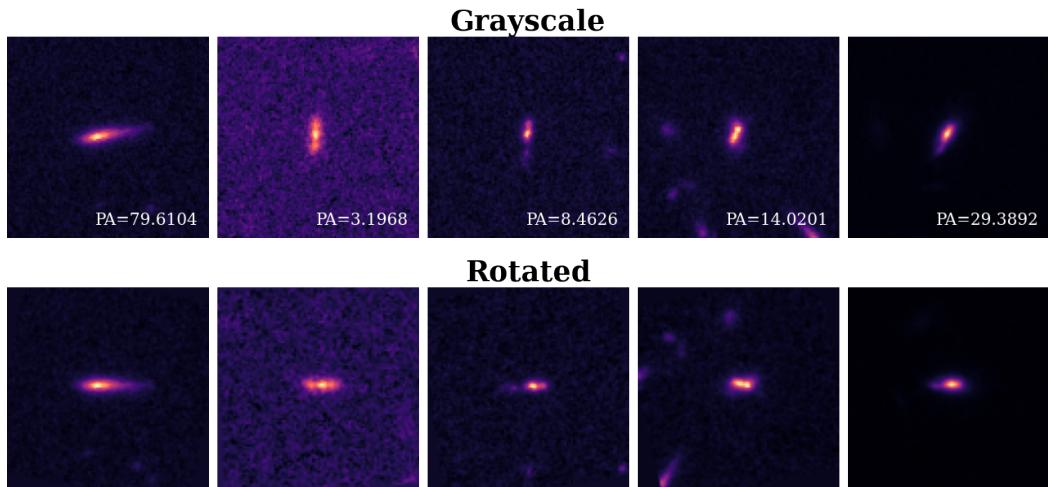
citra galaksi asli. Beberapa kasus tersebut ditunjukkan dalam Gambar III.18.



Gambar III.18: Sampel galaksi yang tidak halus dibersihkan oleh aplikasi **Galclean**. Gambar pada baris pertama menunjukkan data galaksi setelah melalui pembersihan dengan aplikasi **Galclean**, sementara gambar pada baris kedua menunjukkan data asli galaksi.

Rotating Data

Salah satu informasi yang diperoleh dari pemodelan parameterik galaksi menggunakan **Galfitm** adalah nilai *position angle* (PA) setiap galaksi. Parameter ini dapat dijadikan acuan untuk merotasi citra galaksi sehingga diperoleh keseragaman dalam arah PA galaksi. Hal ini dapat dilakukan untuk meminimalisir pengelompokan morfologi galaksi berdasarkan arah PA-nya. Namun perlu menjadi catatan bahwa nilai **Galfitm** tidak sepenuhnya akurat dalam menentukan nilai PA, karena sulit untuk menemukan nilai PA galaksi khususnya untuk galaksi dengan bentuk yang tak beraturan atau memiliki struktur *clump*. Selain itu, *fitting* parameterik yang dilakukan menghasilkan nilai PA yang berbeda untuk masing-masing filter. Dalam beberapa kasus, galaksi di salah satu filter tampak lebih redup daripada citranya di filter yang lain. Hal ini akan berdampak pada nilai PA galaksi. Oleh karena itu, proses *rotating* citra galaksi dilakukan berdasarkan nilai median dari *list* PA untuk ketiga filter. Gambar III.19 menunjukkan sampel galaksi sebelum dan setelah dilakukan proses rotasi berdasarkan nilai PA.



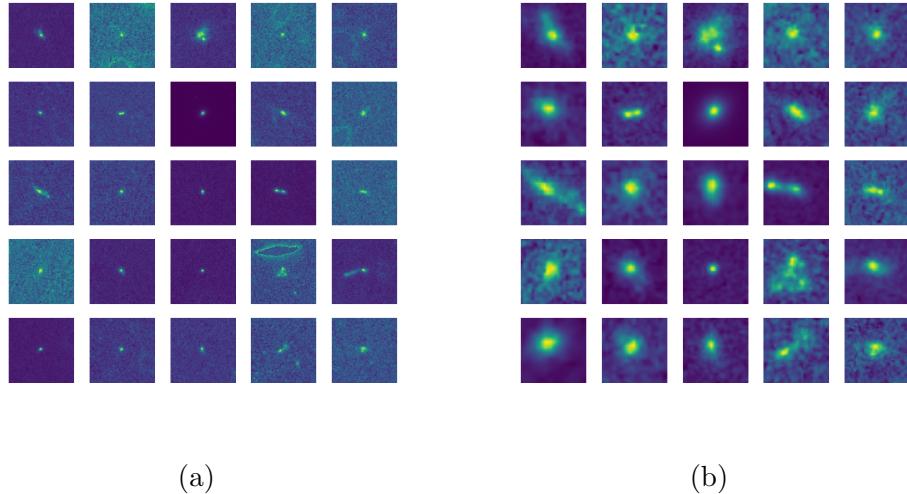
Gambar III.19: Sampel galaksi sebelum dan setelah melalui proses rotasi berdasarkan nilai *position angle*. Gambar pada baris pertama menunjukkan citra galaksi asli, sembari gambar pada baris kedua menunjukkan citra galaksi setelah melalui proses *rotating*.

Rescaling Data

Pada bagian pembersihan data sebelumnya disebutkan bahwa salah satu cara meminimalisir objek terang di dekat galaksi target adalah dengan memangkas ukuran citra semaksimal mungkin sehingga pada setiap citra hanya terdapat galaksi target yang akan dipelajari bentuknya. Namun pada kenyataannya, pemangkasan tersebut berpotensi membuat beberapa citra galaksi menjadi terpotong. Hal ini dapat terjadi terhadap galaksi-galaksi yang memiliki bentangan sudut yang besar. Oleh karena itu, salah satu cara yang dapat dilakukan untuk meminimalisir masalah ini adalah dengan melakukan *rescaling* data dengan memanfaatkan informasi radius efektif galaksi yang diperoleh dari proses *fitting Galfitm*.

Galaksi-galaksi dengan nilai radius efektif $R_e \leq 2pix$ akan diperbesar ukuran citranya dengan skala $\sim 2/R_e$. Sementara untuk galaksi dengan $2pix < R_e \leq 4pix$ akan diperbesar ukuran citranya dengan skala $\sim 4/R_e$. Untuk galaksi dengan $4pix < R_e \leq 6pix$ akan diperbesar ukuran citranya dengan skala $\sim 6/R_e$. Galaksi dengan $6pix < R_e \leq 7pix$ akan diperbesar ukuran citranya dengan skala $\sim 7/R_e$. Sementara itu, galaksi dengan ukuran $R_e > 7pix$ tidak melalui proses *rescaling* karena ukurannya relatif sudah cukup besar. Untuk setiap citra yang telah di-*resize* tersebut, selanjutnya akan dipangkas dengan ukuran yang seragam. Pengaturan diatas didasari oleh hasil *sampling trial and*

error yang dilakukan oleh penulis. Batasan yang diberikan diatas memberikan ukuran galaksi yang cukup seragam. Batasan tersebut diimplementasikan untuk galaksi dengan radius efektif yang dibatasi pada 10 piksel. Gambar III.20 menunjukkan perbandingan galaksi sebelum dan setelah dilakukan *rescaling* dan pemangkasan.



Gambar III.20: Perbandingan data sampel galaksi awal (panel a) dan galaksi yang sama setelah dilakukan *rescaling* dan dipangkas (panel b).

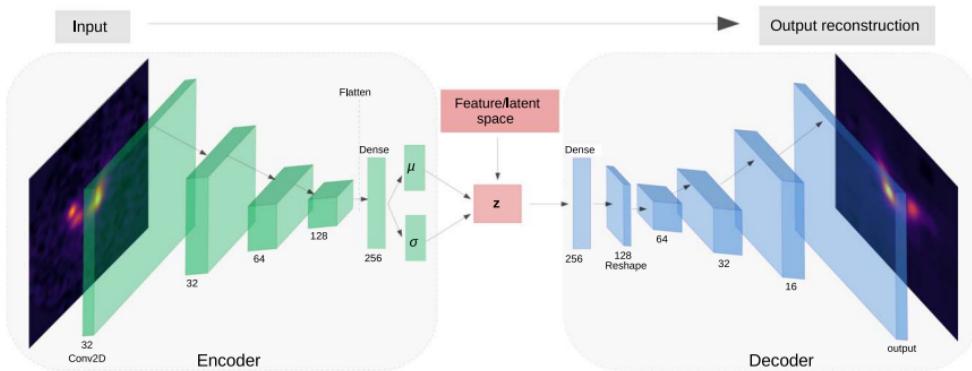
Normalisasi Data

Dataset yang dibuat sebelumnya memiliki nilai yang merentang dari nilai negatif yang besar hingga nilai positif yang juga besar, meski dominan berada pada rentang -1 sampai 1 . Sementara itu, pada proses VAE pada umumnya data harus memiliki nilai pada rentang $0 - 1$ (untuk tipe data *float*) atau $0 - 255$ (untuk tipe data *integer*). Oleh karena itu, perlu dilakukan normalisasi sehingga dataset memiliki nilai pada rentang yang sesuai dengan input arsitektur VAE. Salah satu cara yang dapat dilakukan adalah dengan membangun citra RGB menggunakan paket `make_lupton_rgb`, jika dataset yang akan digunakan dalam bentuk RGB, atau langsung melakukan normalisasi jika dataset yang digunakan dalam format *grayscale*.

III.3.4 Proses Pengkodean (*Encoding*)

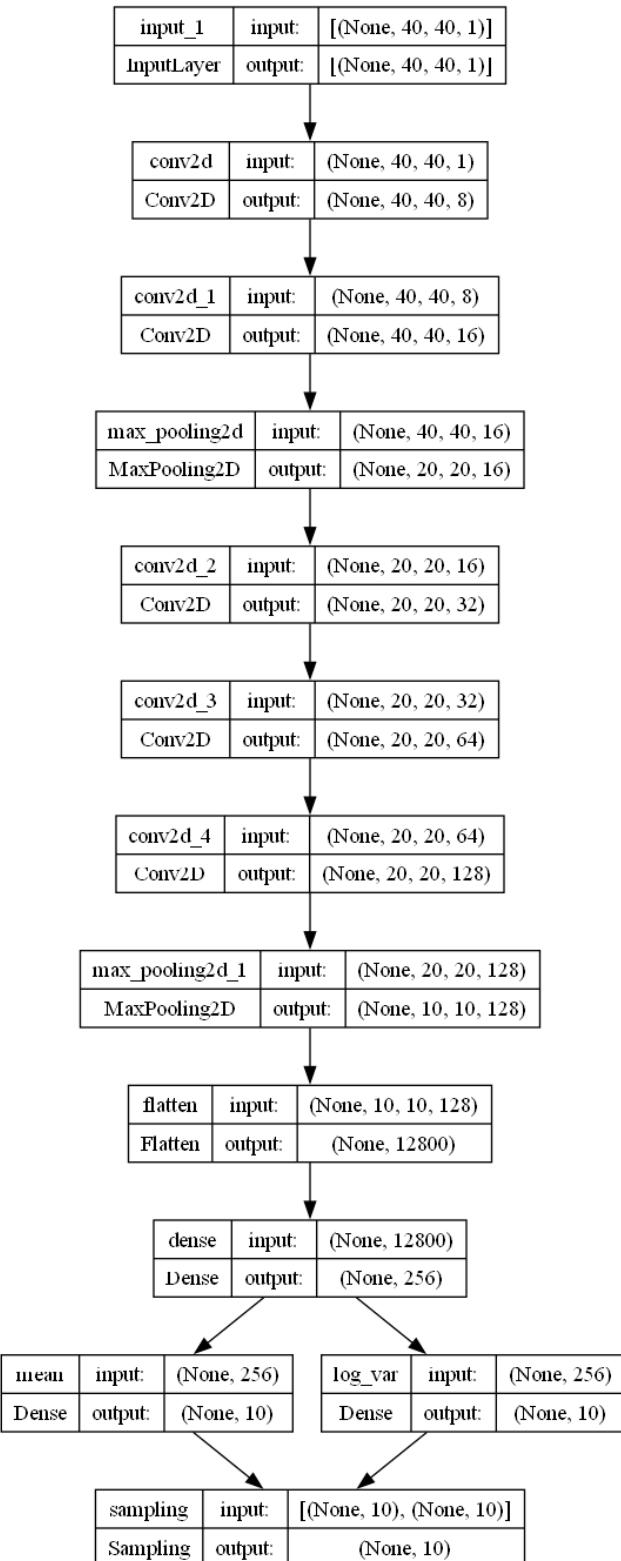
Sebagaimana telah sempat disebutkan pada Subbab II.2.2, pengelompokan galaksi akan dilakukan terhadap fitur-fitur data yang telah diekstrak setelah

melewati lapisan *encoder*. Proses *encoding* merupakan proses mengubah dimensi data yang besar menjadi lebih kecil, dengan tujuan menyaring fitur-fitur penting dari data. Sejumlah fitur penting tersebut selanjutnya berada dalam sebuah ruang laten (*latent space*) berdimensi n , bergantung pada seberapa banyak fitur yang akan dipertahankan dalam proses ekstraksi fitur. Selanjutnya, data akan direkonstruksi kembali berdasarkan fitur-fitur dalam ruang laten. Dengan demikian, ketika akan dilakukan klasifikasi morfologi dengan metode *unsupervised learning*, mesin hanya mempelajari fitur-fitur penting dari data. Secara lebih rinci, Gambar III.21 menunjukkan skema proses VAE pada penelitian Tohill dkk. (2024).

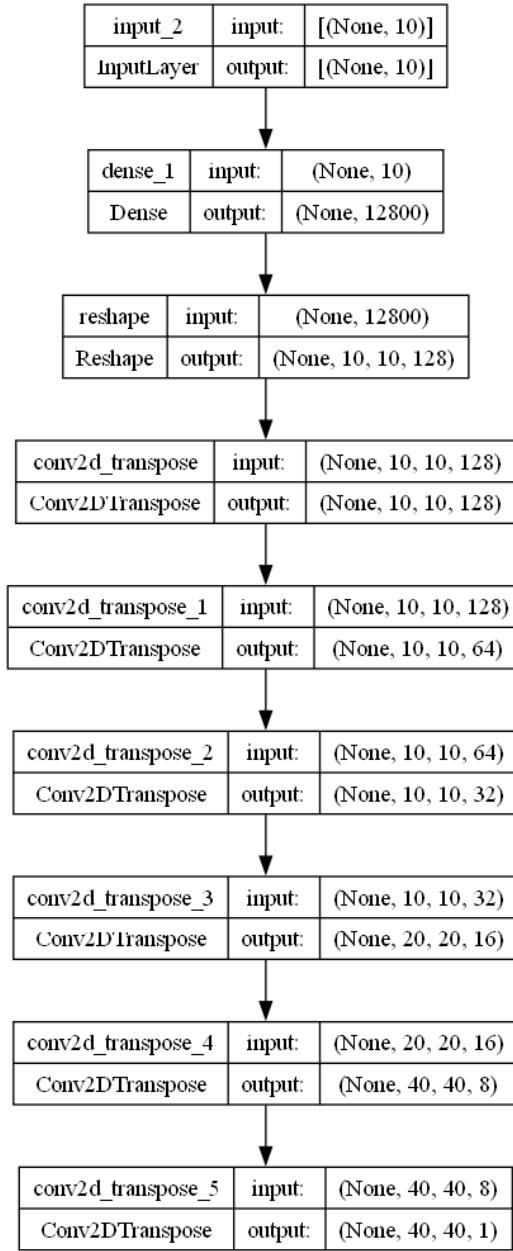


Gambar III.21: Skema proses *variational autoencoder*.
Sumber: Tohill dkk. (2024)

Pada penelitian ini, struktur *encoder* yang akan digunakan terhadap galaksi jauh ditunjukkan pada Gambar III.22 dan struktur *decoder* ditunjukkan pada Gambar III.23. Namun, kedua arsitektur ini dapat diubah untuk meminimalisir nilai *loss*.



Gambar III.22: Arsitektur *encoder* untuk galaksi jauh.



Gambar III.23: Arsitektur *decoder* untuk galaksi jauh.

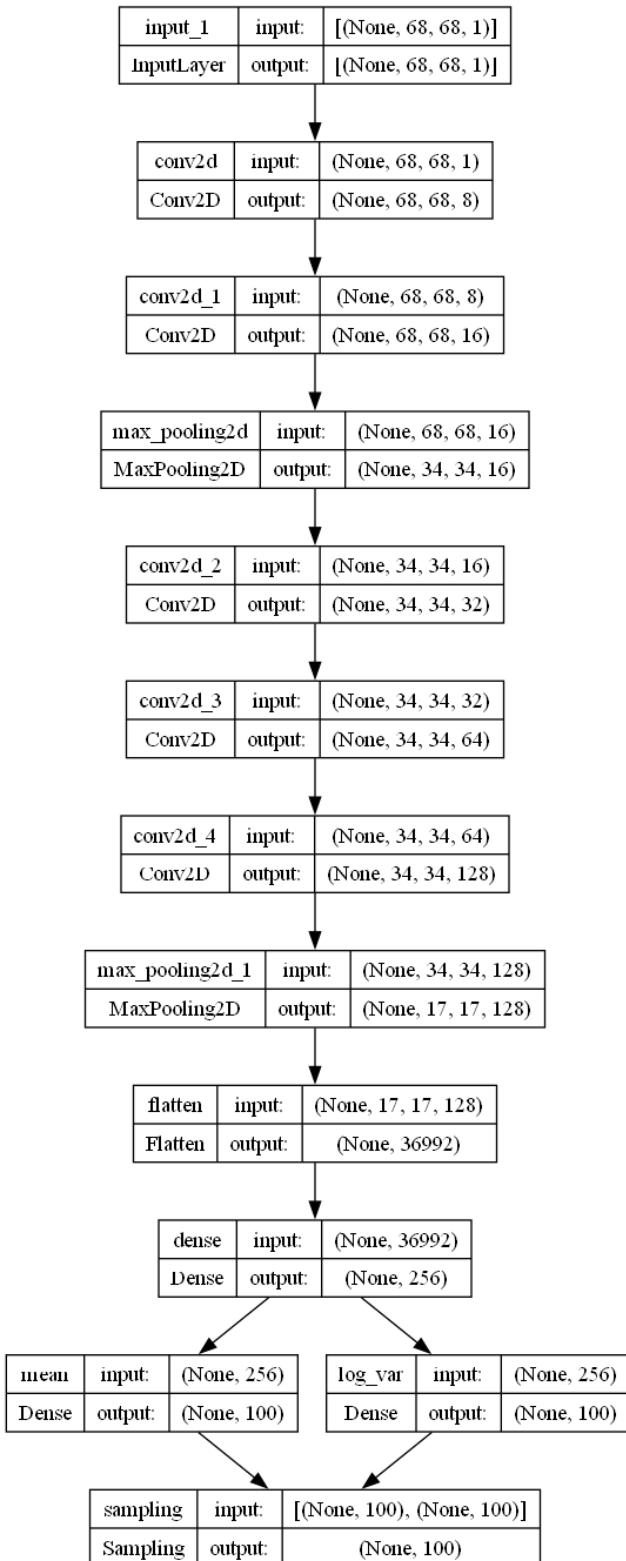
Dari arsitektur yang ditunjukkan pada Gambar III.22, *output* yang dipperoleh merupakan representasi data dalam 10 parameter laten. Seluruh parameter tersebut diharapkan merepresentasikan karakteristik morfologi galaksi yang menjadi objek kajian ini. Untuk bisa menelaah lebih jauh mengenai ke-10 parameter laten tersebut, akan digunakan metode reduksi dimensi sebagaimana dijelaskan dalam Subbab II.2.4 dengan menggunakan PCA dan UMAP. Tujuan dari reduksi dimensi ini adalah untuk mengamati pengelompokan yang mungkin dibentuk berdasarkan 10 parameter laten yang diekstrak dari setiap data. Membangun plot distribusi dalam 10 dimensi mustahil untuk dilaku-

kan, sehingga melalui metode PCA maupun UMAP, akan dilakukan transformasi linear (untuk metode PCA) dan transformasi non-linear (untuk metode UMAP) untuk membangun dua parameter baru yang merupakan kombinasi dari 10 parameter. Dari dua parameter baru tersebut, maka pengelompokan yang mungkin terbentuk lebih mudah diamati melalui plot dua dimensi.

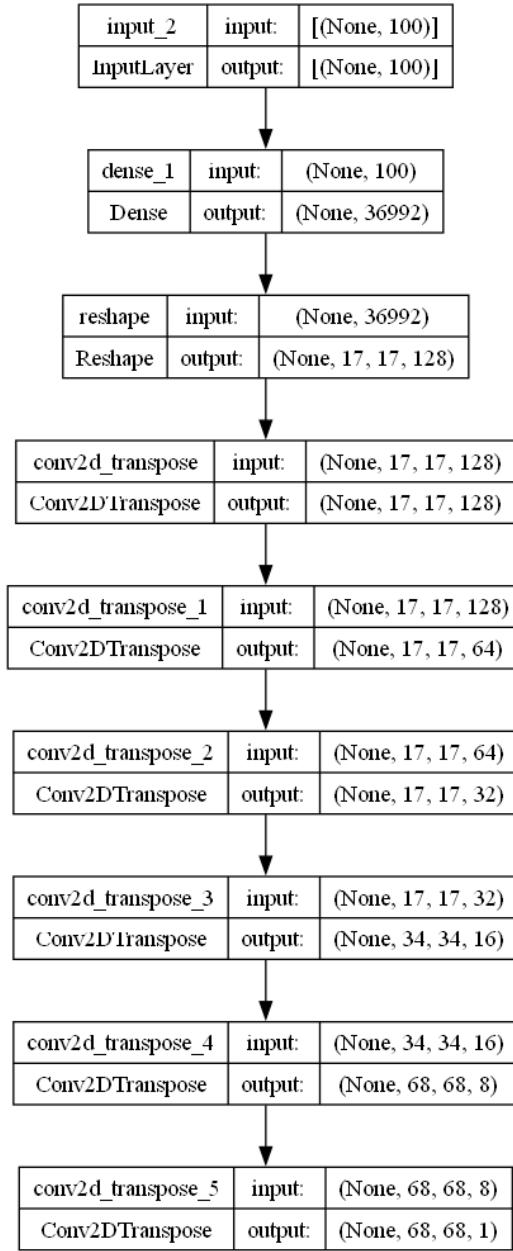
Sementara itu, galaksi dekat memiliki diameter sudut yang lebih besar dibandingkan galaksi jauh, sehingga data galaksi dekat memiliki ukuran data yang lebih besar dibandingkan ukuran data galaksi jauh. Maka, input data yang dimasukkan ke dalam arsitektur *encoder* perlu diubah menyesuaikan ukuran data galaksi dekat. Dengan demikian, arsitektur *encoder* dan *decoder* yang akan digunakan untuk galaksi dekat tidak dapat langsung menggunakan arsitektur seperti pada Gambar III.22 dan III.23.

Selain berdasarkan ukuran input data yang berbeda, galaksi dekat memiliki struktur yang lebih detil dibandingkan galaksi jauh, seperti keberadaan lengank-lengan spiral di dalam galaksi. Struktur yang detil seperti ini seringkali tidak berhasil direpresentasikan hanya dalam 10 parameter laten. Oleh karena itu, *output* dari *encoder* akan merepresentasikan galaksi dekat dalam lebih dari 10 parameter laten.

Gambar III.24 dan III.25 menunjukkan arsitektur *encoder* dan *decoder* yang digunakan untuk merekonstruksi data galaksi dekat. Selain dari perubahan terkait ukuran input data dan jumlah parameter laten, lapisan-lapisan dan urutan dari arsitektur *autoencoder* galaksi dekat tetap sama dengan arsitektur *autoencoder* untuk galaksi jauh.



Gambar III.24: Arsitektur *encoder* untuk galaksi dekat



Gambar III.25: Arsitektur *decoder* untuk galaksi dekat

III.3.5 Pengelompokan Morfologi Galaksi Menggunakan Metode *Unsupervised Learning*

Pada penelitian ini, penulis akan melakukan pengelompokan morfologi galaksi dengan beberapa metode *unsupervised learning*. Pemilihan *unsupervised learning* dibandingkan *supervised* maupun *reinforcement learning* adalah karena belum adanya teori yang *well-established* dalam menjelaskan morfologi galaksi *redshift* tinggi. Beberapa metode *unsupervised learning* yang dapat digunakan

an untuk melakukan pengelompokan diantaranya adalah *hierachial clustering* dan *K-means clustering*, sebagaimana yang sebelumnya telah dijelaskan dalam Subbab II.2.3. Pada penelitian ini, pengelompokan galaksi menggunakan metode *hierarchical clustering* dilakukan dengan menggunakan definisi jarak *euclidean*, dan jarak antarklaster dihitung dengan *ward linkage*.

Similarity

Salah satu cara untuk menilai hasil pengelompokan morfologi adalah dengan melihat nilai kemiripan antara galaksi-galaksi di dalam klaster, maupun kemiripan galaksi-galaksi antarklaster. Nilai kemiripan galaksi di dalam klaster menunjukkan seberapa seragam galaksi-galaksi yang dikelompokan oleh mesin ke dalam sebuah klaster. Sementara itu, kemiripan antarklaster menunjukkan seberapa besar perbedaan antara satu klaster dengan klaster yang lain.

Terdapat beberapa metode perhitungan *similarity* yang dapat dilakukan, dan dua diantaranya adalah dengan menghitung nilai *cosine similarity* dan *structural similarity index*. *Cosine similarity* adalah metode perhitungan kemiripan dua tensor dari nilai kosinus sudut yang dibentuk kedua tensor. Tensor sendiri didefinisikan sebagai *array* multidimensi yang merupakan bentuk umum dari skalar, vektor, dan matriks dalam dimensi yang lebih tinggi. *Cosine similarity* dihitung dalam persamaan berikut.

$$\cos \theta = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \quad (\text{III.1})$$

dengan \mathbf{A} dan \mathbf{B} sebagai dua tensor yang akan dicari nilai kemiripannya. Pada penelitian ini, tensor \mathbf{A} dan \mathbf{B} merupakan vektor berisi parameter laten dalam n dimensi.

Sementara itu, *structural similarity index* (SSIM) merupakan metode untuk menemukan nilai *similarity* dengan membandingkan dua gambar secara langsung. Nilai SSIM memperhitungkan tiga aspek, yakni simetri, kecerahan, dan kontras antara kedua gambar. Dengan kata lain, apabila *cosine similarity* membandingkan dua citra galaksi yang direpresentasikan dalam sebuah tensor, SSIM membandingkan citra galaksi secara langsung. SSIM banyak digunakan untuk membandingkan kualitas dua gambar atau video yang berbeda. Salah satu pemanfaatan SSIM adalah dalam proses kompresi ukuran gambar atau video. Untuk menguji representasi laten dari data, *cosine similarity* menjadi metode perhitungan *similarity* yang cocok untuk penelitian ini.

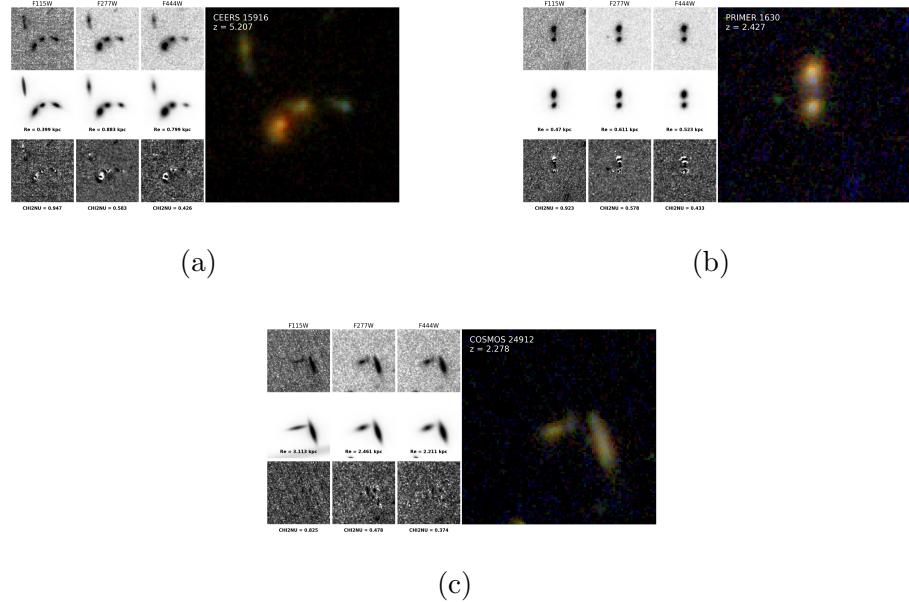
BAB IV

HASIL DAN ANALISIS

Pada Subbab ini akan dijelaskan beberapa hasil dari penelitian ini, mulai dari hasil pemodelan parametrik galaksi menggunakan *multi Sérsic fitting*, dilanjutkan dengan analisis distribusi ukuran galaksi terhadap beberapa parameter, seperti massa, *redshift*, serta panjang gelombang. Bagian selanjutnya adalah membahas hasil pengelompokan galaksi di *redshift* dekat, dan dilanjutkan dengan hasil pengelompokan galaksi di *redshift* tinggi.

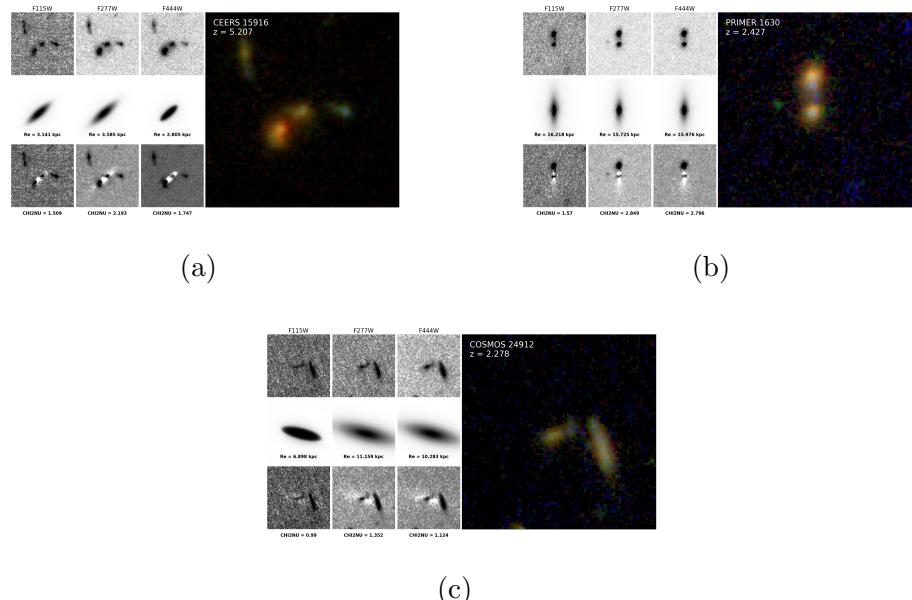
IV.1 Hasil Pemodelan Parametrik

Sebagaimana dijelaskan pada Subbab III.3.2, dalam penelitian ini dilakukan *fitting* kecerlangan galaksi menggunakan fungsi Sérsic. Namun untuk menangani data yang terkontaminasi oleh galaksi tetangga dalam satu citra, maka akan digunakan *fitting* menggunakan *multi Sérsic*. Perlu ditekankan bahwa *fitting multi Sérsic* yang dimaksud dalam penelitian ini bukan bertujuan untuk melakukan *fitting* terhadap komponen-komponen galaksi (misalnya *bulge* dan piringan) secara terpisah.



Gambar IV.1: Contoh beberapa galaksi hasil *fitting multi Sérsic*: CEERS15916 (panel a), PRIMER1630 (panel b), COSMOS24912 (panel c). Baris pertama menunjukkan data pada ketiga filter, baris kedua menunjukkan *best-fit* model, baris ketiga menunjukkan residual *fitting*, serta gambar sebelah kanan menunjukkan citra RGB yang dibangun dari ketiga filter.

Gambar IV.1 menunjukkan beberapa galaksi hasil *fitting multi Sérsic*. Contoh hasil *fitting* tersebut menunjukkan bahwa *fitting multi Sérsic* memberikan hasil *fitting* yang lebih baik dibandingkan *fitting single Sérsic*, terutama dalam mengatasi citra galaksi yang terkontaminasi. Sebagai perbandingan, Gambar IV.2 menunjukkan sampel galaksi yang sama ketika di-*fitting* hanya dengan menggunakan *single Sérsic*.



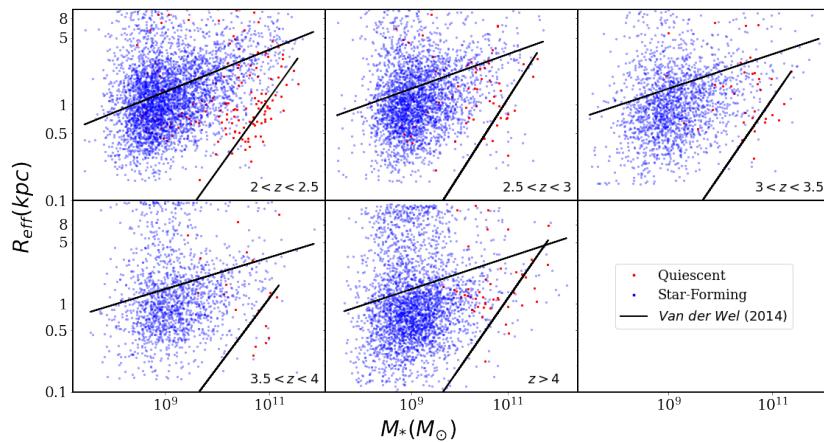
Gambar IV.2: Contoh beberapa galaksi hasil *fitting single Sérsic*: CERS15916 (panel a), PRIMER1630 (panel b), COSMOS24912 (panel c). Baris pertama menunjukkan data pada ketiga filter, baris kedua menunjukkan model *best-fit* model, baris ketiga menunjukkan residual *fitting*, serta gambar sebelah kanan menunjukkan citra RGB yang dibangun dari ketiga filter.

IV.1.1 Hubungan Massa Terhadap Radius Efektif

Salah satu parameter *output* dari *fitting* kecerlangan galaksi adalah nilai radius efektif galaksi. Gambar IV.3 menunjukkan plot hubungan radius efektif galaksi terhadap massa dari katalog *EAZY*. Gambar tersebut menunjukkan hubungan massa-radius untuk galaksi *quiescent* dan galaksi *star forming*. Pada plot tersebut ditunjukkan juga hubungan massa-radius yang diperoleh pada penelitian van der Wel dkk. (2014).

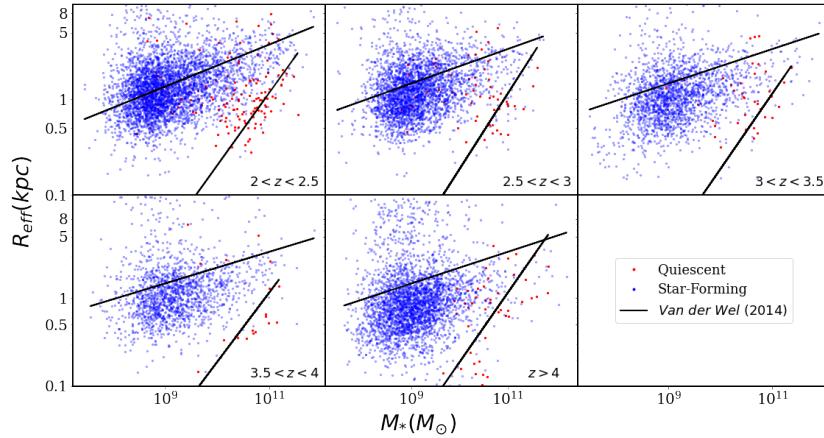
Data yang ditampilkan pada Gambar IV.3 merupakan data yang telah dilakukan seleksi objek non-galaksi dan galaksi dengan hasil *fitting Galfitm* yang tidak memodelkan galaksi target karena objek tidak terdeteksi saat dilakukan ekstraksi sumber. Total galaksi yang ditunjukkan dalam Gambar IV.3 adalah sebanyak 14456 galaksi. Data dikelompokkan ke dalam lima *bin redshift*, dengan *redshift* yang dipakai adalah z_{spec} apabila tersedia, dan z_{phot} apabila tidak ada informasi nilai z_{spec} .

F115W



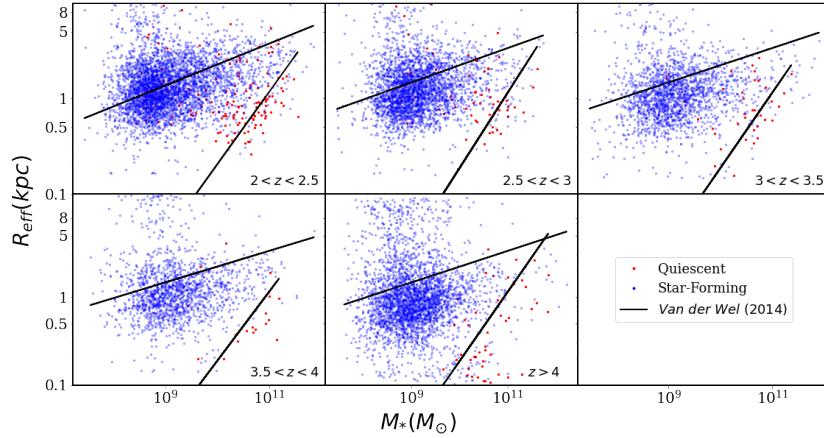
(a)

F277W



(b)

F444W



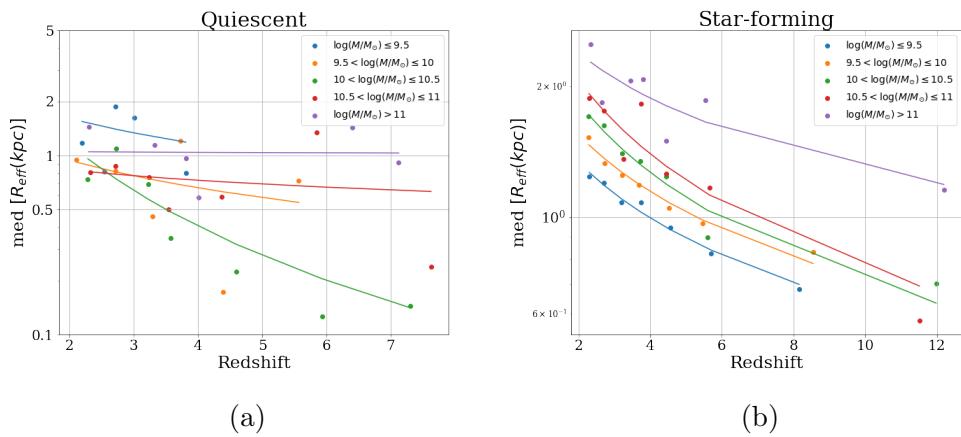
(c)

Gambar IV.3: Plot hubungan radius efektif galaksi terhadap massa, dari hasil *fitting* pada filter F115W (panel a), filter F277W (panel b), dan filter F444W (panel c).

Sebaran massa-radius dari data penelitian ini menunjukkan kecocokan terhadap penelitian terdahulu, dimana galaksi *quiescent* memiliki hubungan massa-radius yang lebih tajam dibandingkan galaksi *star forming*. Hal ini dapat disebabkan karena galaksi *quiescent* atau galaksi *early-type* memiliki bentuk elips yang dominan, sehingga penambahan massa terdistribusi secara mera- ta. Sementara itu, pada galaksi *star forming* atau galaksi *late-type*, massa yang terkonsentrasi di pusat galaksi membuat penambahan massa tidak terlalu mendominasi peningkatan radius efektif galaksi.

IV.1.2 Hubungan Radius Efektif Terhadap *Redshift*

Gambar IV.4 menunjukkan plot hubungan radius efektif galaksi terhadap *redshift* untuk galaksi *quiescent* dan galaksi *star forming*. Sama seperti data pada Gambar IV.3, total data yang digunakan untuk membangun plot pada Gambar IV.4 galaksi adalah sejumlah 14456 galaksi. Untuk membangun plot pada gambar tersebut, data dikelompokkan ke dalam tujuh bin *redshift* ($2 < z \leq 2.5$, $2.5 < z \leq 3$, $3 < z \leq 3.5$, $3.5 < z \leq 4$, $4 < z \leq 5$, $5 < z \leq 7$, dan $z > 7$). Nilai *redshift* yang dipakai adalah z_{spec} apabila tersedia, dan z_{phot} apabila informasi z_{spec} tidak tersedia. Selain dikelompokkan ke dalam tujuh bin *redshift*, data juga dikelompokkan ke dalam lima massa, sebagaimana ditunjukkan dalam Gambar IV.4.



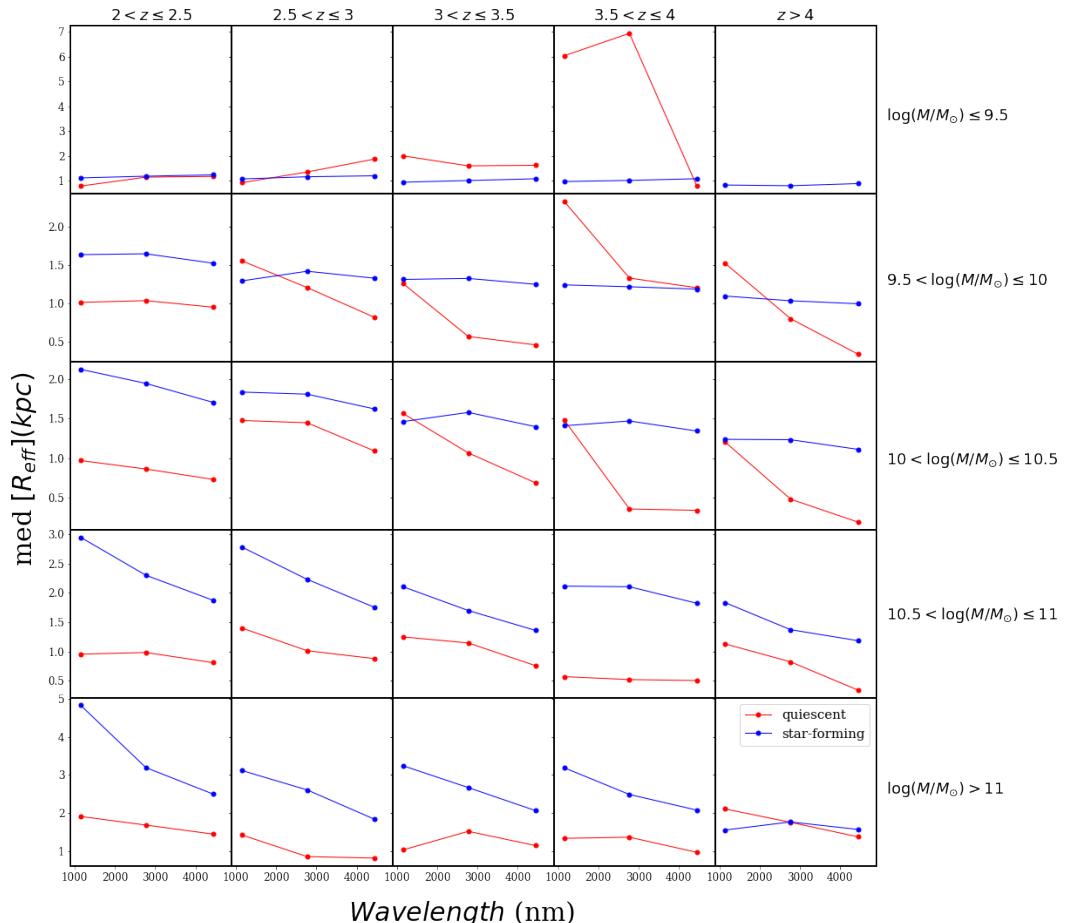
Gambar IV.4: Plot hubungan radius efektif galaksi terhadap *redshift* untuk galaksi *quiescent* (panel a) dan galaksi *star forming* (panel b).

Secara umum, hubungan yang jelas tampak pada galaksi *star forming*. Semakin bertambahnya *redshift*, galaksi-galaksi *star forming* tampak memiliki ukuran yang semakin kecil. Sementara itu, pada galaksi *quiescent*, data terli-

hat lebih menyebar, namun tren yang sama terlihat jelas untuk galaksi *quiescent* dengan $10 < \log(M/M_\odot) \leq 10.5$.

IV.1.3 Hubungan Radius Efektif Terhadap Panjang Gelombang

Hubungan radius efektif galaksi terhadap panjang gelombang ditunjukkan pada Gambar IV.5. Dengan membagi data ke dalam lima rentang *redshift* dan lima rentang massa, plot tersebut menunjukkan nilai median dari distribusi radius efektif galaksi dari *fitting* tiga panjang gelombang. Secara umum, tampak tren radius efektif galaksi semakin kecil seiring dengan penambahan panjang gelombang. Terlihat pula bahwa ukuran galaksi *quiescent* pada umumnya lebih kecil dibandingkan galaksi *star forming*. Sampel galaksi *quiescent* pada massa rendah cukup sedikit, sehingga tren radius efektif galaksi terhadap panjang gelombang untuk galaksi massa rendah perlu diwaspadai.



Gambar IV.5: Plot hubungan radius efektif galaksi terhadap panjang gelombang.

Radius efektif yang lebih kecil pada panjang gelombang panjang disebabkan karena dua hal, yaitu gradien populasi bintang dan *dust attenuation* (Baes dkk., 2024). Bintang-bintang tua dengan warna yang lebih merah biasanya berada di dekat pusat galaksi, sehingga radius efektif galaksi yang terukur lebih kecil. Sementara itu, bintang-bintang muda dengan warna yang lebih biru tersebar di bidang galaksi, sehingga radius efektif pada panjang gelombang yang lebih pendek tampak lebih besar. Di sisi lain, *dust attenuation* juga berkontribusi sebesar 20% terhadap hubungan massa terhadap panjang gelombang (Baes dkk., 2024). Debu di dalam galaksi menyerap foton, sehingga membuat profil kecerlangan permukaan galaksi menjadi lebih datar. Oleh karenanya, radius efektif galaksi akan menjadi lebih besar karena radius efektif galaksi definisikan sebagai radius yang melingkupi setengah dari kecerlangan galaksi. Efek *dust attenuation* lebih tinggi pada panjang gelombang pendek, sehingga radius efektif galaksi pada panjang gelombang pendek menjadi sedikit lebih tinggi.

IV.1.4 Rapat Jumlah Galaksi Berdasarkan Radius Efektif

Analisis lainnya yang dibuat berdasarkan informasi radius efektif galaksi dari hasil *fitting* parametrik adalah terkait rapat jumlah (*number density*) galaksi pada berbagai ukuran. Rapat jumlah memberikan informasi banyaknya galaksi pada berbagai ukuran, di berbagai rentang *redshift*. Gambar IV.6 menunjukkan distribusi ukuran galaksi pada berbagai *redshift*. Sama seperti sebelumnya, total galaksi untuk membangun plot pada gambar tersebut sejumlah 14456 galaksi. Plot ini dibangun dengan membagi data radius efektif galaksi dalam beberapa *bin*, dengan rentang *bin* mengikuti aturan Scott (Scott, 2010). Data juga dibagi ke dalam tujuh *bin redshift*, dengan menggunakan nilai z_{spec} apabila tersedia, tetapi apabila tidak tersedia digunakan nilai z_{phot} .

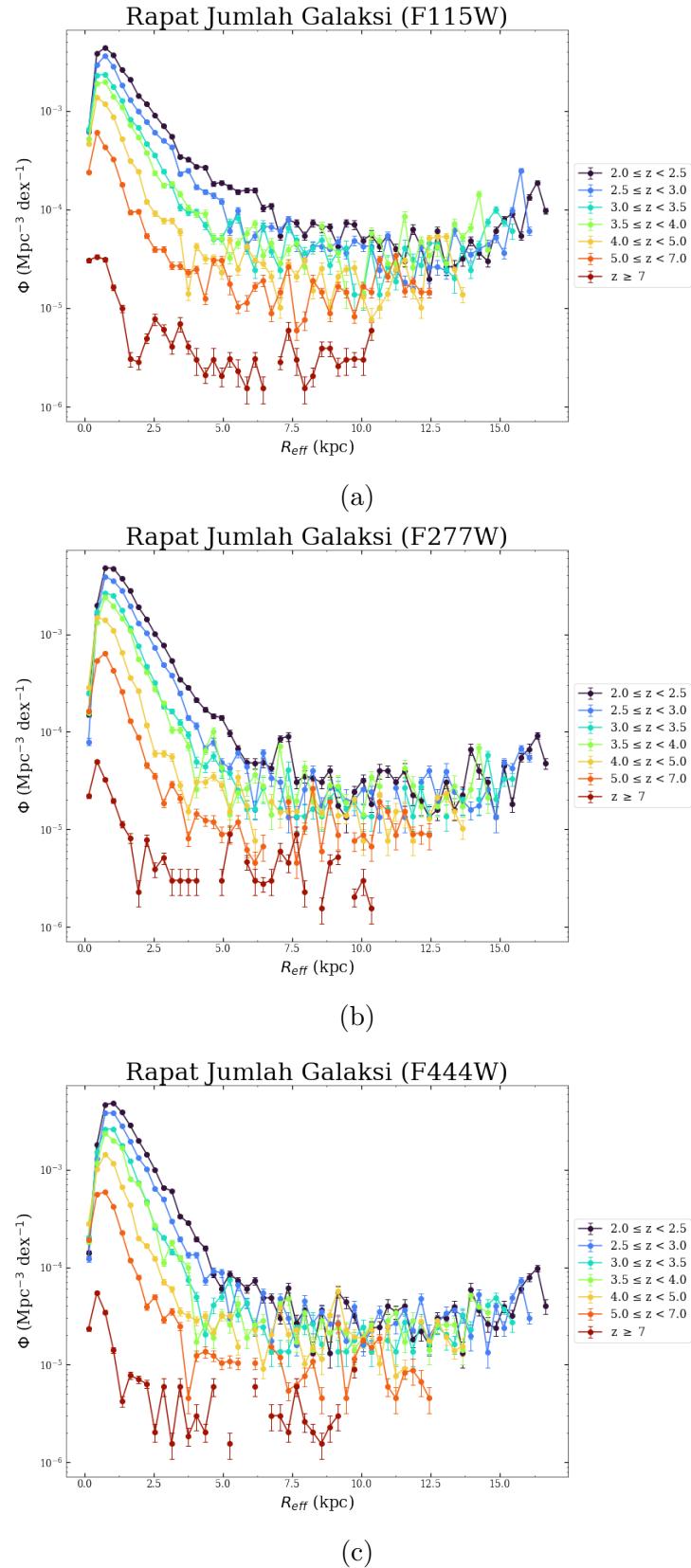
Rapat jumlah dihitung dengan menggunakan persamaan IV.1.

$$\Phi = \frac{N}{V \times \Delta_{Reff}} \quad (\text{IV.1})$$

dengan Φ , N , V , dan Δ_{Reff} masing-masing sebagai rapat jumlah, jumlah galaksi pada setiap *bin*, volume untuk setiap *bin redshift*, dan lebar *bin* radius efektif. Volume setiap *bin redshift* dihitung dengan persamaan IV.2 Galat yang ditampilkan dihitung dari galat jumlah galaksi dibagi dengan volume.

$$V = \frac{4\pi}{3} \frac{\Omega}{\Omega^{langit}} (D_{cov}^3(z_{up}) - D_{cov}^3(z_{low})) \quad (\text{IV.2})$$

Pada persamaan IV.2, Ω didefinisikan sebagai luas medan pandang survei, Ω^{langit} sebagai luas sudut langit yang bernilai 41253 deg^2 , dan D_{cov} yaitu jarak *comoving*, pada batas atas dan batas bawah *redshift* pada setiap *bin*.



Gambar IV.6: Plot hubungan rapat jumlah galaksi pada berbagai radius efektif, dari hasil *fitting* pada filter F115W (panel a), filter F277W (panel b), dan filter F444W (panel c).

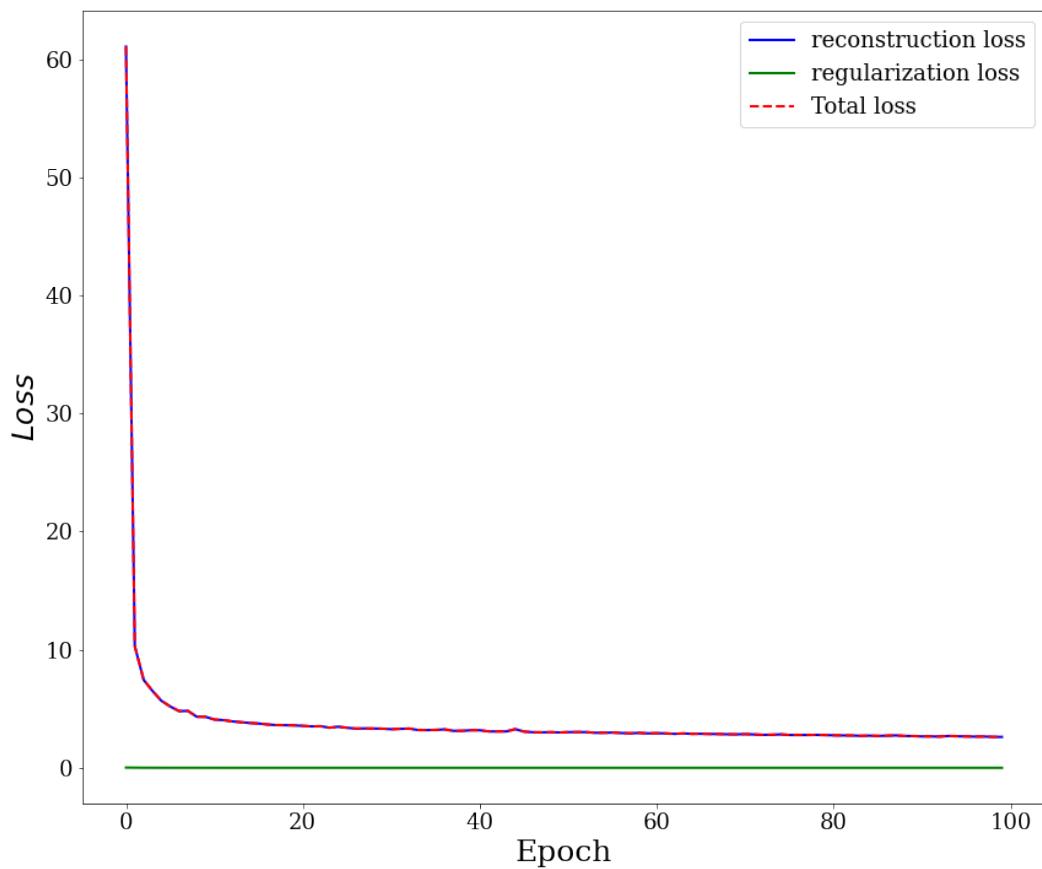
Berdasarkan Gambar IV.6, terlihat bahwa pada sebagian besar *bin redshift*, puncak distribusi berada di radius efektif sekitar $0.5 - 1$ kpc, lalu menurun secara gradual.

IV.2 Pengelompokan Galaksi Dekat

Agar dapat memastikan bahwa arsitektur VAE dan metode *clustering* yang digunakan untuk mengelompokan galaksi jauh sudah tepat, dengan metode yang sama akan dilakukan pengelompokan galaksi dekat yang telah diketahui kategori morfologinya. Terkait hal ini dapat dilihat kembali pada Subbab III.1.4. Karena tidak terdapat informasi radius galaksi maupun *position angle* di dalam katalog galaksi yang digunakan, maka data fotometri galaksi hanya akan melalui proses pembersihan dengan aplikasi `Galclean` dan pemangkasan ukuran data, tanpa proses *rotating* dan *rescaling*. Dataset galaksi dekat memiliki ukuran 68×68 piksel, dengan total data sejumlah 14059 galaksi, dalam format *grayscale*.

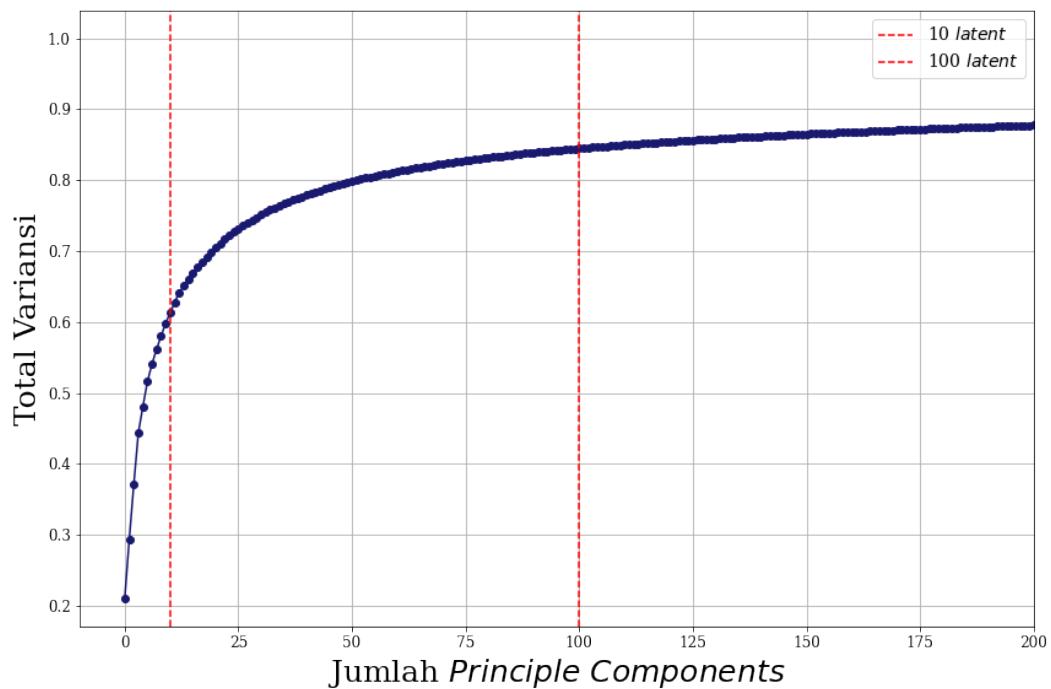
IV.2.1 Hasil Autoencoding

Berdasarkan arsitektur yang ditunjukkan pada Gambar III.24 dan III.25, nilai *loss* setelah 100 *epoch* berada di angka ~ 2.6 . Gambar IV.7 menunjukkan bahwa nilai *reconstruction loss* menurun signifikan dalam beberapa *epoch* pertama, lalu relatif konstan hingga *epoch* terakhir. Hal ini dapat diartikan bahwa data hasil rekonstruksi semakin mendekati dengan data input seiring dengan bertambahnya *epoch*. Sementara itu, nilai *regularization loss* sejak awal cukup rendah dan tidak menunjukkan penurunan yang signifikan. Hal ini menunjukkan bahwa parameter laten sejak awal didorong untuk mengikuti distribusi normal.



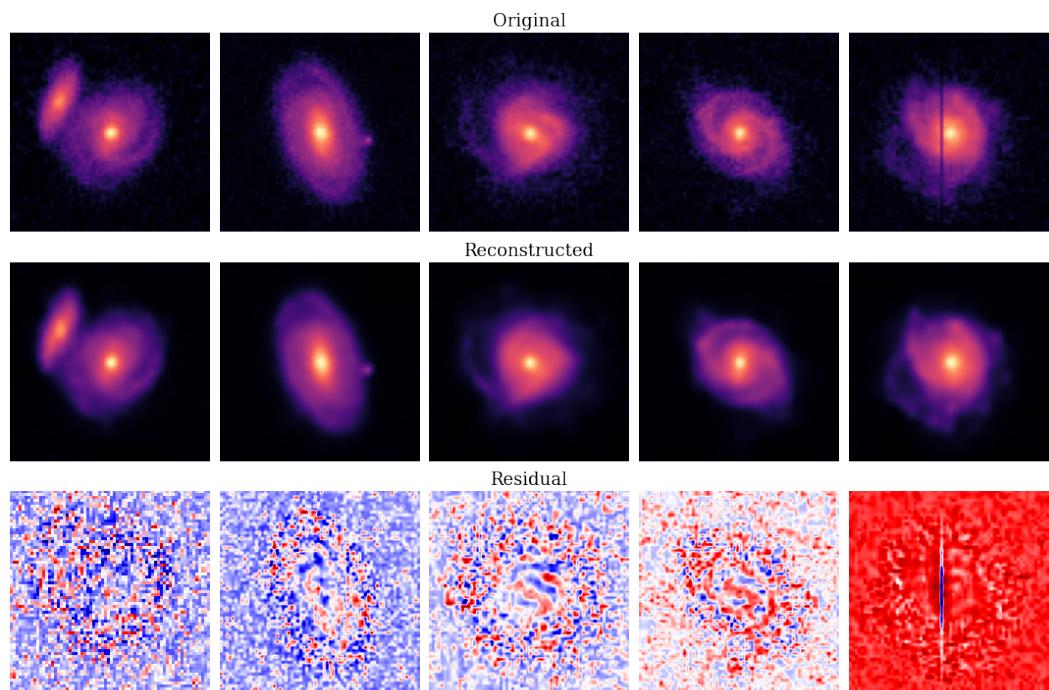
Gambar IV.7: Plot pergerakan nilai *loss* pada setiap *epoch* selama proses *autoencoding* untuk galaksi dekat

Untuk menentukan jumlah parameter laten yang tepat untuk dapat merekonstruksi data tanpa menghilangkan fitur-fitur penting didalamnya, dapat dilakukan dengan memanfaatkan metode PCA. PCA akan membangun sejumlah *principle components* yang merupakan kombinasi linear dari data. Jumlah *principle components* yang bisa memberikan nilai variansi data secara signifikan menjadi indikasi jumlah dimensi laten yang dapat dipilih. Sebagaimana tujuan dari PCA itu sendiri untuk membangun sumbu *principle component* yang bisa menangkap variansi terbesar di dalam data. Gambar IV.8 menunjukkan total variansi yang dapat direpresentasikan oleh sejumlah *principle components*.

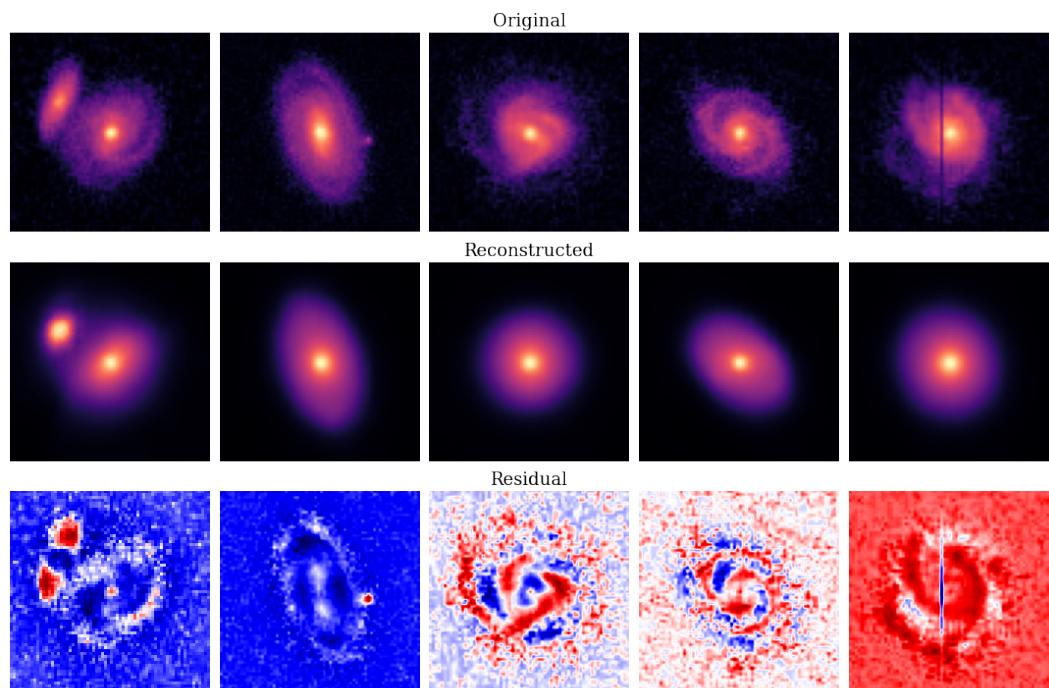


Gambar IV.8: Total variansi data galaksi dekat untuk berbagai jumlah *principle components*.

Apabila dibandingkan secara visual, data yang direkonstruksi dari 100 parameter laten telah merepresentasikan data input dengan cukup baik. Jika dilihat dari Gambar IV.8, 100 *principle components* akan merepresentasikan lebih dari 80% variansi pada data. Gambar IV.9 menunjukkan perbandingan data input dengan data yang direkonstruksi dari 100 parameter laten, serta residual yang dihitung dari data input dikurangi dengan data rekonstruksi. Visualisasi dari 100 parameter laten yang merepresentasikan data dapat dilihat pada bagian lampiran. Sebagai perbandingan, Gambar IV.10 menunjukkan hasil rekonstruksi data apabila data input hanya direpresentasikan hanya dalam 10 parameter laten. Terlihat fitur-fitur lengan spiral tidak tampak, sehingga galaksi spiral tampak seperti galaksi elips.



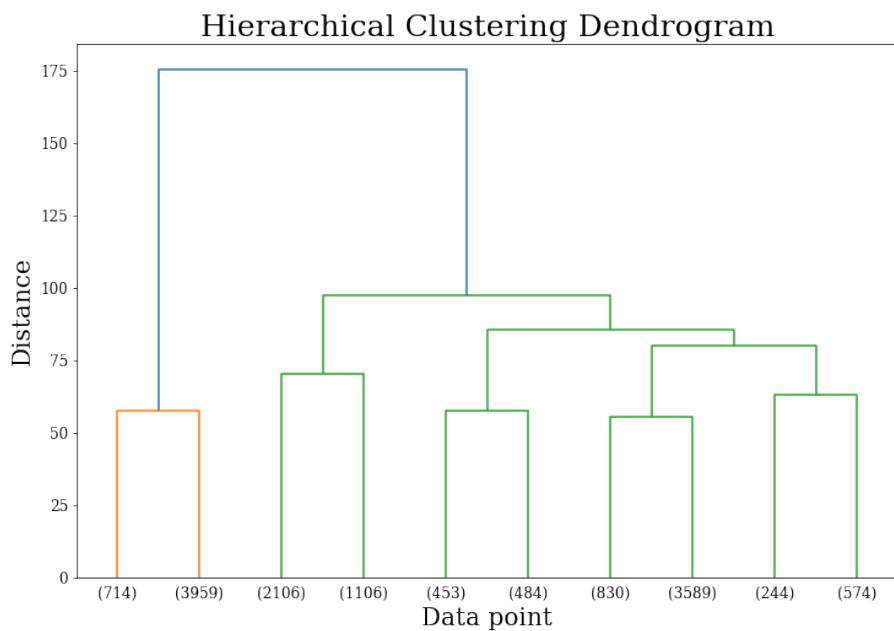
Gambar IV.9: Beberapa sampel *reconstructed images* galaksi dekat yang dibangun dari 100 parameter laten dibandingkan terhadap data input dan residualnya.



Gambar IV.10: Plot perbandingan *reconstructed images* terhadap data input dan residualnya, untuk galaksi dekat dengan 10 parameterlatenten.

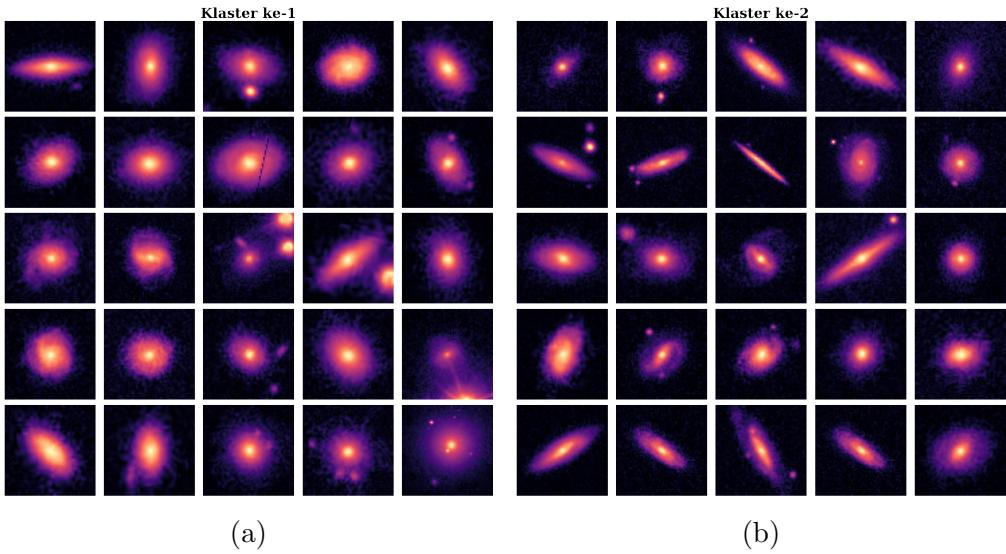
IV.2.2 *Hierarchical Clustering*

Pengelompokan galaksi dekat dengan metode *hierarchical clustering* memberikan dendrogram seperti ditunjukkan pada Gambar IV.11. Pada gambar tersebut, dendrogram dipotong hingga terdapat 10 klaster, meski tidak ada penentuan jumlah klaster yang sebenarnya berdasarkan pemotongan ini. Penentuan jumlah klaster dapat dilihat dari jarak antarklaster.



Gambar IV.11: Dendrogram hasil *hierarchical clustering* untuk galaksi dekat.

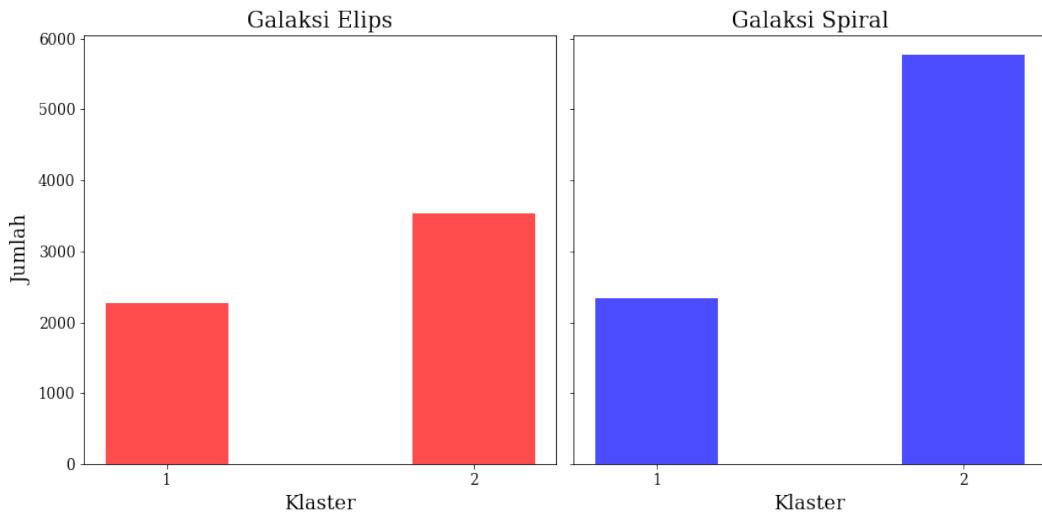
Dari Gambar IV.11, terlihat bahwa galaksi dekat dapat dikelompokan ke dalam dua klaster, yakni klaster berwarna jingga dan klaster berwarna hijau. Jarak antara kedua klaster tersebut berada di angka 175, dan klaster lainnya baru terbentuk pada jarak yang lebih kecil, yaitu di sekitar 90.



Gambar IV.12: Sampel 25 galaksi dekat dari pengelompokan dengan metode *hierarchical clustering*. Masing-masing menunjukkan klaster pertama (panel a), dan klaster kedua (panel b).

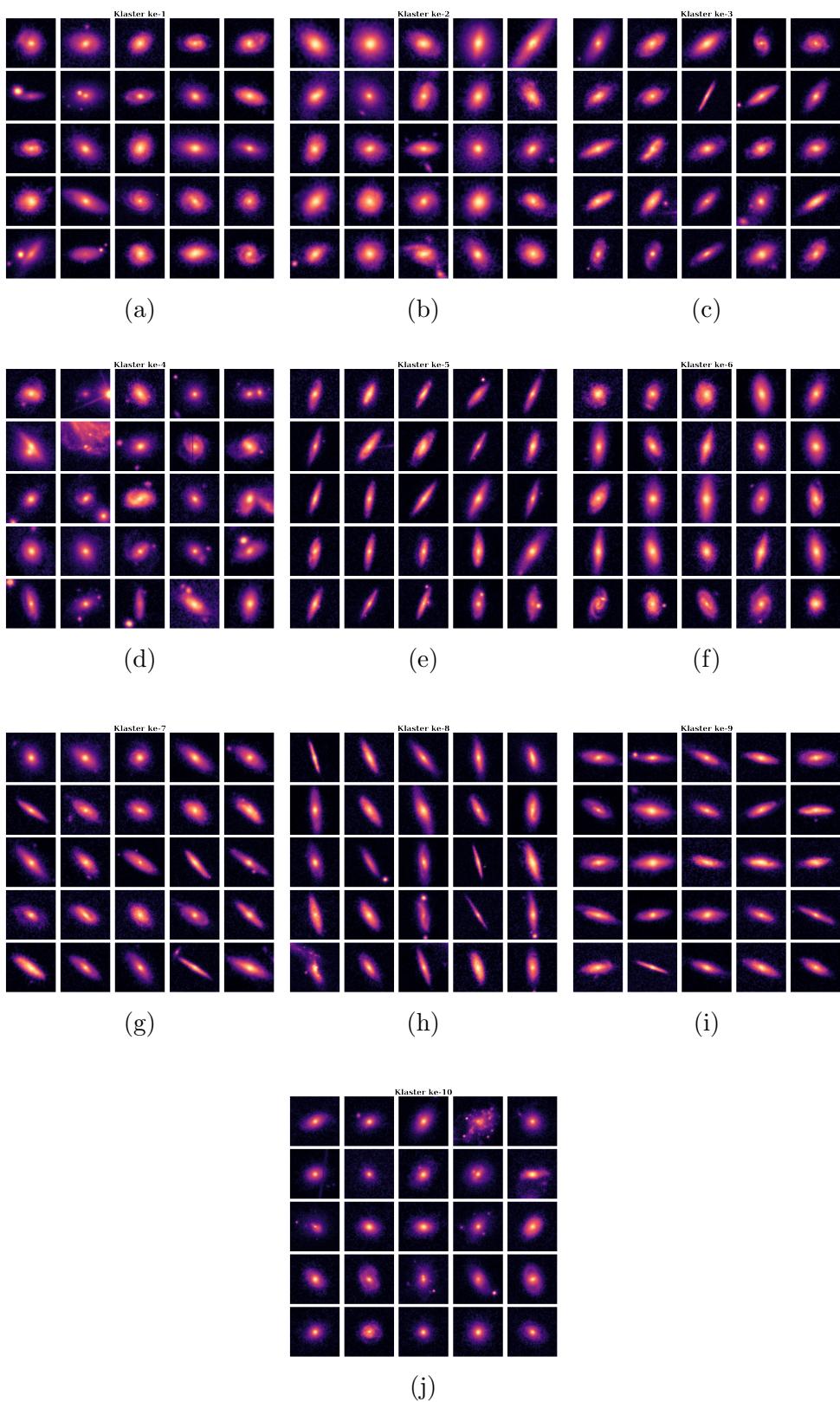
Berdasarkan Gambar IV.12, galaksi dekat tidak berhasil sepenuhnya dike-lompokan berdasarkan morfologi ke dalam tipe galaksi elips dan spiral, meski terlihat ada sedikit kecenderungan klaster pertama berisi galaksi elips, semen-tara klaster kedua berisi galaksi spiral. Namun, pada masing-masing klaster masih terlihat adanya galaksi elips dan spiral bersamaan.

Selain dilihat secara visual, pengelompokan galaksi dengan metode ini juga dapat dibandingkan dengan informasi kategori morfologi dari katalog. Gambar IV.13 menunjukkan distribusi morfologi galaksi berdasarkan katalog *Galaxy Zoo* terhadap hasil pengelompokan menggunakan metode *hierarchical clustering*. Berdasarkan gambar tersebut, tampak bahwa hasil pengelompokan menggunakan metode *hierarchical clustering* belum berhasil mengelompokan galaksi dekat langsung ke dalam kelompok galaksi elips dan spiral. Hal ini dapat dilihat dari galaksi elips banyak berada di klaster 2, begitu pula untuk galaksi spiral.



Gambar IV.13: Sebaran data galaksi berdasarkan morfologi katalog terhadap hasil *hierarchical clustering*.

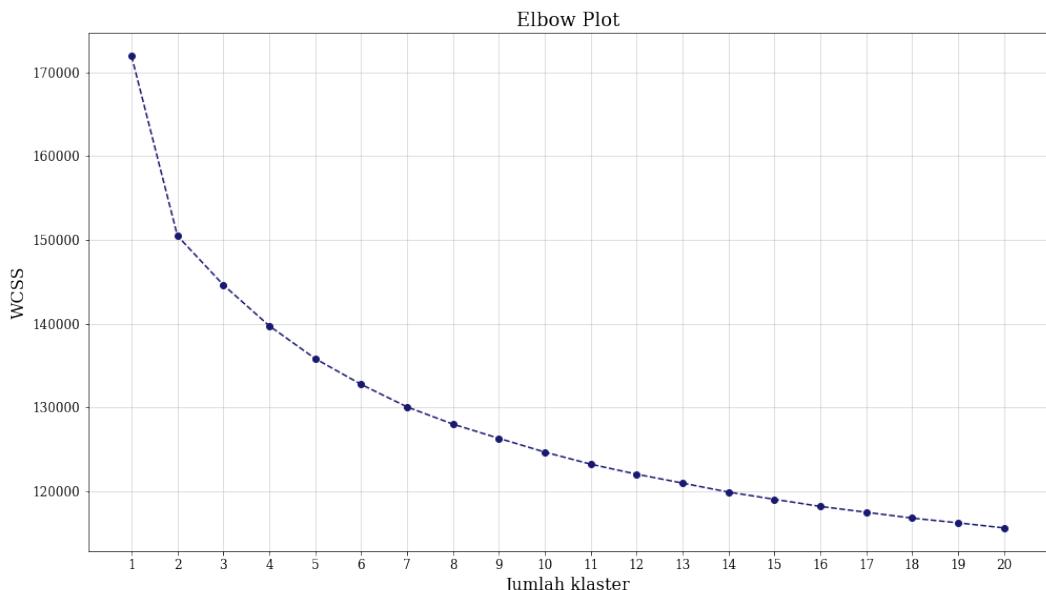
Jika dilihat dalam jumlah klaster yang lebih banyak, misalnya 10 klaster, sampel galaksi dari masing-masing klaster dapat dilihat pada Gambar IV.14. Terlihat bahwa pengelompokan dilakukan berdasarkan arah *position angle* galaksi. Selain itu, beberapa klaster tampak berisi galaksi-galaksi dengan posisi *face-on* seperti pada klaster ke-1, klaster ke-2, klaster ke-4, dan klaster ke-10. Sementara itu, klaster lainnya tampak berisi galaksi-galaksi *edge-on*. Klaster 10 juga tampak memiliki pusat galaksi yang lebih redup dibandingkan klaster lainnya, namun dengan ukuran galaksi yang sama. Dengan demikian, pengelompokan galaksi cukup berhasil dilakukan namun tidak hanya berdasarkan struktur atau morfologi galaksi. Terdapat faktor lain yang mengontaminasi, seperti *position angle* dan kecerlangan galaksi. Hal ini menunjukkan perlunya dilakukan *pre-processing* data yang lebih *robust* sebelum menerapkan algoritma VAE.



Gambar IV.14: Sampel galaksi dekat hasil pengelompokan dengan metode *hierarchical clustering* ke dalam 10 klaster.

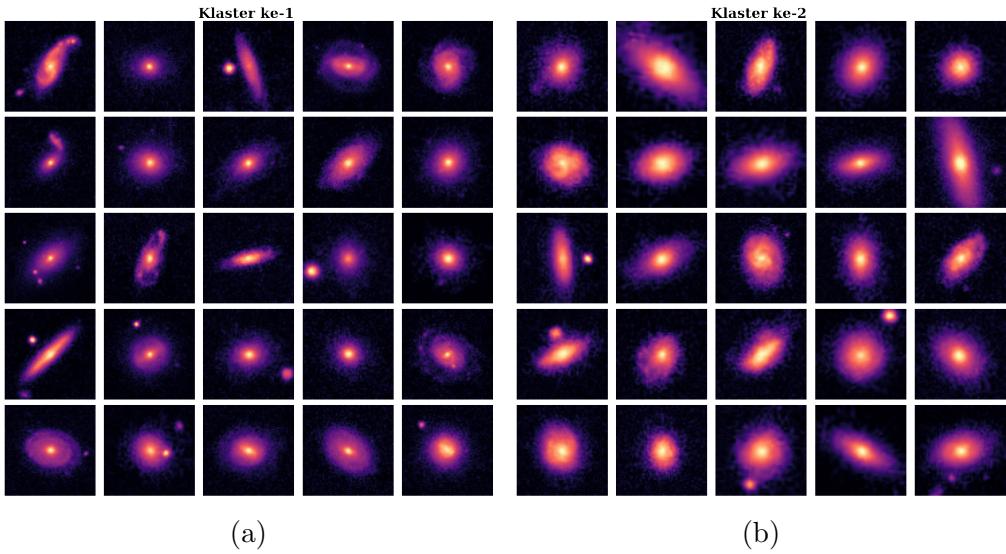
IV.2.3 *K-Means Clustering*

Selain dengan metode *hierarchical clustering*, pengelompokan juga akan dilakukan dengan metode *k-means clustering*. Pertama akan dibuat *elbow plot* untuk melihat jumlah klaster terbaik untuk mengelompokan data galaksi dekat. Gambar IV.15 menunjukkan *elbow plot* untuk data galaksi dekat. Berdasarkan gambar tersebut, tekukan terlihat signifikan pada jumlah klaster dua, sehingga untuk metode ini data akan dikelompokan ke dalam dua klaster.



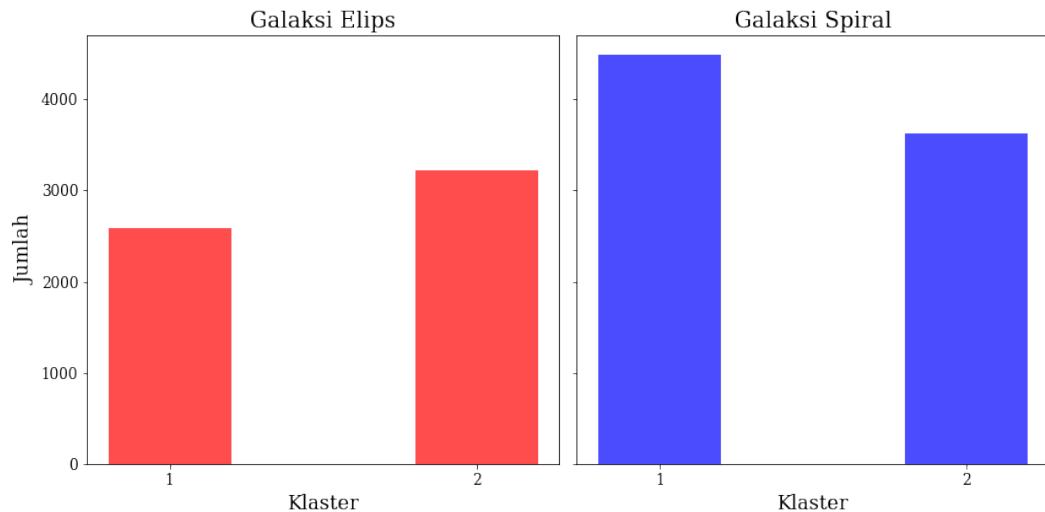
Gambar IV.15: *Elbow plot* untuk menentukan jumlah klaster dalam *K-Means clustering* terhadap data galaksi dekat.

Gambar IV.16 menunjukkan sampel galaksi dekat hasil pengelompokan menggunakan metode *k-means clustering*. Berdasarkan gambar tersebut, pengelompokan tidak tampak berdasarkan bentuknya, melainkan berdasarkan kecerlangan permukaan galaksi. Klaster pertama tampak lebih redup dibandingkan dengan klaster kedua. Hal ini menunjukkan bahwa pengelompokan dengan metode *k-means clustering* belum berhasil mengelompokan galaksi dekat langsung ke dalam kategori galaksi elips dan spiral.



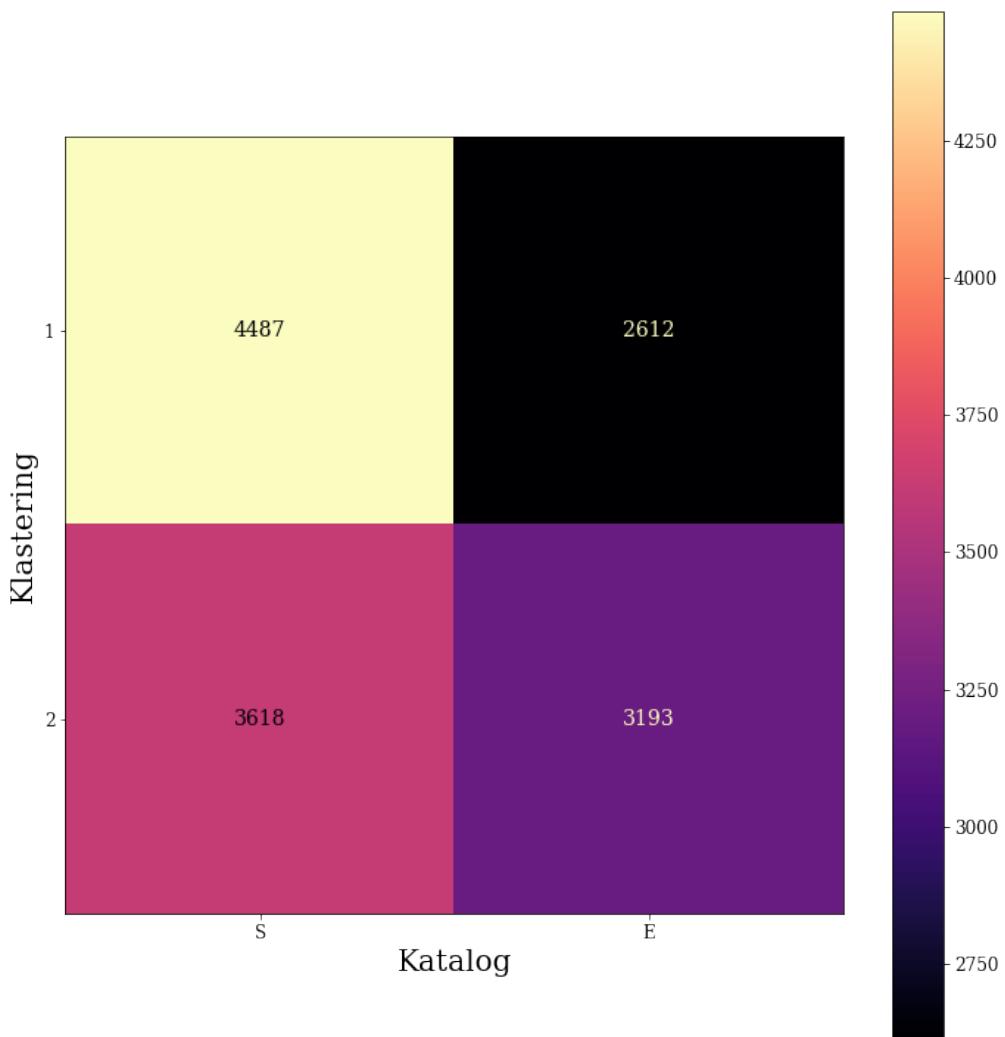
Gambar IV.16: Sampel galaksi dekat dari pengelompokan dengan metode *k-means clustering*. Masing-masing menunjukkan klaster pertama (panel a), klaster kedua (panel b), dan klaster ketiga (panel c).

Sama seperti sebelumnya, hasil pengelompokan dengan metode *k-means clustering* juga akan dibandingkan dengan informasi morfologi galaksi dari katalog *Galaxy Zoo*. Gambar IV.17 menunjukkan distribusi morfologi galaksi berdasarkan katalog terhadap hasil pengelompokan menggunakan metode *k-means clustering*. Tampak bahwa galaksi elips lebih banyak dikategorikan dalam klaster kedua, sementara galaksi spiral lebih banyak dikategorikan dalam klaster pertama. Tidak seperti sebaran hasil pengelompokan *hierarchical clustering* pada Gambar IV.13 yang menunjukkan kedua tipe galaksi dominan dalam satu klaster yang sama, metode *k-means* menunjukkan tipe galaksi yang berbeda dominan pada klaster yang berbeda.



Gambar IV.17: Sebaran data galaksi berdasarkan morfologi katalog terhadap hasil *k-means clustering*.

Berdasarkan Gambar IV.13 dan IV.17, dapat dilihat bahwa dengan metode *clustering* yang berbeda akan menghasilkan kelompok galaksi yang berbeda meski dari satu data yang sama. Meski demikian, dengan kedua metode tersebut pengelompokan belum berhasil memisahkan data langsung ke dalam dua kategori galaksi, yakni galaksi elips dan galaksi spiral. Namun, apabila klaster dibuat lebih banyak, karakteristik masing-masing klaster akan terlihat lebih jelas.



Gambar IV.18: *Confusion matrix* pengelompokan galaksi dekat dengan metode *k-means clustering*.

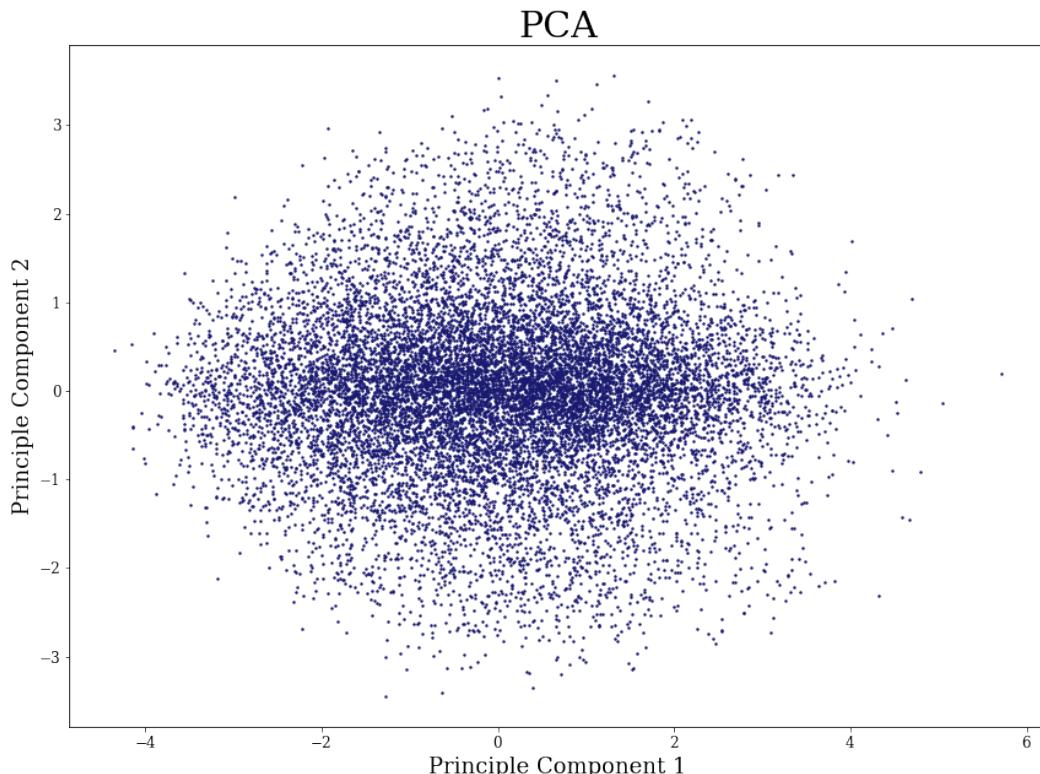
Jika dengan menggunakan metode *hierarchical clustering* sebelumnya kita tidak dapat menentukan galaksi elips dan galaksi spiral berada di klaster pertama atau klaster kedua, dengan metode *k-means clustering*, klaster pertama tampak cukup berkorelasi dengan galaksi spiral, sehingga diasumsikan klaster kedua berkorelasi dengan galaksi elips. Jika dilihat dari *confussion matrix* pada Gambar IV.18, maka akurasi dari pengelompokan galaksi dekat menggunakan metode *k-means clustering* sebesar 0.552 dan presisi mencapai 0.632. Nilai akurasi dan presisi akan dihitung melalui persamaan IV.3 dan IV.4.

$$akurasi = \frac{Jumlah\ prediksi\ tepat}{Total\ prediksi} \quad (\text{IV.3})$$

$$presisi = \frac{\text{Jumlah prediksi kategori } X \text{ tepat}}{\text{Total prediksi kategori } X} \quad (\text{IV.4})$$

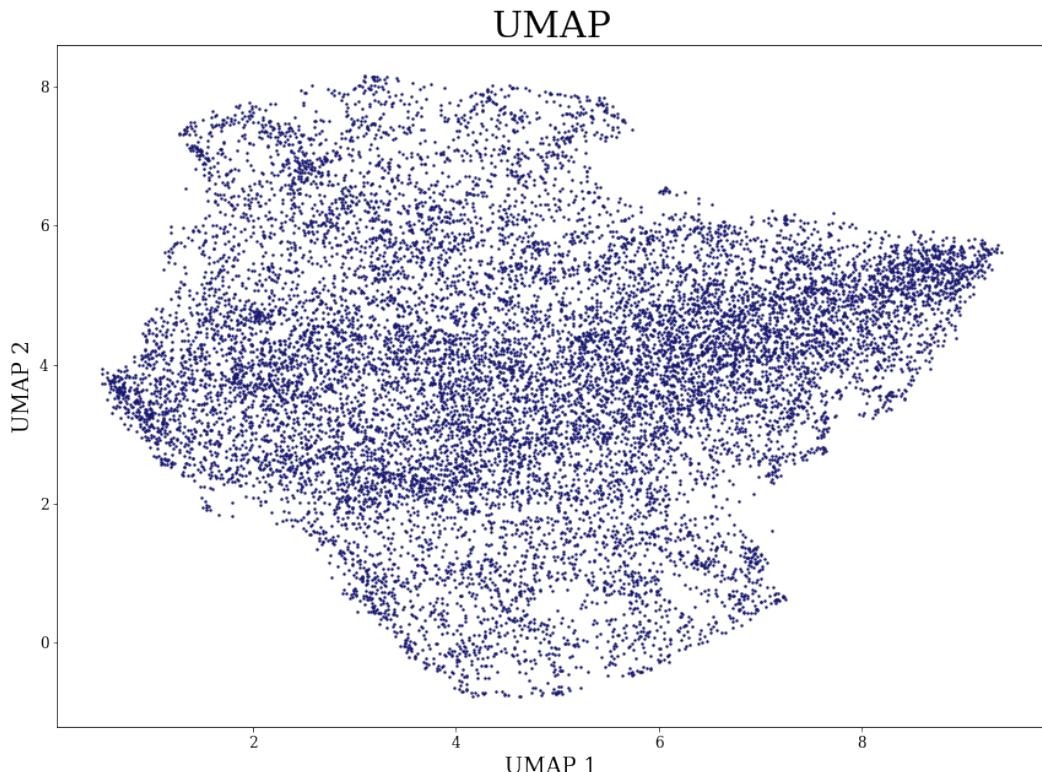
IV.2.4 PCA dan UMAP

Clustering galaksi dekat juga dapat dilihat dari representasi parameter laten ke dalam dua dimensi melalui metode reduksi dimensi. Reduksi dimensi dilakukan dengan metode PCA dan juga UMAP.



Gambar IV.19: Proyeksi parameter laten dalam dua dimensi menggunakan PCA untuk galaksi dekat.

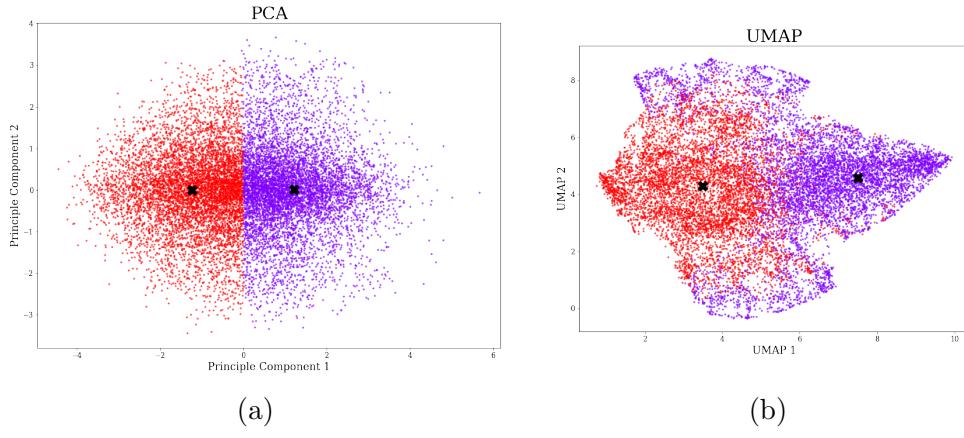
Gambar IV.19 menunjukkan proyeksi parameter laten 100 dimensi yang telah direduksi menjadi hanya dua dimensi dengan menggunakan metode PCA. Dalam plot tersebut sebaran data hanya terlihat sebagai satu klaster, dan tidak terlihat adanya pengelompokan data.



Gambar IV.20: Proyeksi parameter laten dalam dua dimensi menggunakan UMAP untuk galaksi dekat.

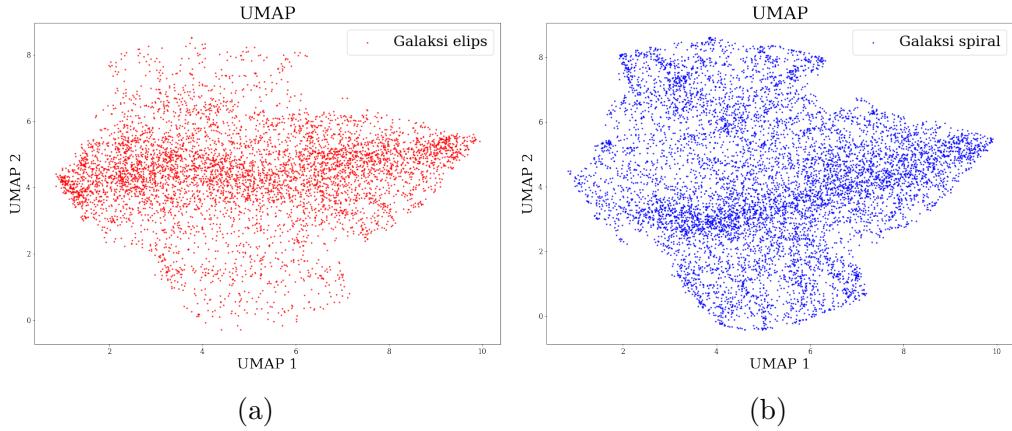
Sementara itu, Gambar IV.20 menunjukkan proyeksi parameter laten 100 dimensi ke dalam dua dimensi menggunakan metode UMAP. Dengan metode ini, persebaran data tampak tidak merata, dimana bagian tengah distribusi tampak lebih padat dibandingkan bagian atas dan bawah.

Pengelompokan dengan metode *k-means clustering* dari Subbab IV.2.3 dapat divisualisasikan dalam plot PCA dan UMAP. Gambar IV.21 menunjukkan hasil pengelompokan galaksi dekat dengan metode *k-means* yang terlihat dari plot PCA dan UMAP, serta posisi *centroid* masing-masing klaster.



Gambar IV.21: Hasil klastering galaksi dekat dengan metode *k-means* yang tampak dari proyeksi parameter laten dengan metode PCA (panel a) dan dengan metode UMAP (panel b).

Jika distribusi galaksi elips dan spiral ditampilkan pada plot representasi parameter laten 100 dimensi, secara umum tidak tampak adanya pengelompokan. Namun ada sedikit perbedaan dalam sebaran distribusi kedua tipe galaksi ini, terutama di area tengah distribusi data. Distribusi kedua tipe galaksi tersebut ditunjukkan pada Gambar IV.22.



Gambar IV.22: Distribusi galaksi elips dan spiral dari representasi parameter laten menggunakan UMAP.

IV.3 Pengelompokan Galaksi Jauh

Pada bagian ini akan dijelaskan hasil pengelompokan morfologi galaksi jauh untuk data sebanyak 14440 galaksi, yakni data dari *field CEERS*, COSMOS, dan PRIMER, dan telah melalui seleksi objek non galaksi dan objek-objek

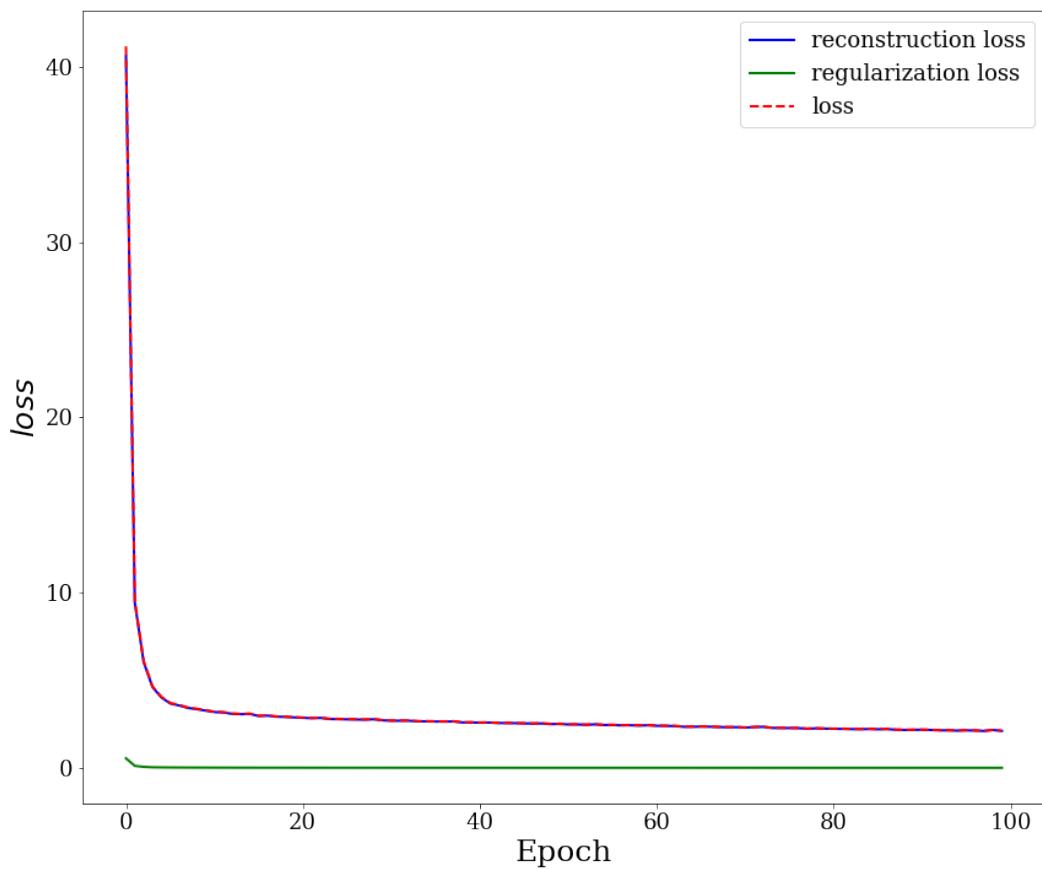
yang sangat redup sebagaimana terlihat pada contoh galaksi pada Gambar III.12, III.13, dan III.14.

IV.3.1 Pengelompokan Galaksi Jauh Secara Global

Pengelompokan galaksi yang dijelaskan pada Subbab ini mengelompokan galaksi pada $r_{eff} \leq 10pix$. Pembatasan ini dilakukan karena ukuran data akan diseragamkan dengan melakukan *rescaling* dan *resizing* data. Dengan demikian, tahapan *pre-processing* data yang diterapkan pada Subbab ini adalah *rotating* data, pemangkasan data, dan normalisasi data. Dengan data yang digunakan adalah data dalam format *grayscale* tanpa mengaplikasikan **Galclean**. Total galaksi dalam dataset ini berjumlah 11064 galaksi, dengan galaksi pada *redshift* terdekat berada di $z = 0.000342$ (z_{spec}) dan galaksi pada *redshift* terjauh berada di $z = 16.834166$ (z_{phot}). Meski di awal seleksi data dilakukan untuk mengambil data hanya pada $z_{phot} > 2$, kenyataannya terdapat sebagian galaksi yang memiliki $z_{spec} < 2$. Oleh karena itu, data *redshift* yang digunakan pada analisis di Subbab ini menggunakan nilai *redshift* spektroskopik, dan jika tidak tersedia akan digunakan nilai *redshift* fotometri.

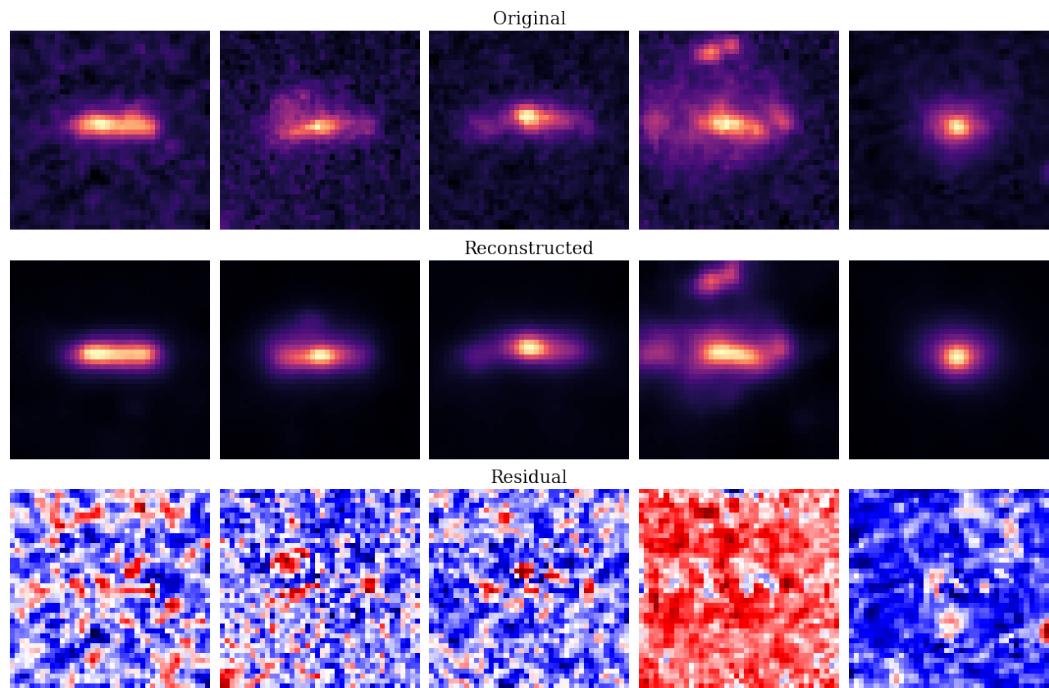
Hasil Autoencoding

Berdasarkan arsitektur yang ditunjukkan pada Gambar III.22 dan III.23, nilai *loss* setelah 100 *epoch* berada di angka ~ 2.5 . Grafik nilai *loss* ditunjukkan pada Gambar IV.23. Nilai *loss* tersebut menunjukkan nilai *reconstruction loss* yang menurun signifikan dalam beberapa *epoch* pertama, lalu relatif konstan hingga *epoch* terakhir. Hal ini menunjukkan bahwa data hasil rekonstruksi secara kuantitatif semakin mirip dengan data asli. Sementara itu, nilai *regularization loss* sejak awal cukup rendah dan tidak menunjukkan penurunan yang signifikan. Hal ini menunjukkan bahwa parameter laten sejak awal didorong untuk mengikuti distribusi normal.



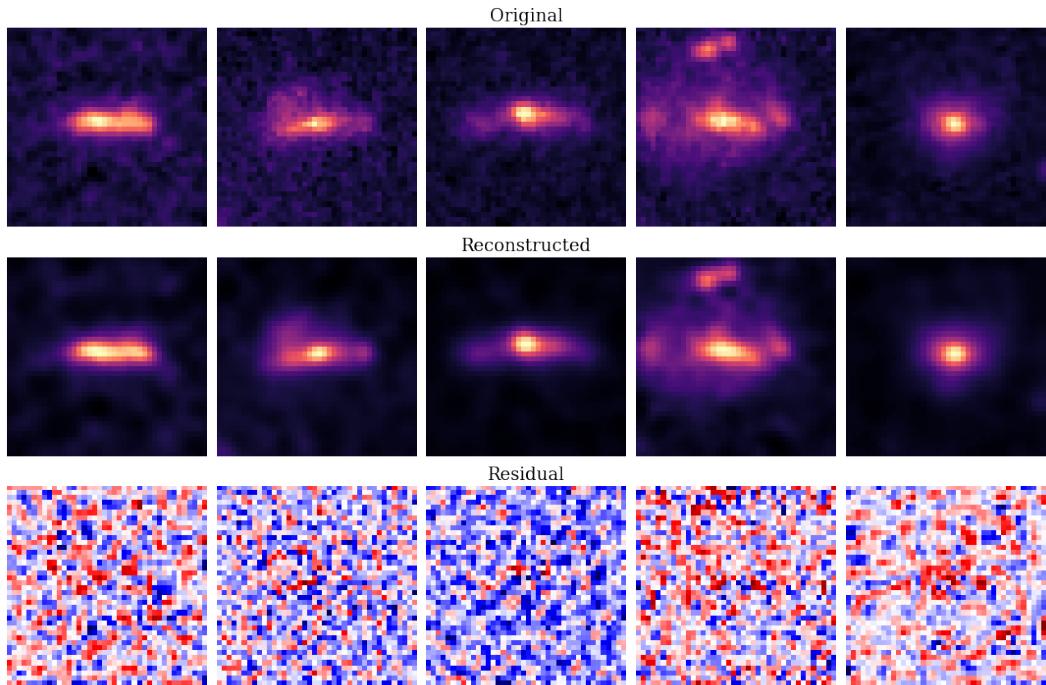
Gambar IV.23: Plot pergerakan nilai *loss* pada setiap *epoch* selama proses *autoencoding* untuk galaksi jauh.

Apabila dibandingkan secara visual, data yang direkonstruksi dari parameter laten telah cukup baik merepresentasikan data asli. Hal ini menunjukkan bahwa parameter laten menyimpan informasi yang memang relevan dengan data. Gambar IV.24 menunjukkan perbandingan data asli dengan data yang direkonstruksi dari fitur-fitur yang telah diekstrak, serta residual yang dihitung dari data asli dikurangi dengan data rekonstruksi.



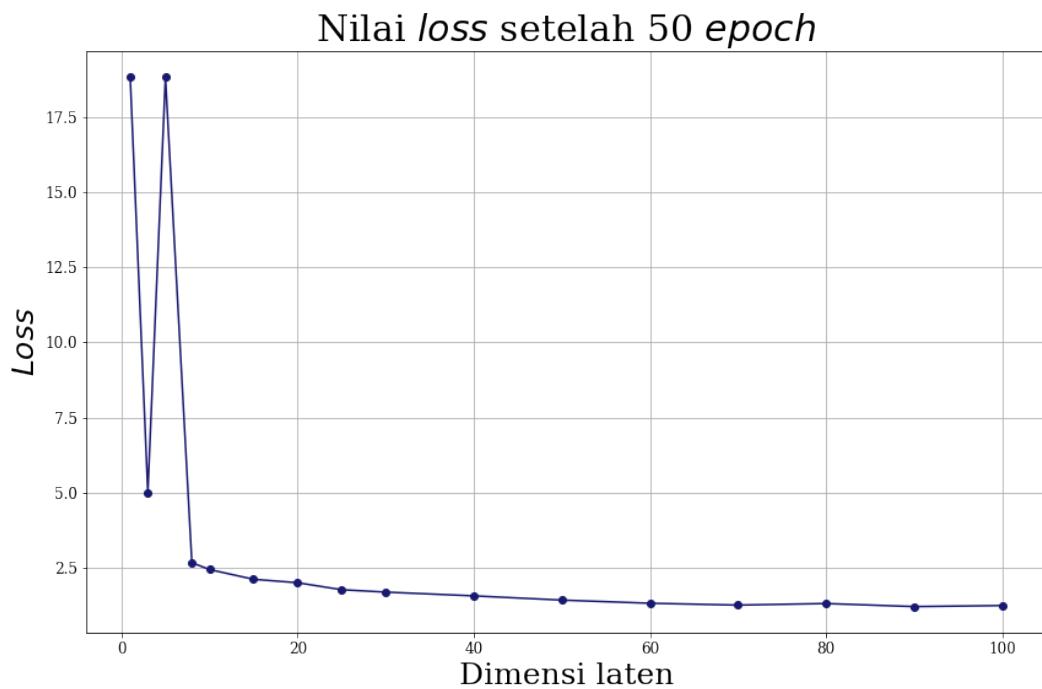
Gambar IV.24: Beberapa sampel *reconstructed images* galaksi jauh yang dibangun dari 10 parameter laten dibandingkan terhadap data input dan residualnya.

Sebagai perbandingan, jika data direpresentasikan dalam parameter laten yang lebih banyak, misalnya 100 parameter laten seperti yang dilakukan terhadap galaksi dekat, data rekonstruksi tetap tampak memiliki *noise* latar belakang yang cukup besar. Gambar IV.25 menunjukkan galaksi yang sama seperti pada Gambar IV.24 jika data direpresentasikan dalam 100 parameter laten.



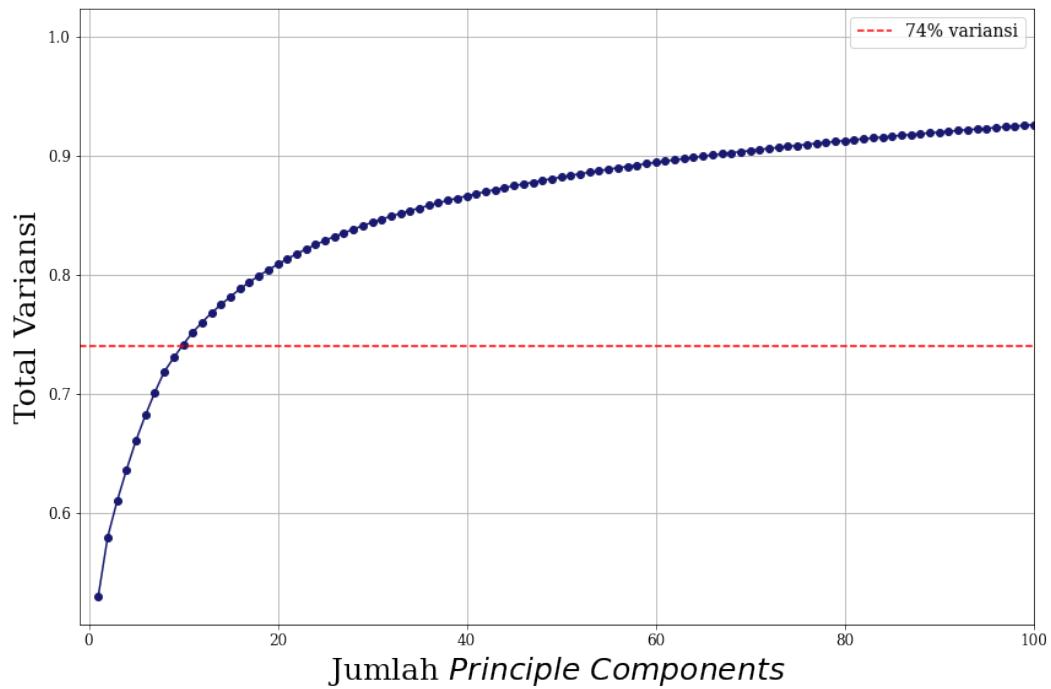
Gambar IV.25: Plot perbandingan *reconstructed images* untuk 100 parameter laten terhadap data input dan residualnya, pada galaksi jauh.

Penentuan jumlah parameter laten yang baik dapat dilakukan dengan membandingkan nilai *loss* pada *epoch* terakhir dari proses VAE dengan berbagai parameter laten. Dengan kata lain, seluruh parameter dan data dibuat sebagai variabel kontrol selain dari dimensi laten sebagai variabel bebas. Gambar IV.26 menunjukkan perbandingan nilai *loss* dari *epoch* ke-50 untuk proses VAE dengan beberapa dimensi laten. Terlihat bahwa semakin besar dimensi laten, nilai *loss* semakin kecil, namun semakin besar dimensi laten, penurunan nilai *loss* semakin tidak signifikan. Perlu diperhatikan bahwa nilai *loss* yang sangat rendah berpotensi membuat mesin ikut merekonstruksi *noise* di dalam data. Oleh karena itu, dimensi laten berukuran 10 dianggap cukup baik untuk hanya menyimpan fitur-fitur penting di dalam data.



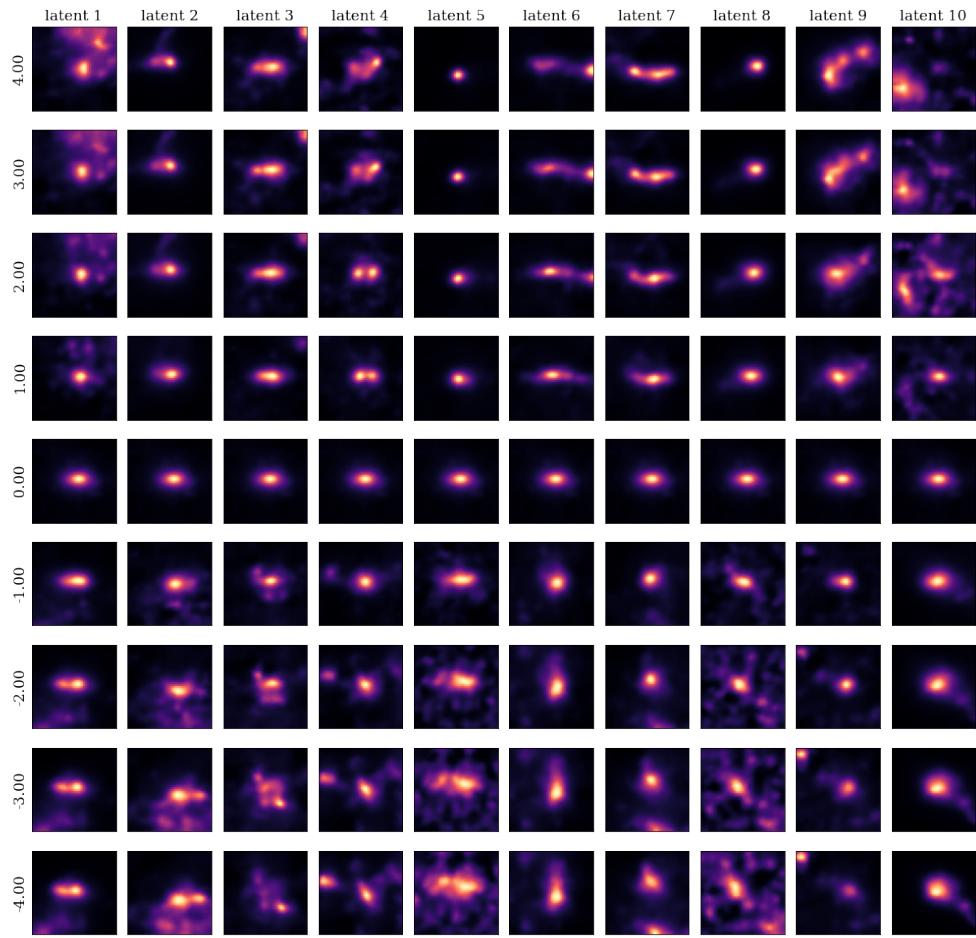
Gambar IV.26: Perbandingan nilai *loss* setelah 50 epoch untuk beberapa nilai dimensi laten.

Selain dengan membandingkan nilai *loss* untuk berbagai dimensi laten, penentuan dimensi laten juga dapat dilihat menggunakan metode PCA seperti yang dilakukan terhadap data galaksi dekat. Metode ini akan membangun sejumlah *principle components* yang merupakan kombinasi linear dari data. Jumlah *principle components* yang bisa memberikan nilai variansi data secara signifikan menjadi indikasi jumlah dimensi laten yang dapat dipilih. Gambar IV.27 menunjukkan bahwa pemilihan dimensi laten 10 sudah dapat merepresentasikan lebih dari 70% variansi data.



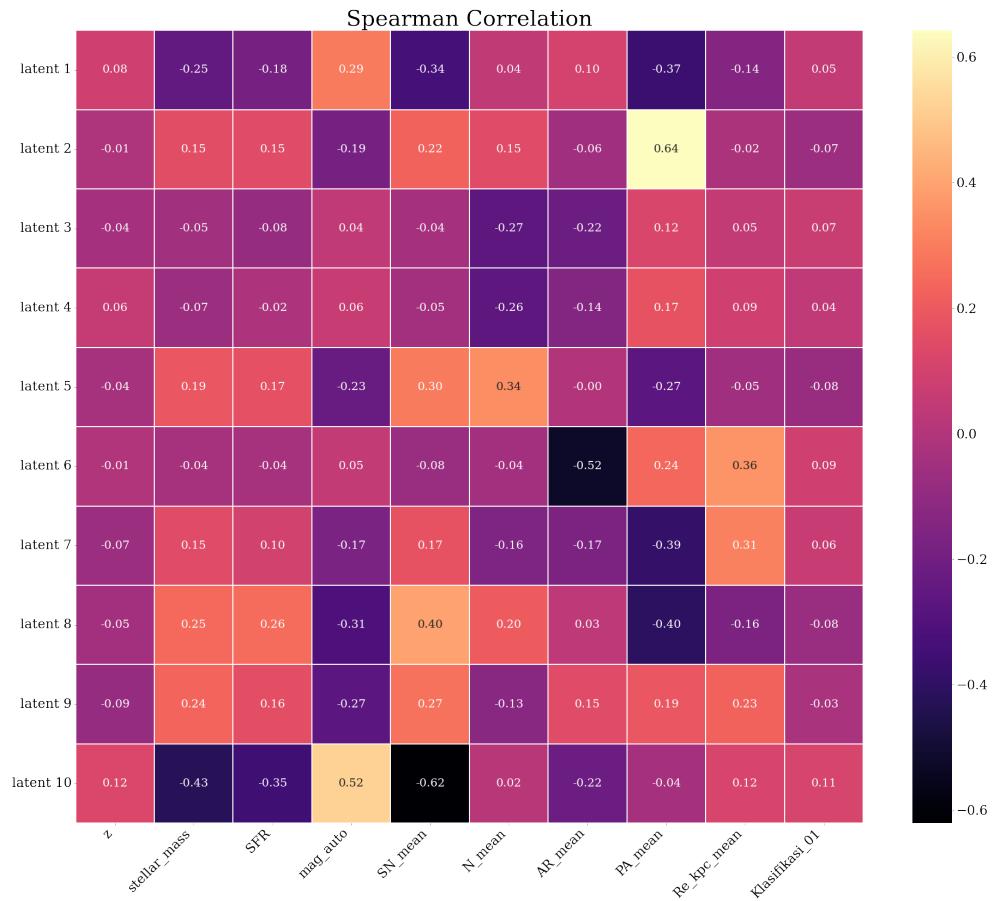
Gambar IV.27: Nilai total variansi untuk berbagai jumlah *principle components*.

Berdasarkan arsitektur *encoder* yang ditunjukkan pada Gambar III.22, data akan direpresentasikan dalam 10 parameter laten. Setiap parameter laten tersebut dapat direpresentasikan secara visual sebagaimana ditunjukkan dalam Gambar IV.28.



Gambar IV.28: Visualisasi dari parameter laten.

Selanjutnya, akan dilihat hubungan antara setiap parameter laten terhadap parameter galaksi, seperti massa, laju pembentukan bintang (SFR), indeks *Sérsic*, dan beberapa parameter lainnya. Namun, beberapa parameter galaksi hasil *fitting Galfitm*, seperti indeks *Sérsic* dan *position angle*, memiliki nilai yang berbeda untuk setiap filter, sehingga dalam analisis ini akan digunakan nilai rata-ratanya. Hubungan antara parameter laten dan parameter galaksi dilihat berdasarkan nilai korelasinya. Digunakan korelasi *spearman* karena dalam kasus ini tidak ada asumsi adanya hubungan linear antara parameter laten terhadap parameter galaksi. Gambar IV.29 menunjukkan matriks korelasi antara parameter laten dengan parameter galaksi.



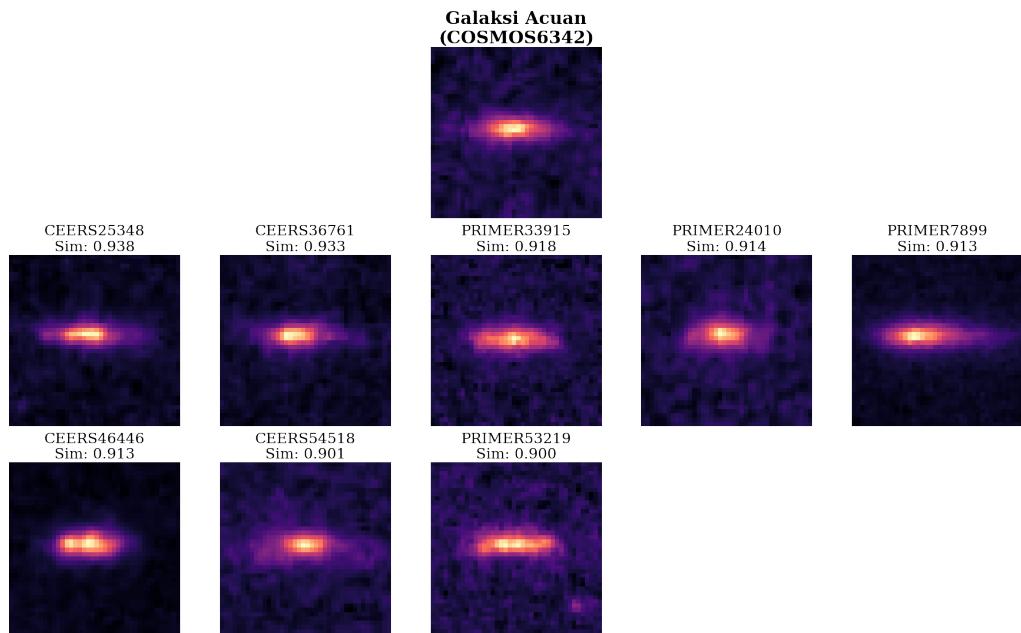
Gambar IV.29: Matriks korelasi parameter laten terhadap parameter galaksi.

Dari matriks korelasi di atas, terlihat bahwa banyak parameter galaksi yang berkorelasi dengan parameter laten 10, seperti massa, SFR, magnitudo, dan S/N. Nilai indeks Sérsic memiliki korelasi tertinggi dengan parameter 5, meski nilai korelasinya hanya 0.34. Nilai *axis ratio* dan radius efektif berkorelasi dengan parameter laten 6. Meski telah dilakukan proses *rotating* data untuk menyeragamkan nilai PA setiap galaksi, namun nyatanya parameter *position angle* masih memiliki korelasi yang cukup besar dengan parameter laten 2. Sementara itu, *redshift* dan klasifikasi galaksi berdasarkan pembentukan bintang tidak menunjukkan korelasi terhadap seluruh parameter laten, terlihat dari nilai korelasi tertingginya yang hanya 0.1.

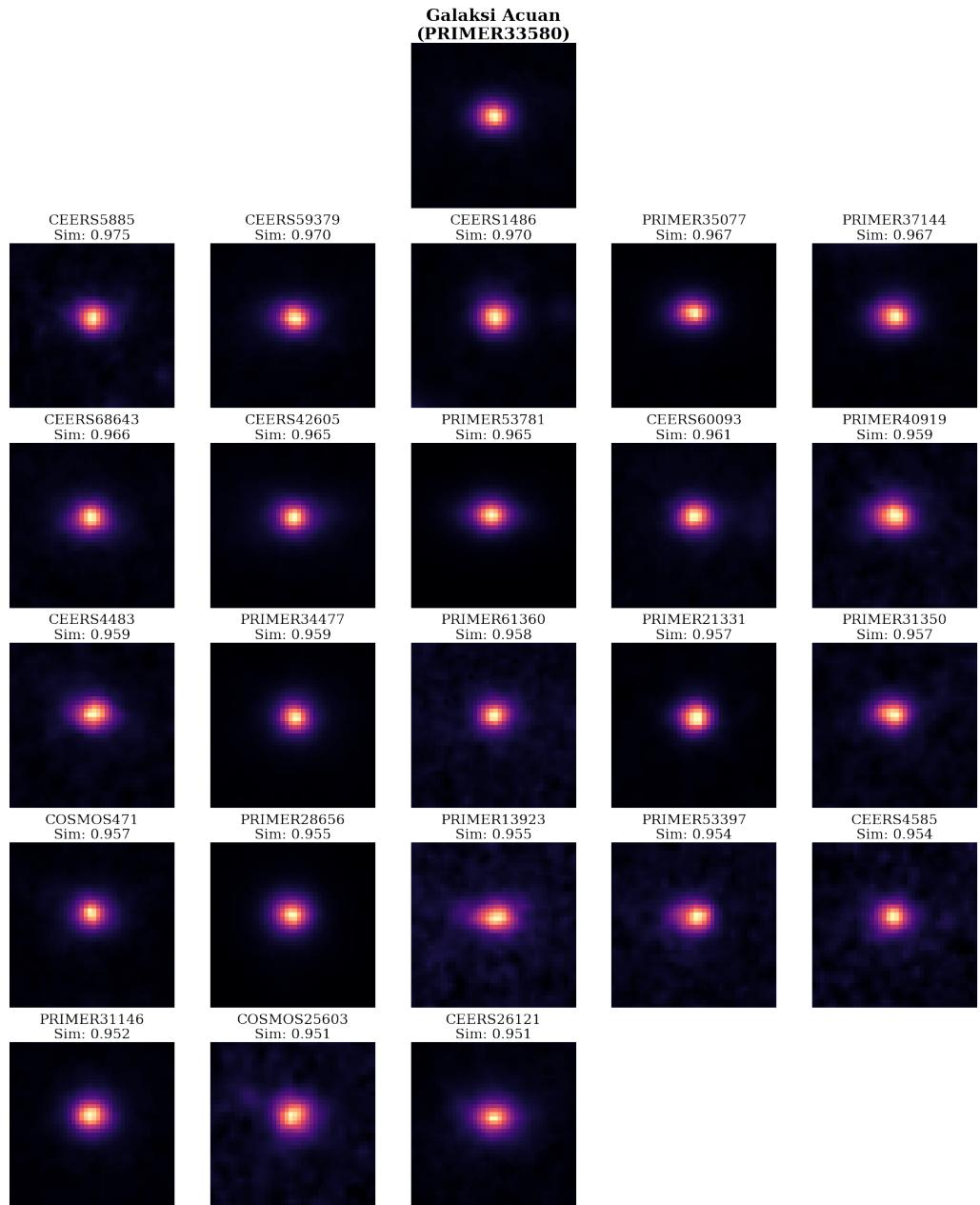
Similarity Cosine

Salah satu cara untuk melihat keberhasilan VAE dalam mengekstrak fitur yang relevan dengan morfologi galaksi adalah dengan membandingkan galaksi-galaksi yang telah diberi label morfologi. Dengan memanfaatkan pelabelan

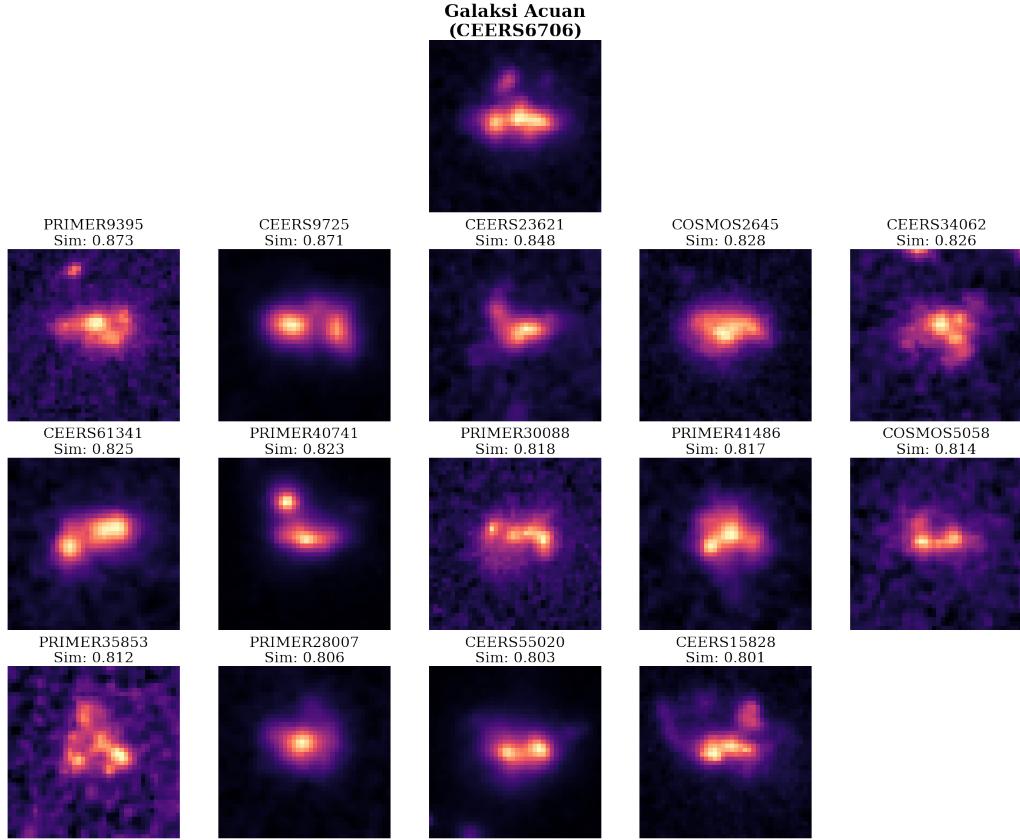
morfologi yang dilakukan pada penelitian Effendi (2024), akan dilihat galaksi-galaksi dengan nilai kemiripan yang tinggi terhadap sampel galaksi yang telah diberi label. Gambar IV.30, IV.31, dan IV.32 masing-masing menunjukkan sampel galaksi dari kategori *disk*, *spheroid*, dan *irregular* pada penelitian tersebut dibandingkan dengan galaksi dalam sampel penelitian ini dengan nilai *cosine similarity* di atas *threshold* tertentu (0.9 untuk galaksi *disk*, 0.95 untuk galaksi *spheroid*, dan 0.8 untuk galaksi *irregular*). Galaksi *irregular* cenderung memiliki nilai *similarity cosine* yang rendah dengan galaksi-galaksi lain karena bentuknya yang tidak beraturan. Sementara itu galaksi *spheroid* dengan *similarity cosine* > 0.9 jumlahnya sangat banyak, sehingga sebagai sampel digunakan *threshold* yang lebih tinggi.



Gambar IV.30: Sampel galaksi kategori *disk* dengan galaksi lain yang memiliki *similarity cosine* > 0.9 .



Gambar IV.31: Sampel galaksi kategori *spheroid* dengan galaksi lain yang memiliki *similarity cosine* > 0.95.

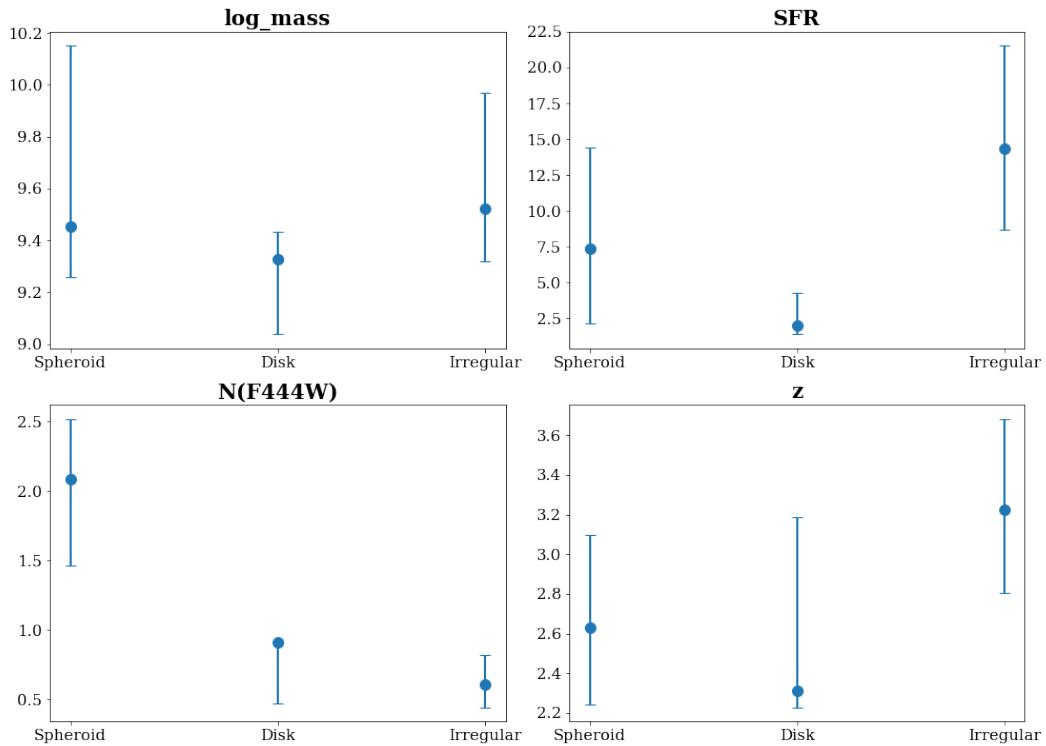


Gambar IV.32: Sampel galaksi kategori *irregular* dengan galaksi lain yang memiliki *similarity cosine* > 0.8.

Secara visual, sampel galaksi pada ketiga contoh galaksi diatas cocok dengan deskripsi morfologi pada masing-masing kategori. Pada masing-masing kategori juga terlihat *noise* tidak memengaruhi nilai *cosine similarity* galaksi. Namun ketiga gambar diatas hanyalah sampel galaksi untuk menunjukkan kemampuan algoritma VAE untuk menangkap fitur-fitur morfologi galaksi. Tiga sampel galaksi di atas tidak bisa memberikan informasi jumlah total galaksi dengan kategori *disk*, *spheroid*, maupun *irregular*.

Untuk sampel ketiga kategori pada Gambar IV.30, IV.31, dan IV.32, dapat dilakukan analisis lebih lanjut mengenai distribusi parameter galaksi antara ketiga kategori tersebut. Gambar IV.33 menunjukkan distribusi empat parameter galaksi untuk ketiga kategori yang ditunjukkan sebelumnya. Dari gambar tersebut terlihat bahwa median nilai massa dari ketiga kategori galaksi tidak jauh berbeda, namun galaksi *disk* memiliki massa yang cenderung lebih rendah dibandingkan kedua kategori lainnya. Galaksi *irregular* memiliki nilai SFR yang lebih tinggi, dan galaksi *irregular* banyak berada di *redshift* yang lebih tinggi dibandingkan dua kategori galaksi lainnya. Sementara itu, indeks *Sérsic* untuk galaksi *spheroid* paling tinggi dibandingkan galaksi *disk*

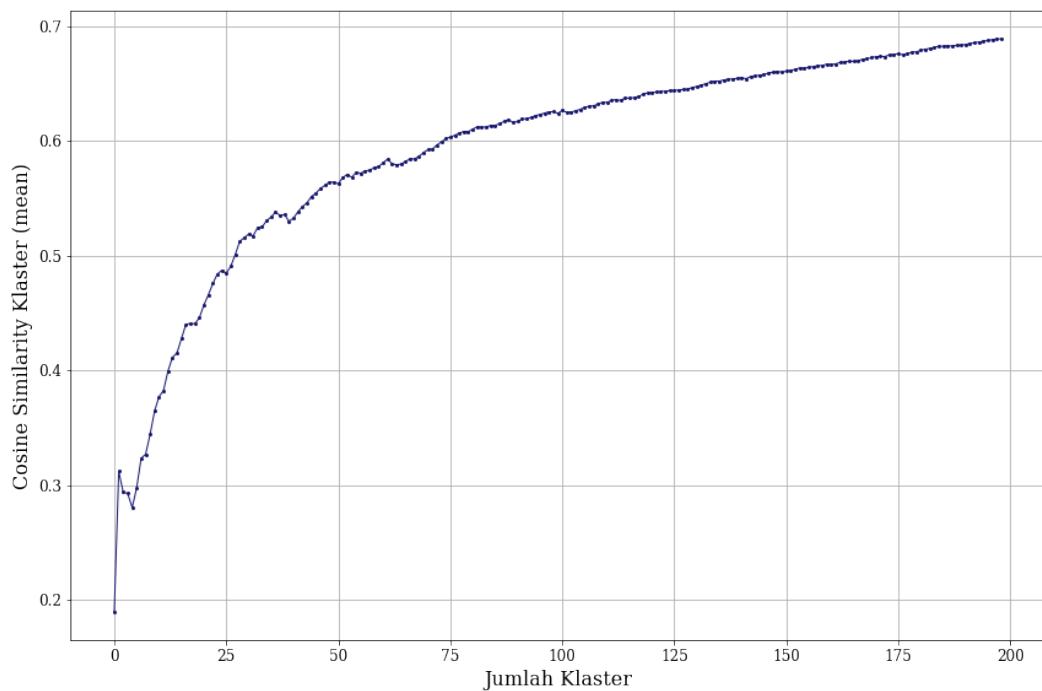
dan *irregular*.



Gambar IV.33: Distribusi beberapa parameter galaksi terhadap tiga sampel kategori galaksi.

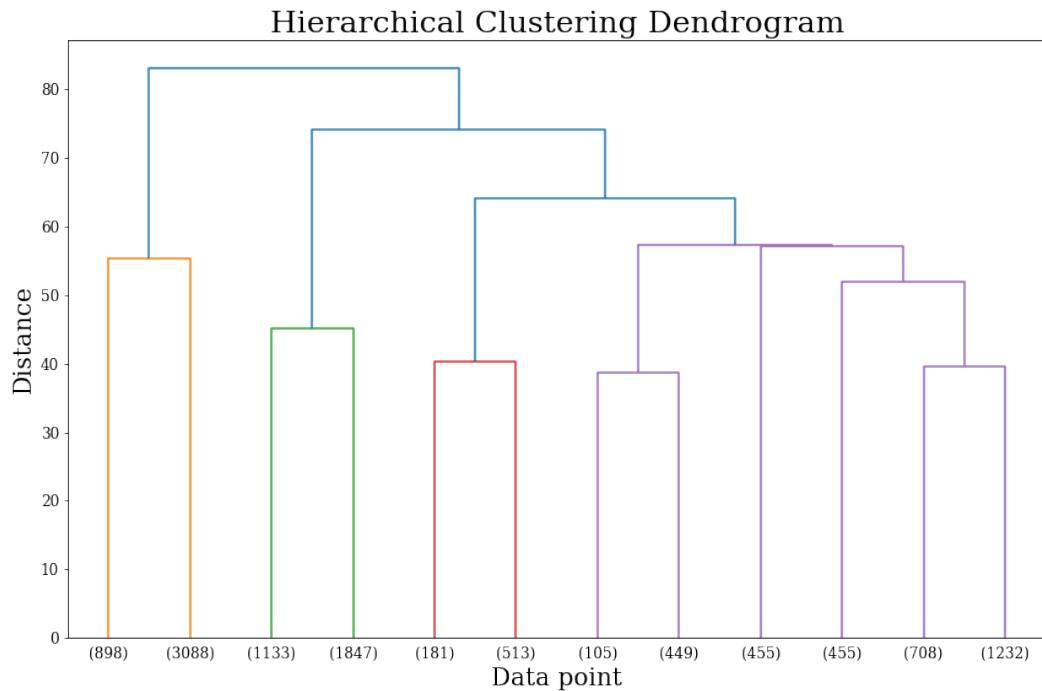
Hierarchical Clustering

Sebelumnya ditunjukkan analisis sampel galaksi dari tiga kategori galaksi yang sudah memiliki label, namun nyatanya belum ada landasan yang jelas mengenai berapa banyak tipe morfologi galaksi di *redshift* tinggi. Gambar IV.34 menunjukkan plot antara nilai rata-rata *cosine similarity*, apabila pengelompokan dipotong hingga jumlah klaster tertentu. Terlihat dari plot tersebut, semakin banyak klaster, kemiripan antar anggota klaster semakin besar.



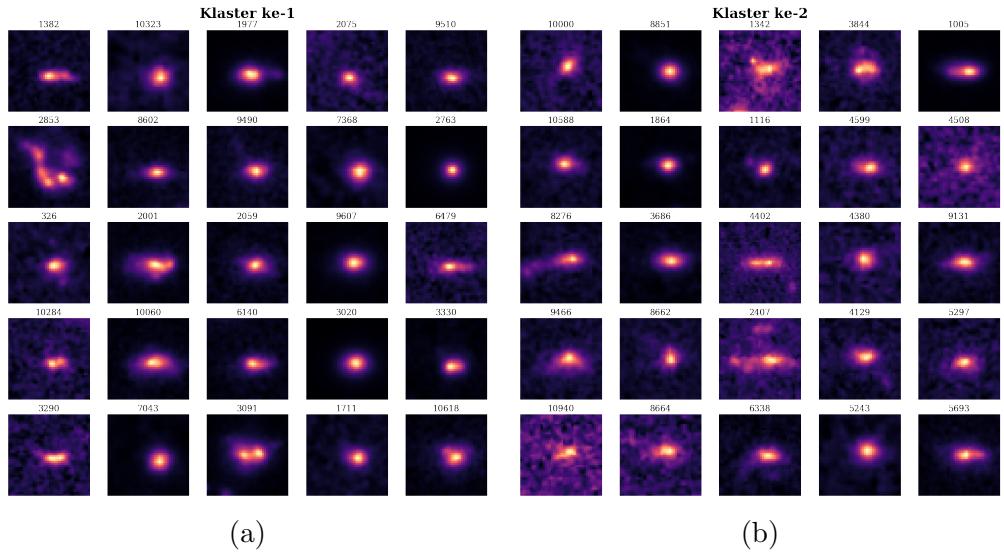
Gambar IV.34: Perbandingan nilai rata-rata *cosine similarity* masing-masing klaster terhadap jumlah klaster.

Untuk bisa melihat pola yang dipelajari mesin dalam mengelompokan galaksi, akan dilihat pengelompokan galaksi ke dalam sejumlah besar klaster. Dendrogram dipotong hingga terdapat 12 klaster, dengan rata-rata *cosine similarity* masing-masing klaster sebesar ~ 0.36 . Gambar IV.35 menunjukkan dendrogram yang menjadi acuan dalam pengelompokan galaksi pada Subbab ini.



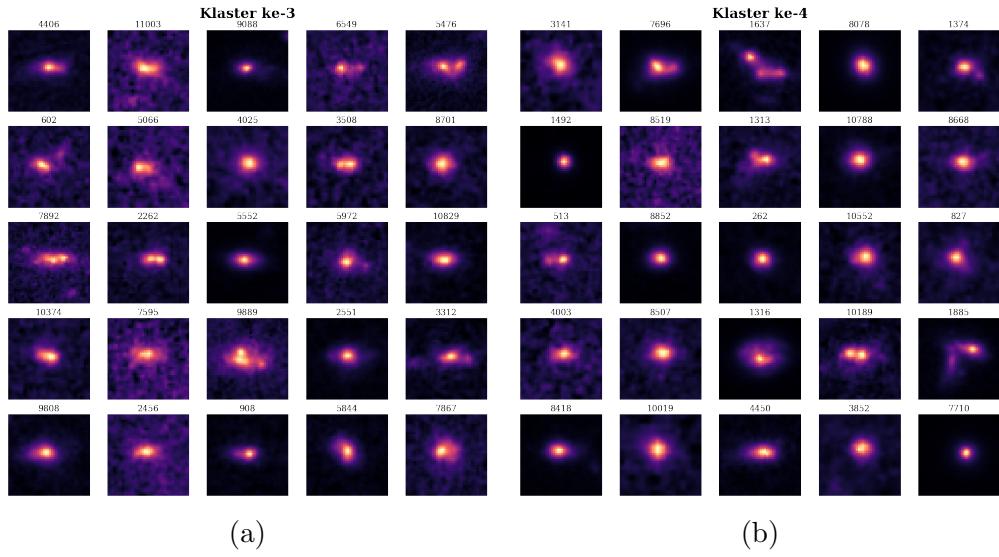
Gambar IV.35: Dendrogram hasil *hierarchical clustering* untuk galaksi jauh.

Berdasarkan dendrogram di atas, pengelompokan akan ditinjau dalam 4 klaster, yakni klaster berwarna jingga, hijau, merah, dan ungu. Pemotongan klaster ke dalam 4 klaster ini berdasarkan batas 70% dari nilai jarak maksimum. Masing-masing klaster berdasarkan warna tersebut memiliki klaster-klaster yang lebih rinci dengan anggota setiap klasternya ditampilkan dalam Gambar IV.36, IV.37, IV.38, dan IV.39.



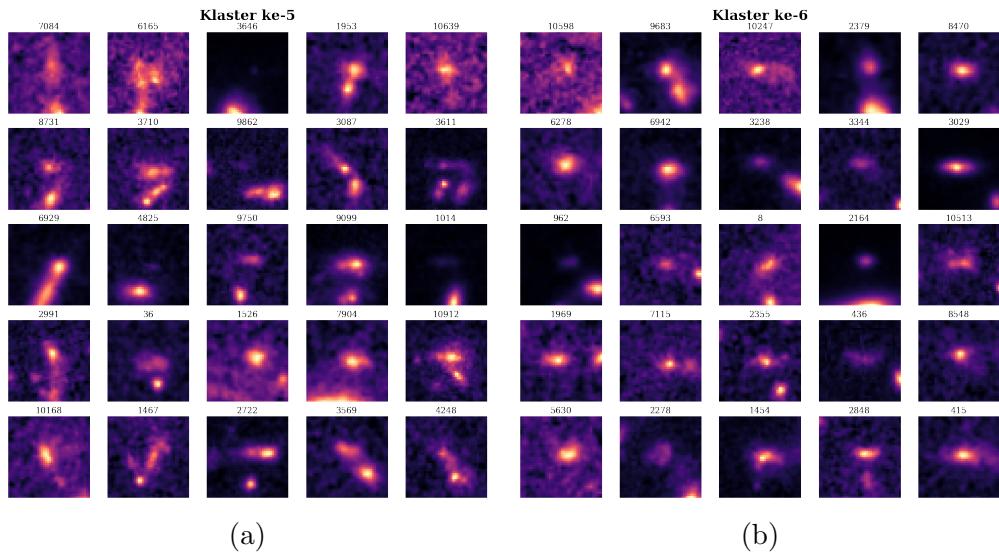
Gambar IV.36: Sampel galaksi jauh dari pengelompokan dengan metode *hierarchical clustering* untuk klaster berwarna jingga.

Klaster jingga pada Gambar IV.36 tampak merupakan klaster dengan posisi galaksi yang tidak tepat di tengah citra, melainkan sedikit bergeser ke arah kanan. Klaster jingga terdiri dari dua klaster, yaitu klaster 1 dan klaster 2. Secara umum, tidak tampak adanya karakteristik khusus yang membedakan antara kedua klaster tersebut, meski ada sedikit kecenderungan *offset* posisi galaksi di klaster 2 lebih kecil dibandingkan galaksi di klaster 1. *Offset* pada sebagian galaksi ini terjadi karena galaksi yang tidak selalu berada di posisi tengah citra. Tahap *pre-processing* berupa *rotating* dan *resizing* data dapat memperbesar *offset* posisi galaksi dalam citra.



Gambar IV.37: Sampel galaksi jauh dari pengelompokan dengan metode *hierarchical clustering* untuk klaster berwarna hijau.

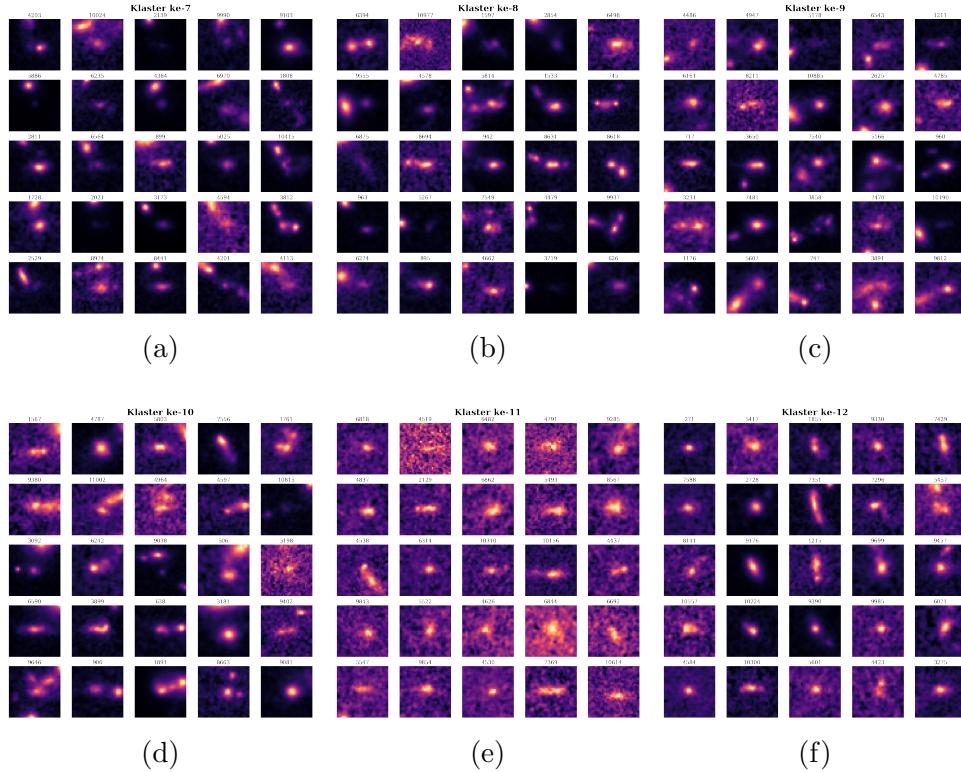
Klaster hijau terdiri dari klaster 3 dan klaster 4. Kedua klaster ini tampak sebagai galaksi yang tidak terkontaminasi objek terang didekatnya. Beberapa galaksi tampak memiliki lebih dari satu inti terang, namun dengan jarak antara keduanya yang cukup berdekatan sehingga dianggap masih menjadi bagian dari galaksi target. Sama seperti sebelumnya, tidak terlihat adanya perbedaan karakteristik yang jelas diantara kedua klaster ini.



Gambar IV.38: Sampel galaksi jauh dari pengelompokan dengan metode *hierarchical clustering* untuk klaster berwarna merah.

Galaksi di dalam klaster merah terdiri dari klaster 5 dan klaster 6. Galak-

si di dalam dua klaster ini tampak sebagai galaksi yang di dalam satu citra terdapat objek terang didekatnya, tetapi tidak tampak sebagai bagian dari galaksi target. Tidak terdapat perbedaan karakteristik yang jelas antara galaksi dalam kedua klaster tersebut.

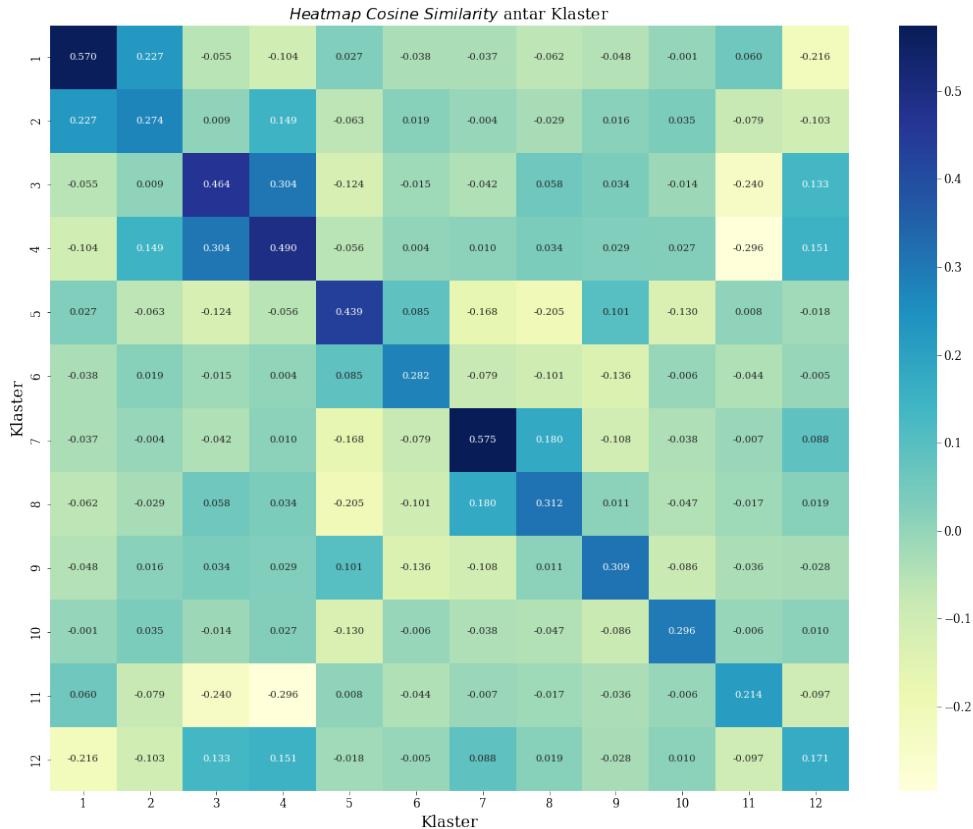


Gambar IV.39: Sampel galaksi jauh dari pengelompokan dengan metode *hierarchical clustering* untuk klaster berwarna ungu.

Klaster ungu terdiri dari klaster 7 hingga klaster 12. Diantara keenam klaster, klaster 11 menjadi klaster yang penampakannya paling berbeda dibandingkan klaster lainnya karena didalamnya berisi galaksi-galaksi dengan *noise* yang tinggi. Meski galaksi dengan *noise* yang besar juga dapat ditemui dalam klaster 1 hingga klaster 12, tetapi klaster 11 tampak seluruhnya berisi galaksi dengan *noise* yang besar. Selain itu, galaksi pada klaster 7 hingga klaster 10 merupakan galaksi yang terdapat objek terang didekatnya, dan tampaknya bukan merupakan bagian dari galaksi target di dalam citra. Klaster 7 juga tampak berisi galaksi yang redup, dan objek terang berada di pojok kiri atas citra.

Secara visual sulit untuk melihat dengan jelas karakteristik galaksi pada masing-masing klaster, serta perbedaan di antara setiap klaster. Secara kuantitatif, kemiripan galaksi dalam masing-masing klaster dapat dihitung dari

nilai *cosine similarity* antar galaksi dalam satu klaster. Di sisi lain, perbedaan antara setiap klaster juga dapat dilihat dengan metode yang sama. Gambar IV.40 menunjukkan *heatmap* berisi nilai *similarity score* antarklaster.



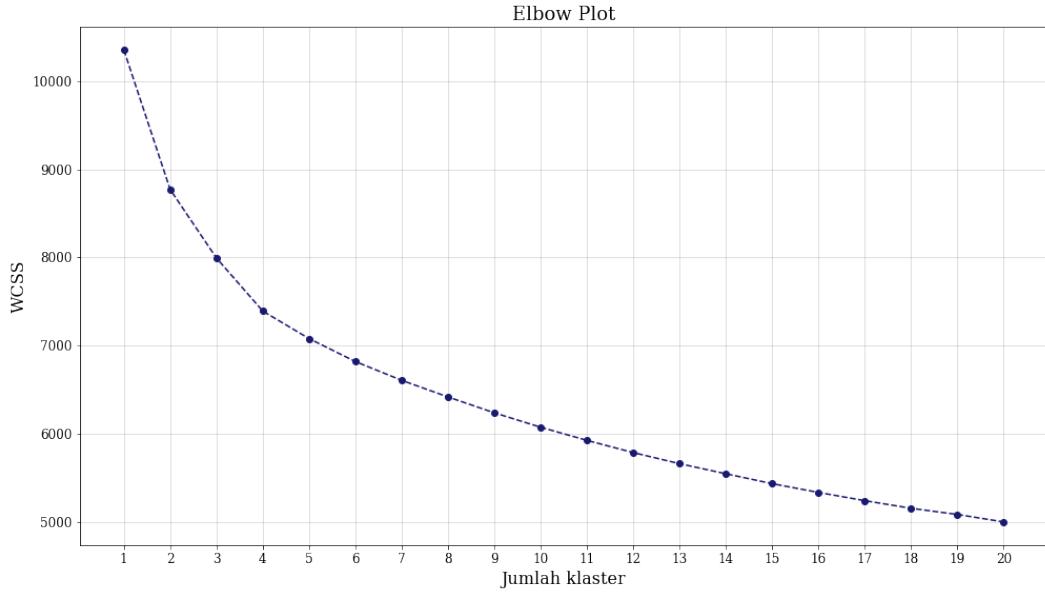
Gambar IV.40: *Heatmap cosine similarity* antar klaster.

Berdasarkan *heatmap* dalam Gambar IV.40, klaster 1 dan klaster 7 memiliki nilai *cosine similarity* terbesar, yang berarti galaksi dalam kedua klaster tersebut paling seragam dibandingkan galaksi dalam klaster lainnya. Di sisi lain, nilai kemiripan antara dua klaster seperti untuk klaster 3 dan klaster 4 tampak lebih besar dibandingkan kemiripan antaranggota klaster 6, klaster 10, klaster 11, dan klaster 12. Hal ini menunjukkan bahwa Klaster 3 dan Klaster 4 memiliki anggota galaksi yang mirip, bahkan lebih mirip dibandingkan antaranggota klaster yang disebutkan sebelumnya.

K-Means Clustering

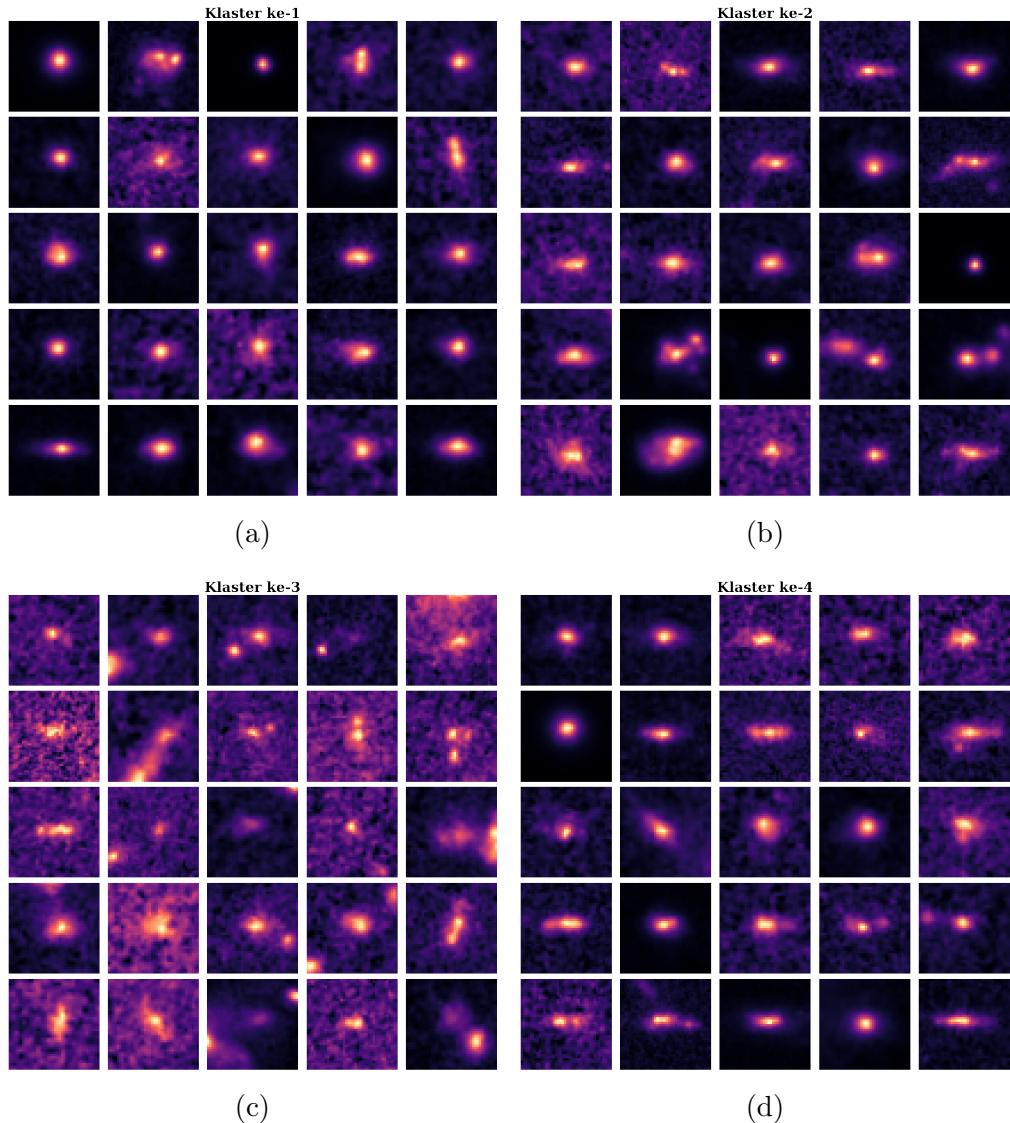
Pengelompokan juga dilakukan dengan metode *K-Means Clustering* untuk membandingkan hasilnya dengan metode sebelumnya. Tetapi, tidak seperti metode sebelumnya dimana pengelompokan dilakukan dengan input parameter laten dalam bentuk vektor 10 dimensi, pengelompokan dengan metode

k-means dilakukan terhadap parameter laten yang dinormalisasi sehingga besar vektor tidak memengaruhi pengelompokan. Mula-mula akan dibuat *elbow plot* untuk melihat jumlah klaster yang sebaiknya dipilih. Berdasarkan plot pada Gambar IV.41, terlihat cukup signifikan pada jumlah klaster 4, sehingga untuk metode *k-means clustering* data akan dikelompokkan ke dalam 4 klaster.



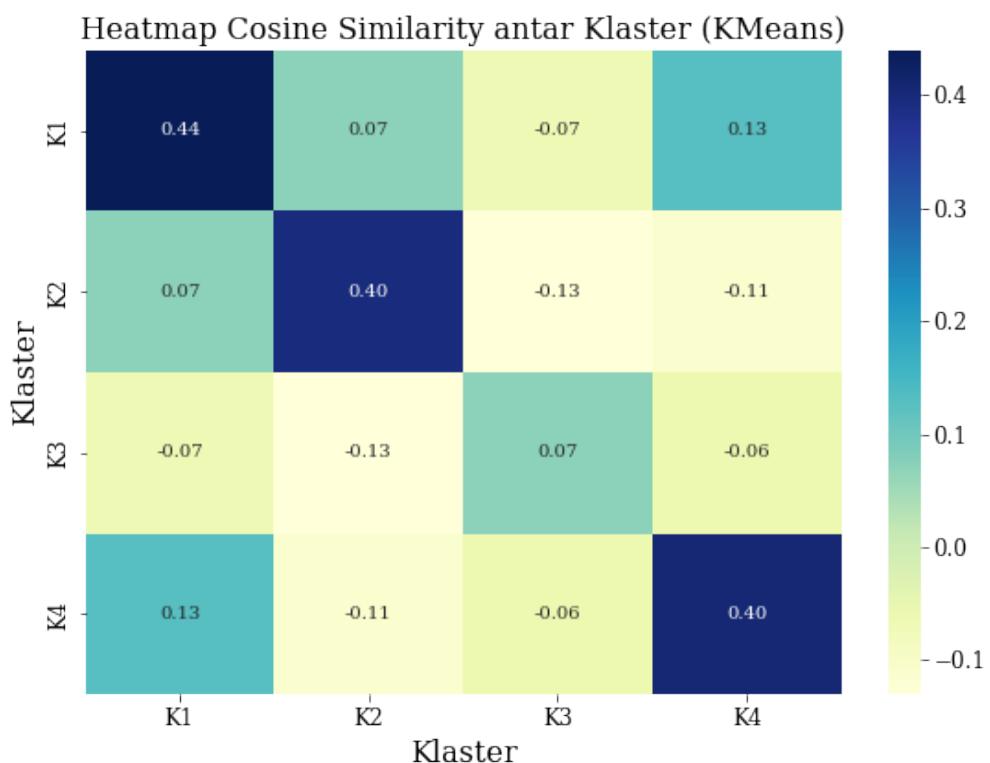
Gambar IV.41: *Elbow plot* untuk menentukan jumlah klaster dalam *K-Means clustering* terhadap data galaksi jauh.

Gambar IV.42 menunjukkan sampel galaksi dari masing-masing klaster hasil pengelompokan menggunakan metode *k-means*. Berdasarkan Gambar IV.42, klaster 3 tampak berisi galaksi dengan *noise* yang paling besar dibandingkan galaksi dalam klaster lainnya. Selain itu, klaster 1 tampak didominasi dengan galaksi dengan bentuk *spheroid*. Galaksi-galaksi dalam klaster 2 dan klaster 4 tidak tampak memiliki karakteristik khusus yang seragam.



Gambar IV.42: Sampel galaksi jauh dari pengelompokan dengan metode *k-means clustering*. Masing-masing menunjukkan klaster pertama (panel a), klaster kedua (panel b), klaster ketiga (panel c), dan klaster keempat (panel d).

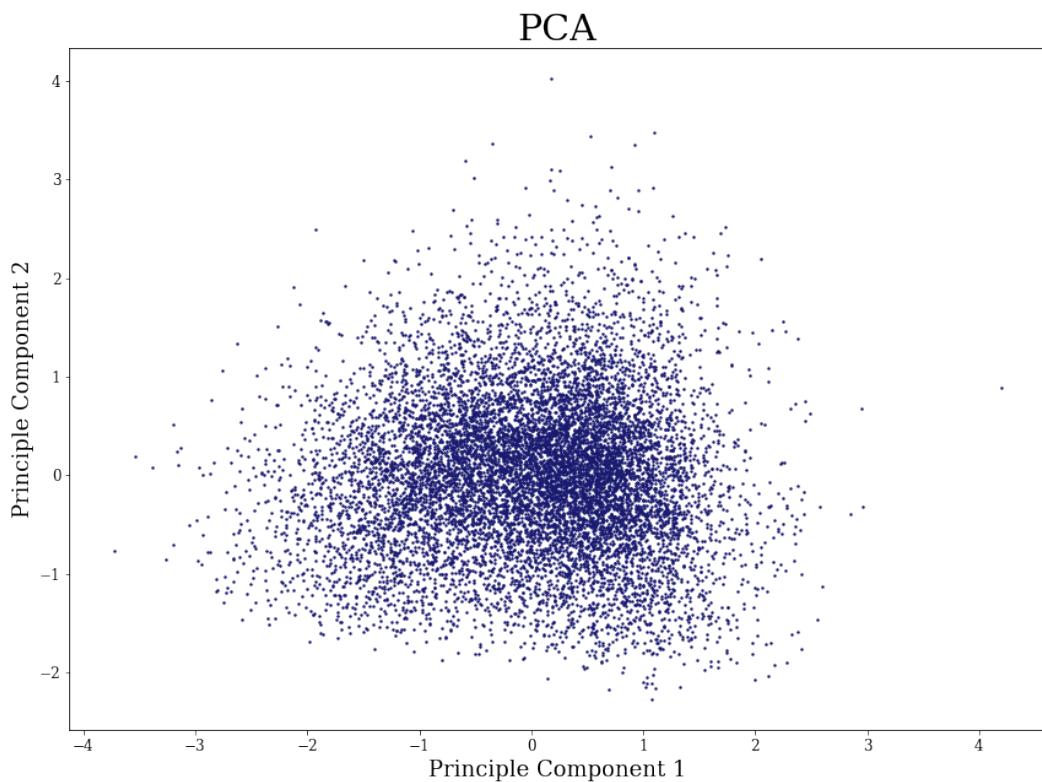
Jika dibandingkan dari nilai *cosine similarity* masing-masing klaster dalam Gambar IV.43, terlihat bahwa klaster 1 memiliki nilai *cosine similarity* yang paling tinggi dibandingkan klaster-klaster lainnya. Hal ini menunjukkan bahwa galaksi dalam klaster 1 lebih seragam dibandingkan galaksi pada klaster lainnya. Sementara itu, klaster 3 memiliki nilai *cosine similarity* yang paling rendah dan mendekati 0, yang berarti galaksi dalam klaster 3 sangat beragam dan tidak memiliki karakteristik khusus.



Gambar IV.43: *Heatmap cosine similarity* antar klaster dari metode *K-Means Clustering*.

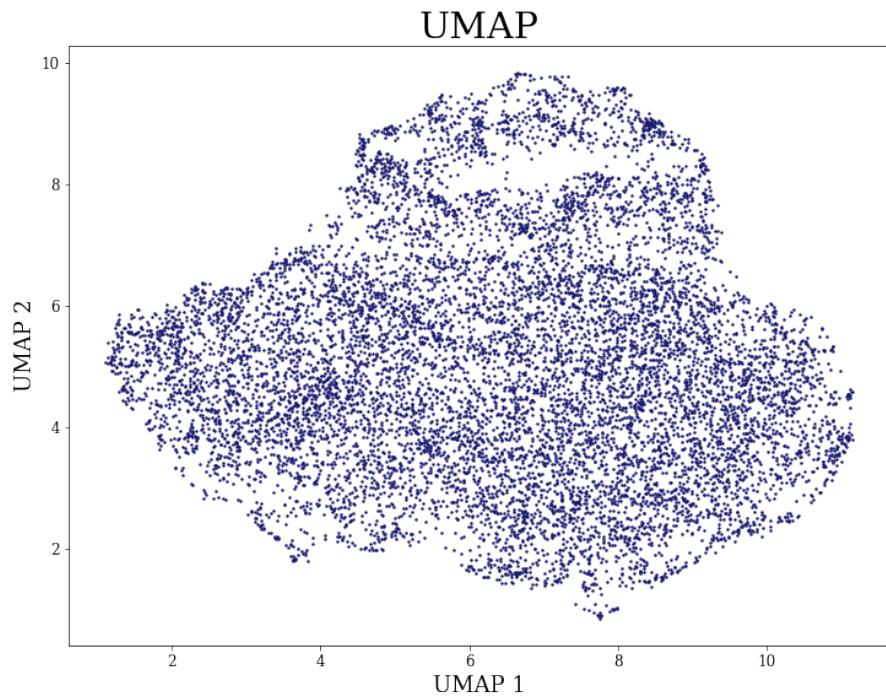
PCA dan UMAP

Selain dilihat berdasarkan sampel galaksi masing-masing klaster, pengelompokan galaksi juga dapat dilihat dengan mereduksi dimensi parameter laten menjadi dimensi yang lebih rendah, lalu melihat ketersebaran data. Reduksi dimensi dilakukan dengan menggunakan metode PCA dan UMAP. Hasil reduksi dimensi ini akan ditampilkan dalam bentuk plot dua dimensi, dan akan menjadi bahan analisis selanjutnya.



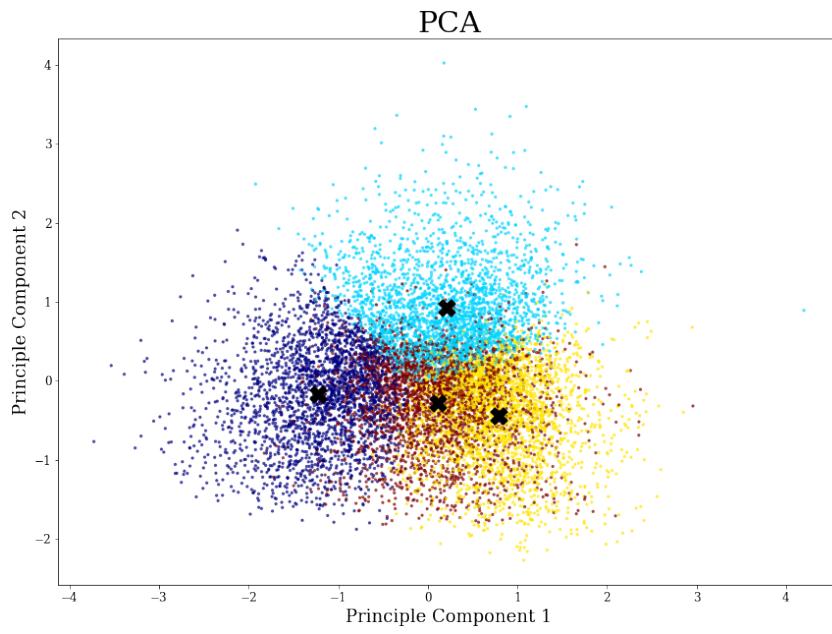
Gambar IV.44: Proyeksi parameter laten dalam dua dimensi menggunakan PCA.

Gambar IV.44 menunjukkan proyeksi parameter laten sepuluh dimensi menjadi hanya dua dimensi dengan menggunakan metode PCA. Berdasarkan gambar tersebut sebaran data hanya menunjukkan bahwa hanya terdapat satu klaster, dan tidak tampak adanya pengelompokan terhadap data galaksi jauh. Sementara itu, Gambar IV.45 menunjukkan proyeksi parameter laten menjadi dua dimensi menggunakan metode UMAP. Dengan metode ini, tampak persebaran data tidak merata meski pengelompokan pun juga tidak tampak jelas.

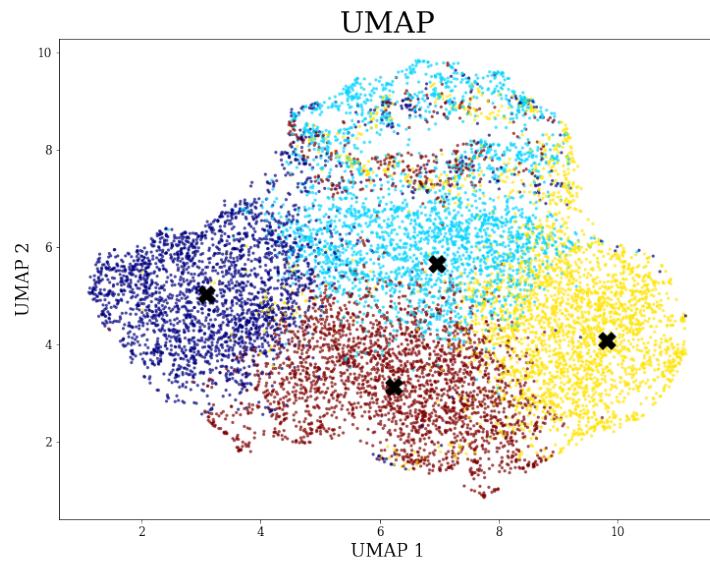


Gambar IV.45: Proyeksi parameter laten dalam dua dimensi menggunakan UMAP.

Proyeksi PCA dan UMAP yang ditunjukkan pada Gambar IV.44 dan IV.45 juga dapat menunjukkan distribusi klaster hasil pengelompokan dengan metode *k-means* yang sebelumnya dibahas pada Subbab IV.3.1. Posisi *centroid* keempat klaster akan ditunjukkan dalam plot proyeksi parameter laten. Distribusi galaksi setiap klaster dapat dilihat dalam plot proyeksi PCA dan UMAP, yang ditunjukkan pada Gambar IV.46.



(a)



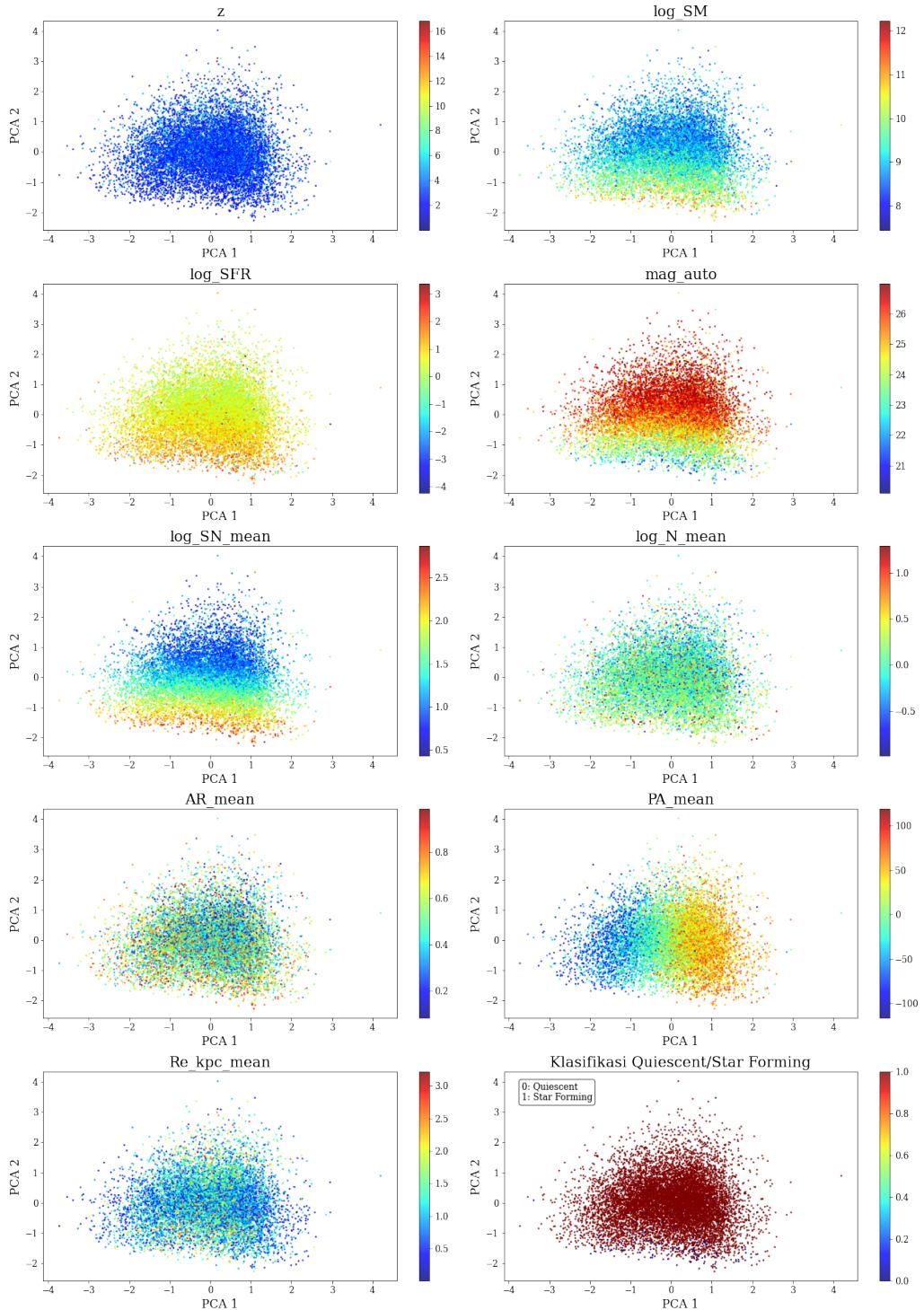
(b)

Gambar IV.46: Hasil klastering dengan metode *K-Means* yang tampak dari proyeksi parameter laten dengan metode PCA (panel a) dan dengan metode UMAP (panel b).

Selain dilihat berdasarkan hasil pengelompokan dengan metode *k-means*, proyeksi parameter laten juga dapat dilihat dengan membandingkannya terhadap parameter galaksi dari katalog dan juga dari hasil *fitting Galfitm*. Hal ini memberi gambaran informasi apa saja yang berhasil dipelajari mesin melalui

data citra galaksi.

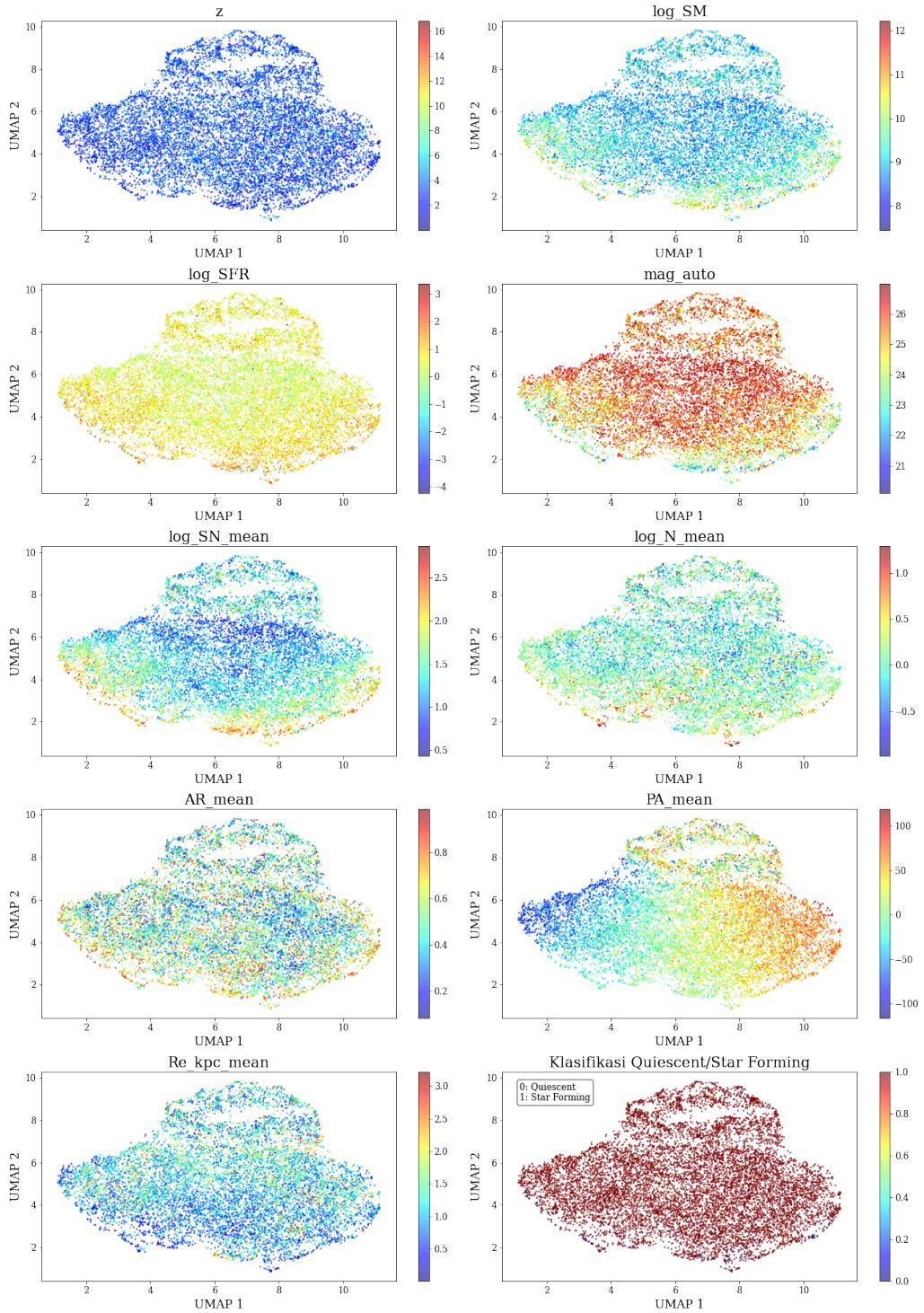
Gambar IV.47 menunjukkan plot hubungan beberapa parameter galaksi terhadap proyeksi parameter laten menggunakan PCA. Dari 10 parameter, terlihat korelasi antara parameter massa, SFR, magnitudo, S/N, dan *position angle*. Sementara itu, parameter lainnya tampak terdistribusi acak dalam proyeksi parameter laten. Untuk parameter-parameter yang terlihat berkorelasi, hal tersebut menunjukkan bahwa arsitektur *encoder* yang dibangun berhasil menyimpan informasi parameter-parameter tersebut, sehingga dalam proyeksinya tampak adanya korelasi. Namun, jika dibandingkan dengan proyeksi parameter laten dari Gambar IV.46 panel (a), terlihat bahwa pengelompokan dengan metode *k-means* tidak mengelompokkan berdasarkan salah satu parameter galaksi. Karena jika pengelompokan berdasarkan salah satu atau beberapa parameter, maka seharusnya akan muncul pola yang sama seperti yang muncul pada Gambar IV.46.



Gambar IV.47: Hubungan beberapa parameter galaksi terhadap proyeksi parameter laten dalam dua dimensi dengan PCA

Sementara itu, Gambar IV.48 menunjukkan hubungan antara parameter galaksi terhadap proyeksi UMAP. Secara umum, hubungan yang terlihat tidak jauh berbeda dengan hubungan yang terlihat dari Gambar IV.47. Beberapa

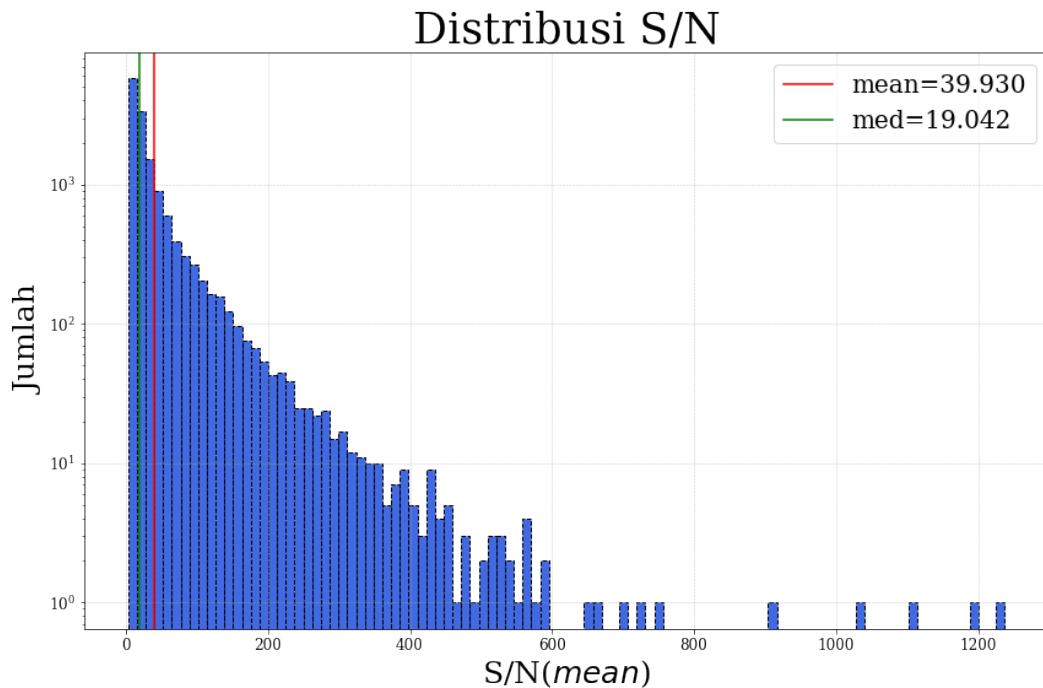
parameter seperti *redshift*, *axis-ratio* dan indeks *Sérsic* tersebar acak dalam proyeksi parameter laten. Hal ini menunjukkan bahwa parameter tersebut tidak direpresentasikan oleh parameter laten dengan cukup signifikan. Sama sebelumnya, tidak tampak adanya pola pengelompokan seperti pada Gambar IV.46 panel (b).



Gambar IV.48: Hubungan beberapa parameter galaksi terhadap proyeksi parameter laten dalam dua dimensi dengan UMAP

IV.3.2 Pengelompokan Galaksi Jauh pada Berbagai Rentang *Redshift*

Hasil pengelompokan dari Subbab IV.3.1 menunjukkan bahwa terdapat kecenderungan pengelompokan galaksi dengan *machine learning* memisahkan data galaksi berdasarkan *noise* yang tinggi dan rendah. Salah satu cara untuk mengatasi hal ini adalah dengan membatasi nilai S/N, sehingga data galaksi memiliki *noise* yang relatif lebih homogen. Untuk menentukan batas nilai S/N yang dipilih, mula-mula akan dilihat distribusi nilai S/N, sebagaimana ditunjukkan pada Gambar IV.49. Untuk menjaga kehomogenan data berdasarkan *noise*-nya, S/N dibatasi pada rentang $S/N > 20$, karena itu berarti bahwa dipilih 50% data dengan S/N yang paling besar, karena nilai median dari distribusi tampak di sekitar nilai tersebut.

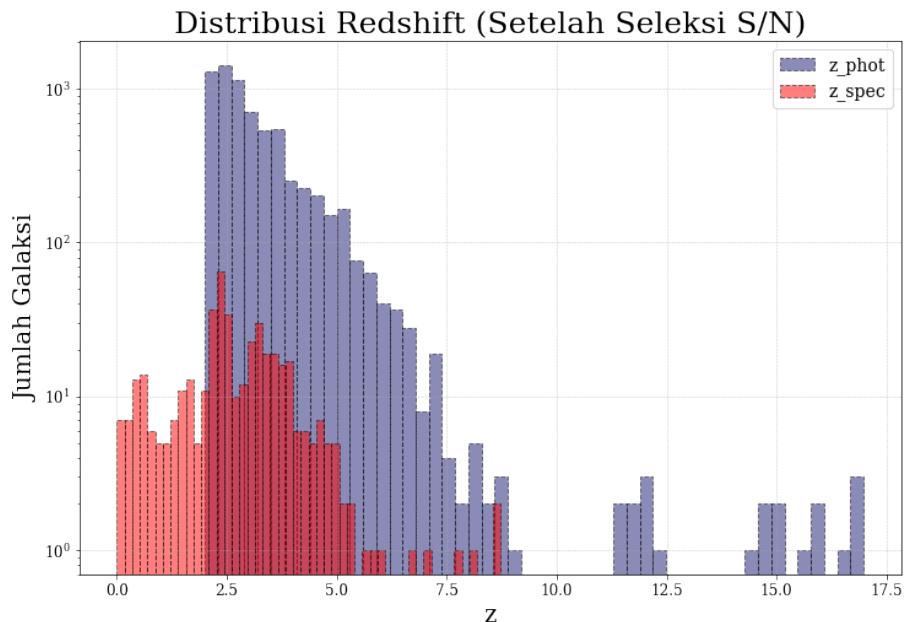


Gambar IV.49: Distribusi nilai rata-rata *signal-to-noise* dari data tiga filter.

Proses *resizing* dan *rescaling* yang dilakukan pada Subbab IV.3.1 berpotensi menghilangkan informasi riil yang terkandung di dalam data galaksi pada rentang *redshift* yang lebar, terutama informasi ukuran setiap galaksi. Untuk itu, salah satu cara untuk mengatasi hal tersebut adalah dengan membagi data ke dalam beberapa *bin redshift* pada rentang yang lebih kecil, kemudian pengelompokan galaksi dilakukan untuk setiap *bin* tersebut. Galaksi-galaksi

pada rentang *redshift* yang lebih kecil seharusnya memiliki ukuran yang lebih seragam dibandingkan galaksi pada rentang *redshift* yang besar.

Dalam penelitian ini, pemilihan galaksi jauh dilakukan berdasarkan nilai *redshift* fotometrik dari katalog *EAZY*. Meski demikian, di dalam katalog tersebut terdapat informasi parameter *redshift* spektroskopik, tetapi jumlahnya jauh lebih sedikit dibandingkan *redshift* fotometrik. Namun, pengukuran *redshift* dengan metode spektroskopi memberikan nilai yang lebih akurat dibandingkan pengukuran dengan metode fotometrik. Oleh karena itu, pada Subbab ini, pembagian galaksi ke dalam beberapa *bin redshift* dilakukan berdasarkan nilai *redshift* spektroskopi. Gambar IV.50 menunjukkan distribusi *redshift* galaksi setelah membatasi S/N > 20.

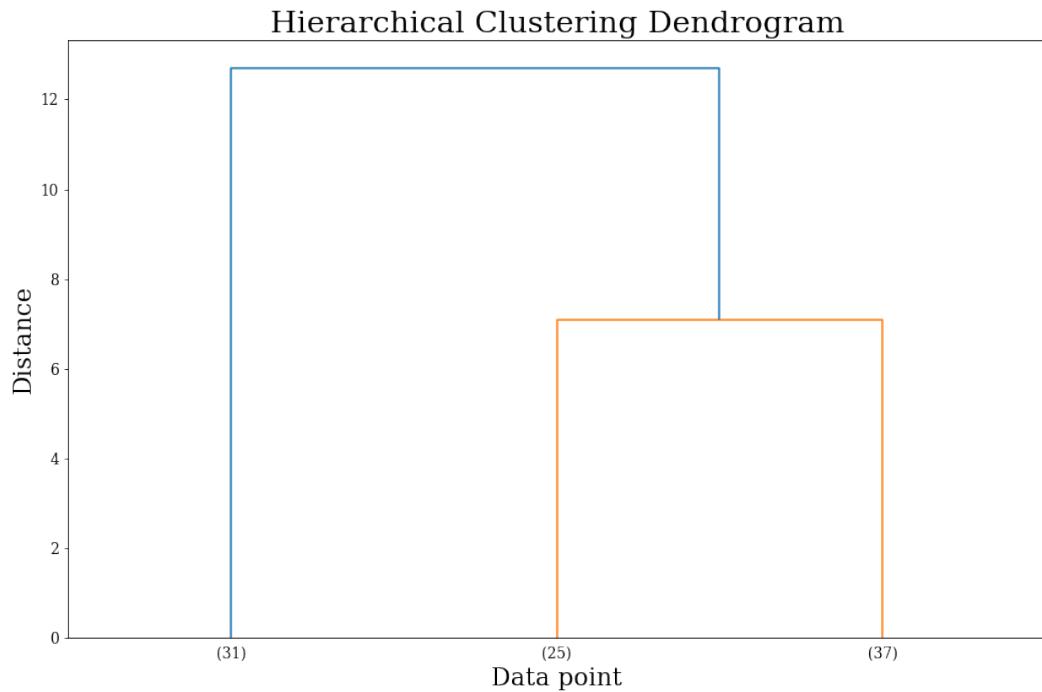


Gambar IV.50: Distribusi *redshift* galaksi setelah dilakukan seleksi S/N > 20.

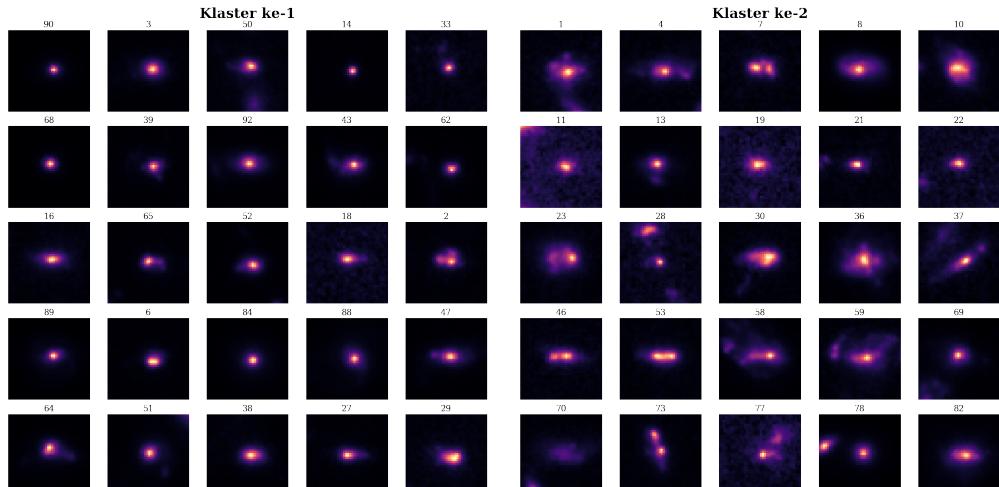
Berdasarkan distribusi *redshift* yang ditunjukkan pada Gambar IV.50, tampak bahwa sebagian galaksi yang telah diseleksi pada $z_{phot} > 2$ berada pada *redshift* yang lebih rendah ($z_{spec} < 2$). Meski demikian, galaksi-galaksi tersebut akan tetap digunakan pada analisis ini. Data galaksi akan dibagi ke dalam 4 rentang *redshift*, yaitu $z_{spec} \leq 2$, $2 < z_{spec} \leq 3$, $3 < z_{spec} \leq 4$, $z_{spec} > 4$.

Bin Pertama ($z_{spec} \leq 2$)

Dataset pada *bin* pertama berisi 93 galaksi, yang dapat dikelompokan berdasarkan dendrogram pada Gambar IV.51. Sampel galaksi dari masing-masing klaster ditunjukkan pada Gambar IV.52.

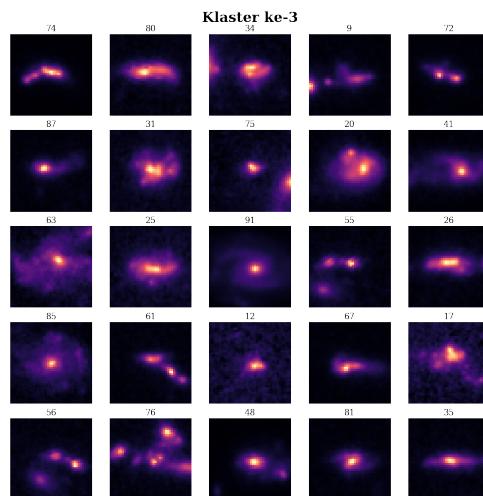


Gambar IV.51: Dendrogram galaksi pada rentang $z_{spec} \leq 2$.



(a)

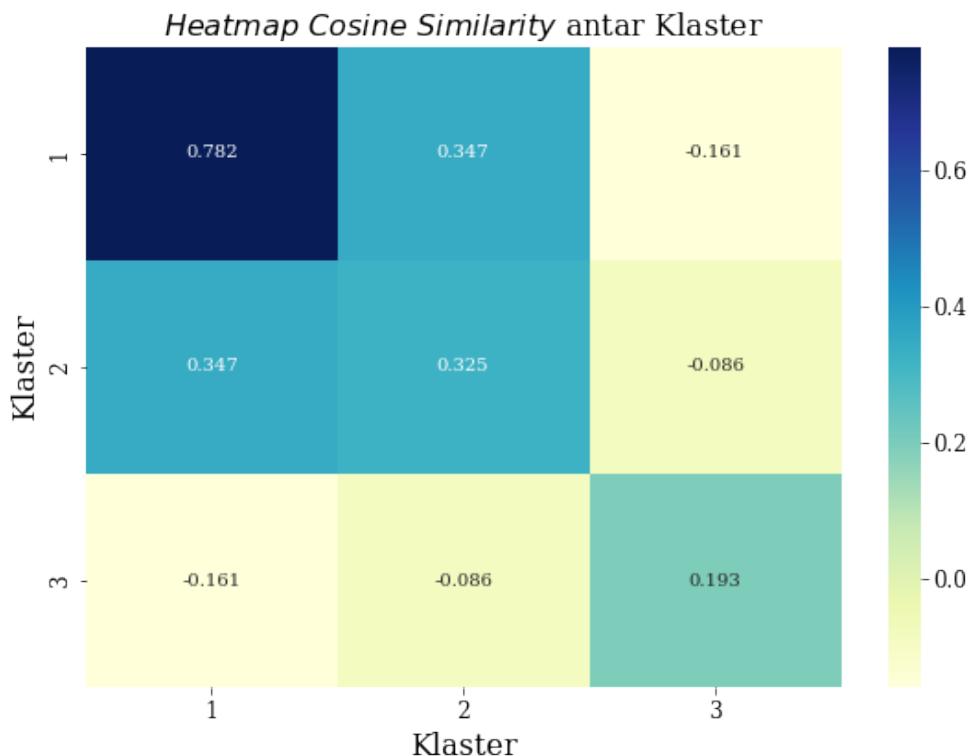
(b)



(c)

Gambar IV.52: Sampel galaksi dari masing-masing klaster pada rentang $z_{spec} \leq 2$.

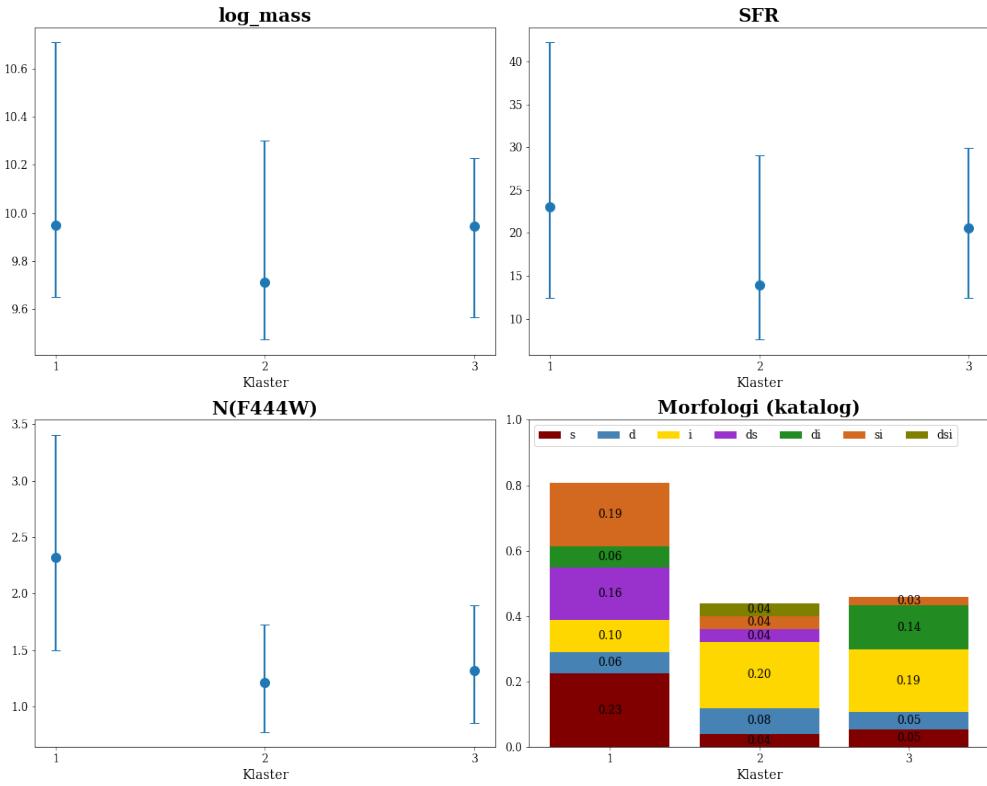
Secara visual, terlihat bahwa klaster pertama berisi galaksi-galaksi yang memiliki bentuk *spheroidal* sederhana dengan ukuran yang relatif cukup kecil. Sementara itu, klaster kedua dan ketiga tampak sebagai galaksi-galaksi dengan struktur yang lebih acak dan memiliki struktur *clump*. Kemiripan antarklaster dapat dilihat dari nilai *cosine similarity* yang ditampilkan dalam bentuk *heatmap* pada Gambar IV.53.



Gambar IV.53: *Heatmap cosine similarity* antarklaster untuk galaksi pada rentang $z_{spec} \leq 2$.

Dalam *heatmap* tersebut, nilai *cosine similarity* untuk klaster yang sama menunjukkan nilai kemiripan galaksi-galaksi dalam satu klaster. Sementara itu, nilai *cosine similarity* untuk klaster yang berbeda menunjukkan nilai kemiripan antara kedua klaster.

Berdasarkan Gambar IV.53, terlihat bahwa nilai kemiripan untuk galaksi-galaksi pada klaster pertama memiliki nilai > 0.78 , yang berarti galaksi-galaksi dalam klaster pertama sangat mirip satu sama lain. Sementara itu, nilai *cosine similarity* untuk klaster kedua dan ketiga cukup rendah, yakni di angka ~ 0.3 dan ~ 0.1 . Hal ini menunjukkan bahwa galaksi-galaksi pada kedua klaster tersebut memiliki bentuk yang beragam.



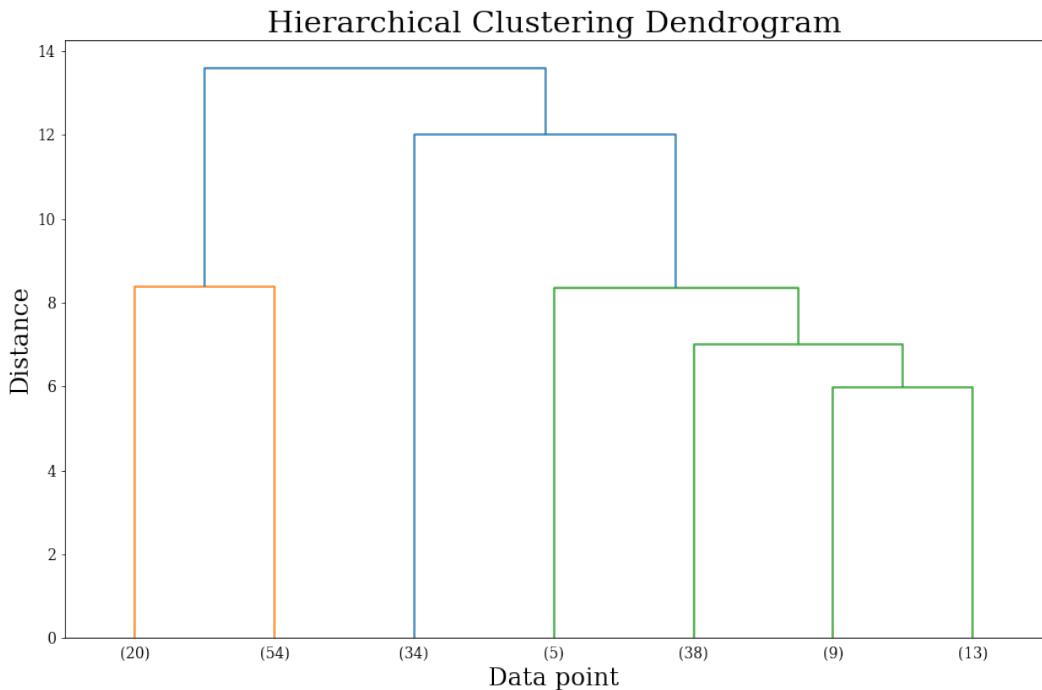
Gambar IV.54: Distribusi beberapa parameter galaksi terhadap klaster pada rentang $z_{spec} \leq 2$.

Distribusi beberapa parameter galaksi untuk masing-masing klaster ditunjukkan pada Gambar IV.54. Terdapat empat parameter yang akan dilihat distribusinya, yaitu massa, SFR, indeks Sérsic pada filter F444W, dan kategori morfologi galaksi melalui inspeksi visual dari penelitian Effendi (2024). Secara umum, massa ketiga klaster cukup seragam dari nilai mediannya. Hal yang sama juga tampak untuk parameter SFR. Indeks Sérsic untuk klaster pertama relatif lebih tinggi dibandingkan kedua klaster lainnya. Sementara itu, dilihat dari sebaran morfologi galaksi dari inspeksi visual, klaster pertama tampak didominasi dengan tipe galaksi *spheroid*, baik *spheroid* murni maupun *spheroid+disk* dan *spheroid+irregular*. Sementara itu, galaksi di klaster kedua dan ketiga didominasi dengan tipe galaksi *irregular*.

Bin Kedua ($2 < z_{spec} \leq 3$)

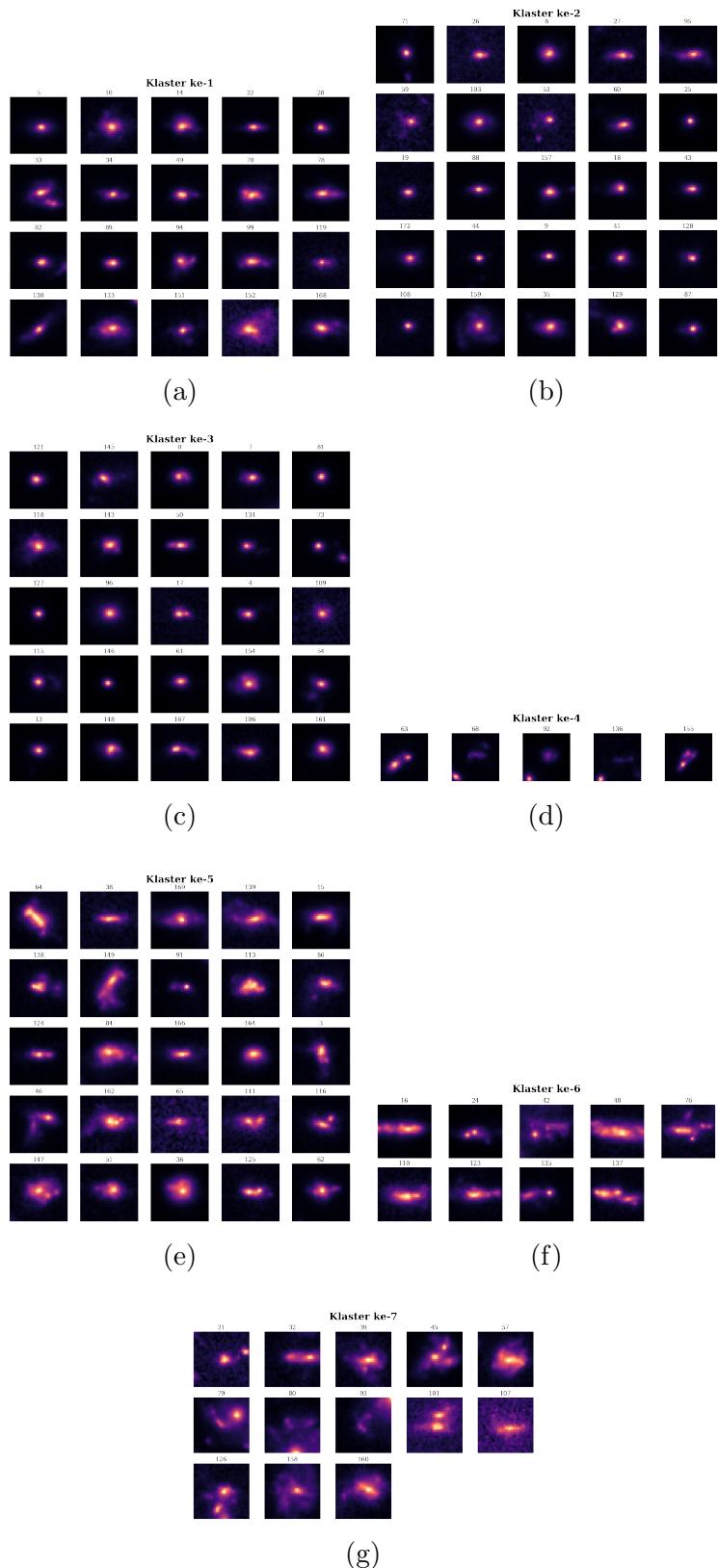
Dalam *bin* kedua terdapat 173 galaksi, dan menjadi *bin redshift* yang berisi paling banyak galaksi. Galaksi dalam *bin* kedua ini dapat dikelompokkan berdasarkan dendrogram pada Gambar IV.55. Karena terdapat lebih banyak galaksi pada *bin* kedua, maka pada *bin* ini akan dilihat sampel galaksi apabila

dikelompokkan ke dalam 7 klaster.



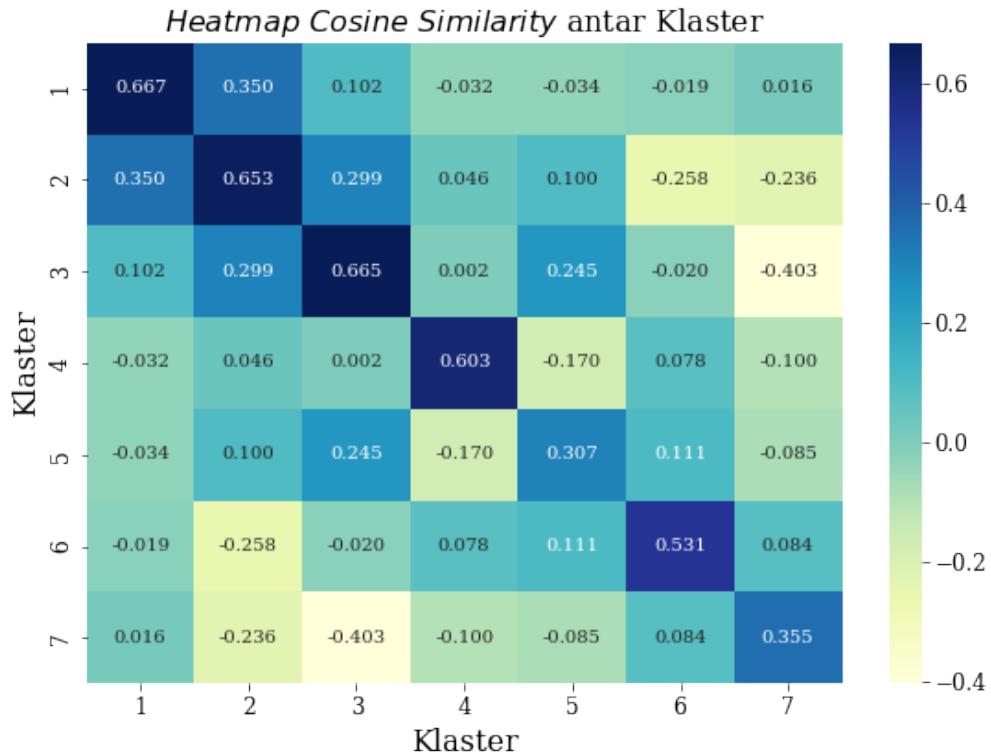
Gambar IV.55: Dendrogram galaksi pada rentang $2 < z_{spec} \leq 3$.

Gambar IV.56 menunjukkan sampel galaksi pada masing-masing klaster, untuk pengelompokan galaksi pada rentang $2 < z_{spec} \leq 3$. Secara visual, tampak klaster kelima hingga klaster ketujuh merupakan klaster yang memiliki struktur *clump* atau galaksi yang berinteraksi. Klaster keempat tampak merupakan klaster berisi galaksi yang redup, dan berada dekat dengan objek terang didekatnya. Sementara itu, klaster-klaster lainnya tampak memiliki bentuk yang *spheroid* yang mirip satu sama lain.

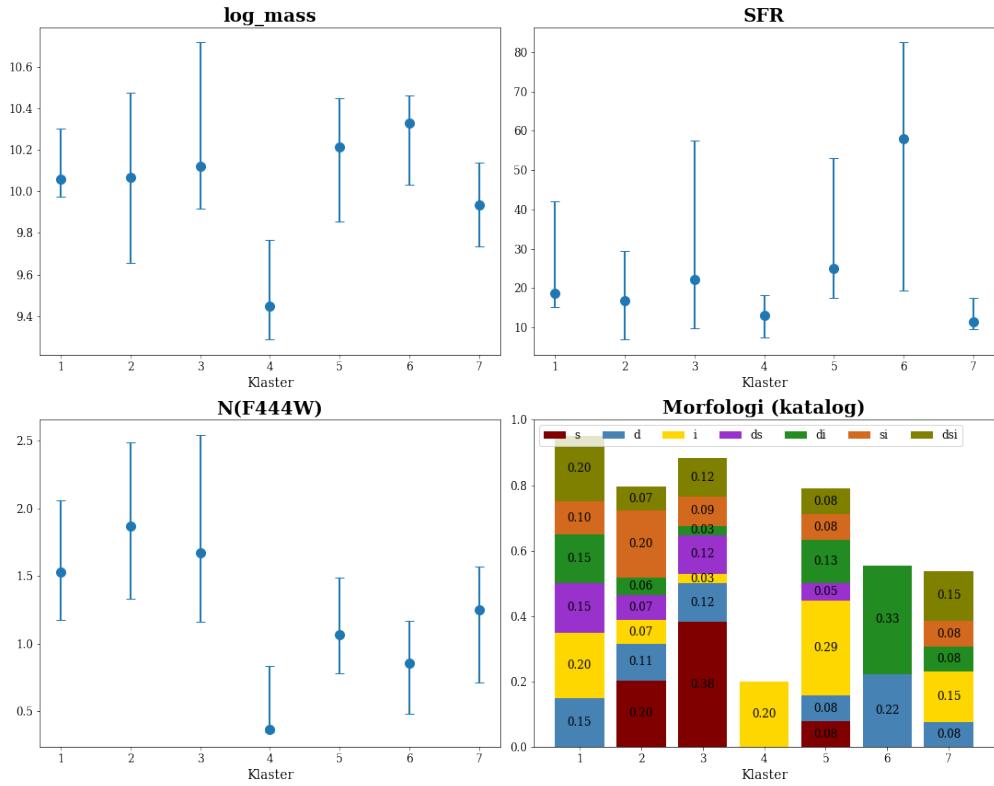


Gambar IV.56: Sampel galaksi dari masing-masing klaster pada rentang $2 < z_{spec} \leq 3$.

Jika dilihat dari nilai *cosine similarity* antarklaster seperti ditunjukkan dalam Gambar IV.57, terlihat bahwa klaster pertama memiliki nilai kemiripan yang cukup besar, namun masih tidak sebesar klaster pertama untuk *bin* sebelumnya. Hal ini dapat disebabkan karena jumlah galaksi dalam klaster ke-1 yang lebih sedikit, sehingga relatif lebih seragam antar satu sama lain.



Gambar IV.57: *Heatmap cosine similarity* antarklaster untuk galaksi pada rentang $2 < z_{spec} \leq 3$.

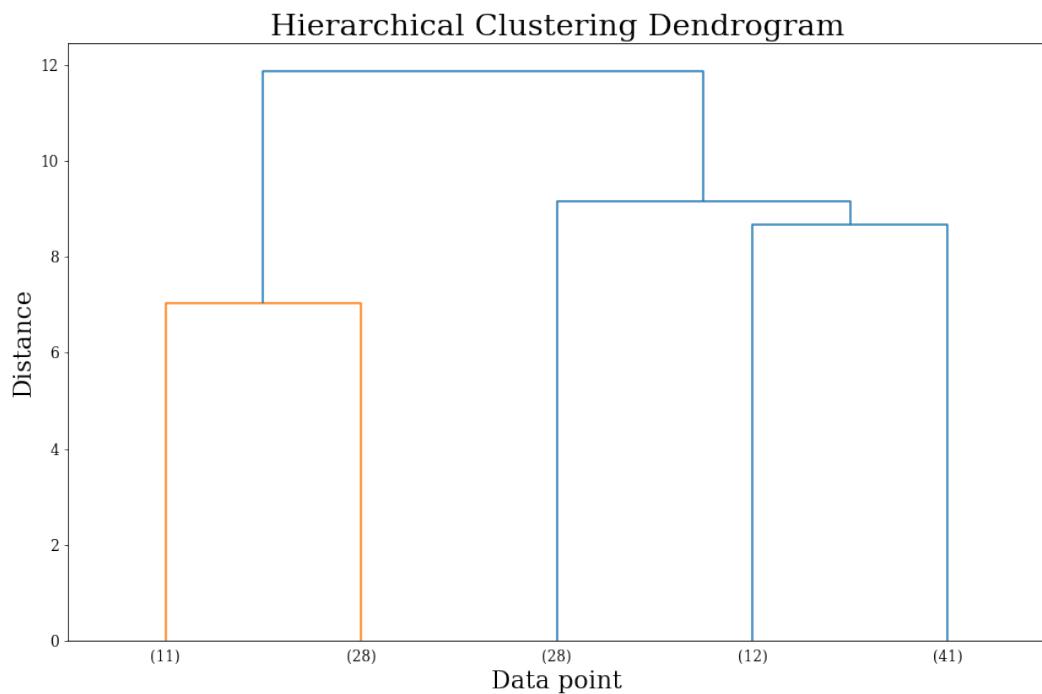


Gambar IV.58: Distribusi beberapa parameter galaksi terhadap klaster pada rentang $2 < z_{spec} \leq 3$.

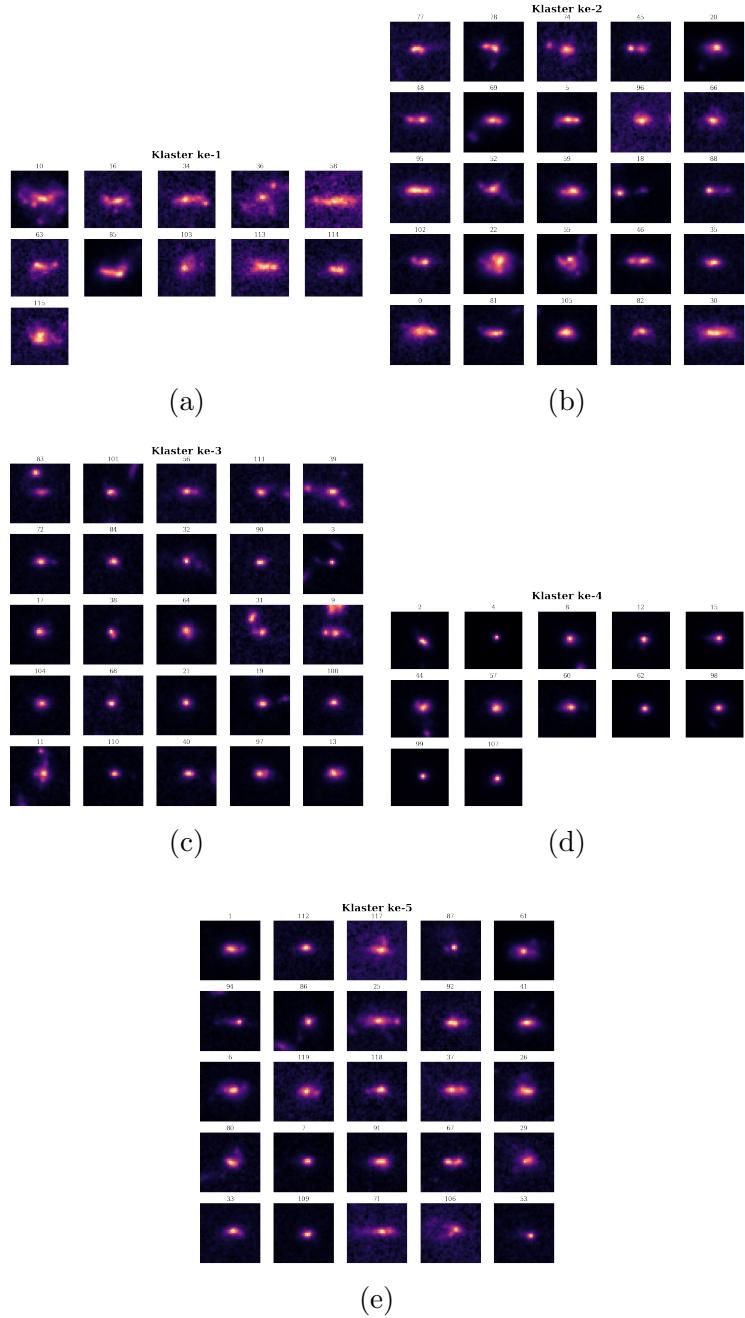
Distribusi beberapa parameter galaksi untuk masing-masing klaster ditunjukkan pada Gambar IV.58. Secara umum, klaster keempat tampak berisi galaksi dengan massa yang paling kecil dibandingkan klaster lainnya. Nilai SFR pada klaster keempat juga cukup rendah. Indeks Sérsic untuk galaksi pada klaster keempat pun paling rendah dibandingkan klaster lainnya. Jika dibandingkan dengan kategori morfologi dari inspeksi visual, klaster ketiga didominasi dengan galaksi *spheroid*, sementara itu galaksi di klaster kelima hingga ketujuh tampak didominasi dengan tipe galaksi *irregular* dan *disk+irregular*.

Bin Ketiga ($3 < z_{spec} \leq 4$)

Terdapat 120 galaksi dalam dataset pada rentang $3 < z_{spec} \leq 4$, dengan dendrogram seperti ditunjukkan pada Gambar IV.59. Sampel galaksi dari masing-masing klaster ditunjukkan dalam Gambar IV.60.

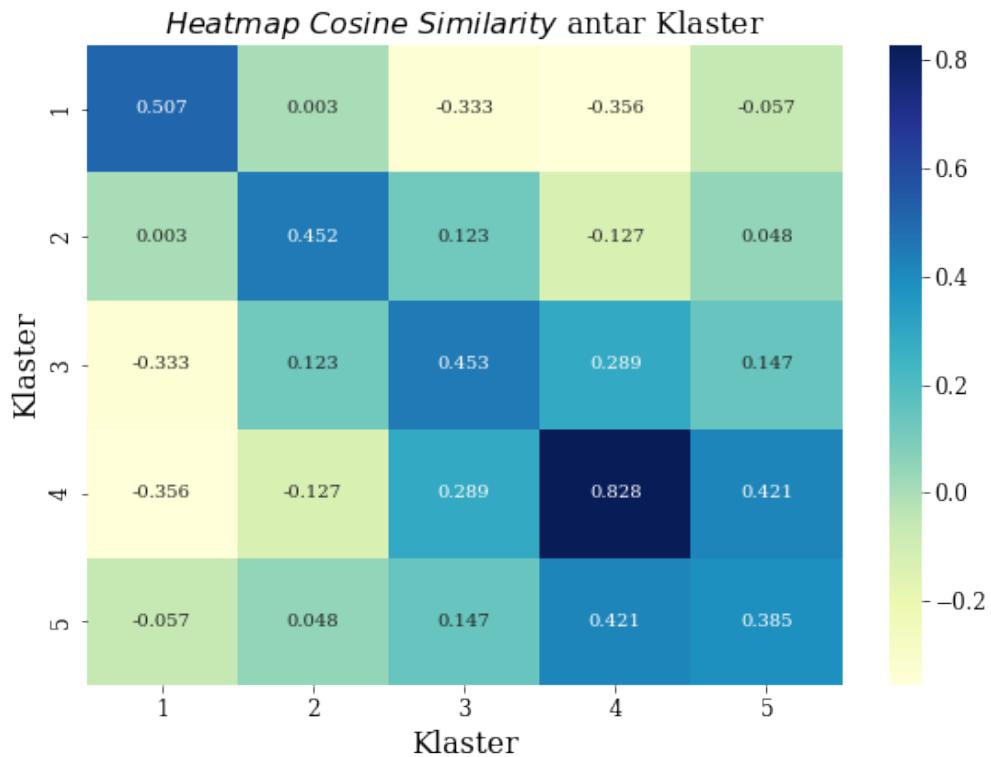


Gambar IV.59: Dendrogram galaksi pada rentang $3 < z_{spec} \leq 4$.



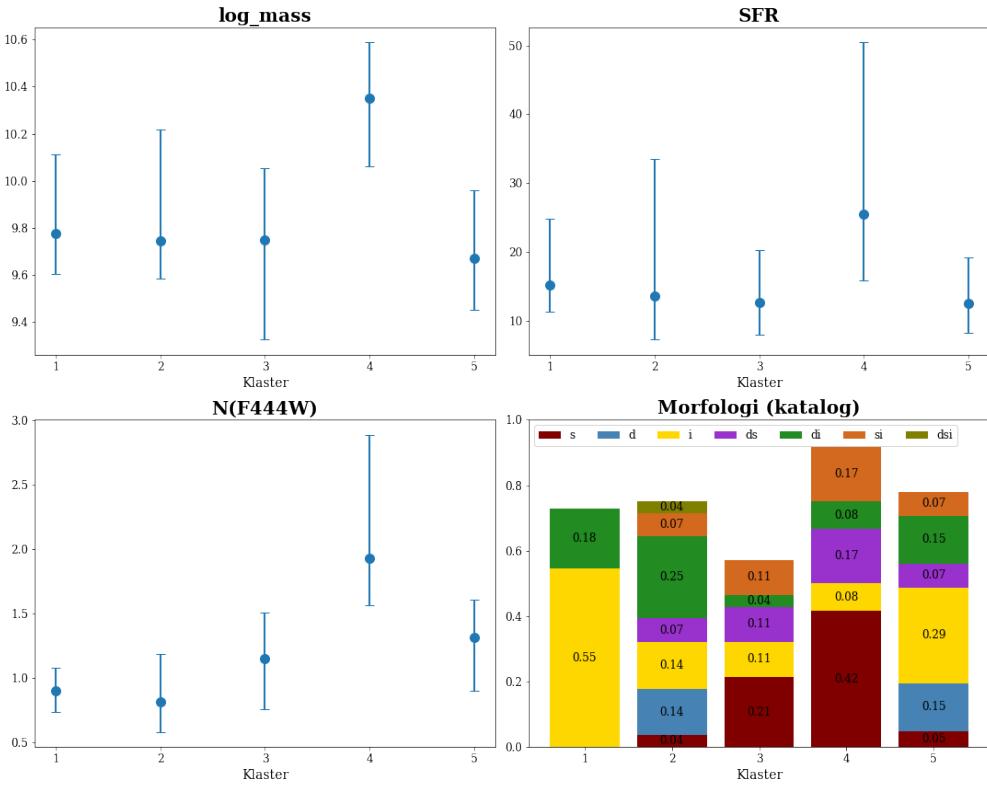
Gambar IV.60: Sampel galaksi dari masing-masing klaster pada rentang $3 < z_{spec} \leq 4$.

Secara visual, klaster pertama dan kedua tampak berisi galaksi-galaksi *clumpy*. Sementara itu klaster ketiga dan keempat merupakan klaster berisi galaksi-galaksi *spheroid* dengan diameter sudut yang tidak terlalu besar. Klaster kelima tampak berisi galaksi dengan bentuk yang beragam dan tampaknya tidak memiliki karakteristik khusus.



Gambar IV.61: *Heatmap cosine similarity* antarklaster untuk galaksi pada rentang $3 < z_{spec} \leq 4$.

Nilai kemiripan antarklaster untuk *bin* ketiga ditunjukkan dalam Gambar IV.61. Klaster keempat memiliki nilai kemiripan yang paling besar, yang berarti galaksi-galaksi dalam klaster keempat cukup seragam.



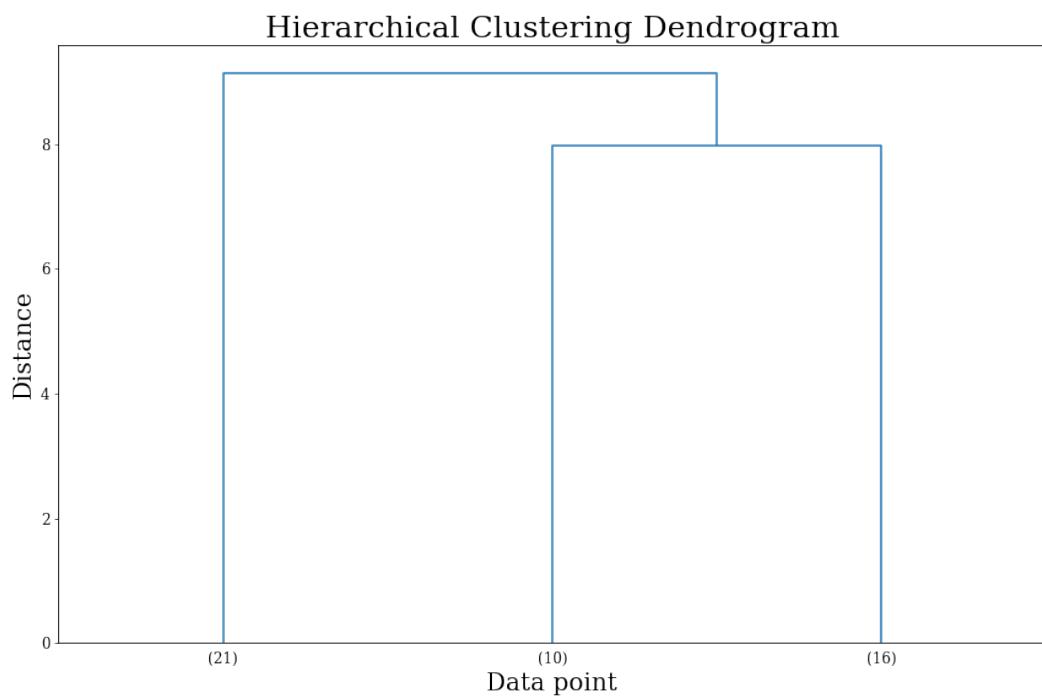
Gambar IV.62: Distribusi beberapa parameter galaksi terhadap klaster pada rentang $3 < z_{spec} \leq 4$.

Distribusi beberapa parameter galaksi untuk masing-masing klaster ditunjukkan pada Gambar IV.62. Secara umum, klaster keempat tampak berisi galaksi dengan massa dan SFR yang paling tinggi. Galaksi pada klaster keempat juga memiliki indeks Sérsic yang paling tinggi dibandingkan klaster lainnya. Jika dibandingkan dengan informasi morfologi galaksi melalui inspeksi visual, klaster keempat tampak didominasi dengan tipe galaksi *spheroid*. Sementara itu, klaster pertama tampak didominasi oleh galaksi tipe *irregular*.

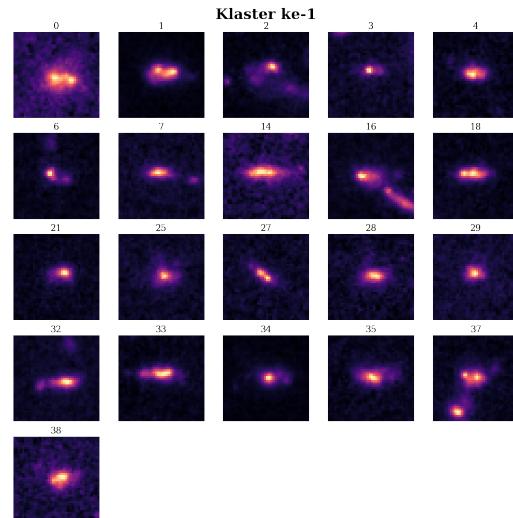
Bin Keempat ($z_{spec} > 4$)

Data pada *bin* keempat menjadi data dengan jumlah galaksi yang paling sedikit dibandingkan *bin redshift* lainnya, yakni hanya sebanyak 47 galaksi. Di dalam *bin* ini, galaksi pada *redshift* tertinggi berada pada $z_{spec} = 8.71$.

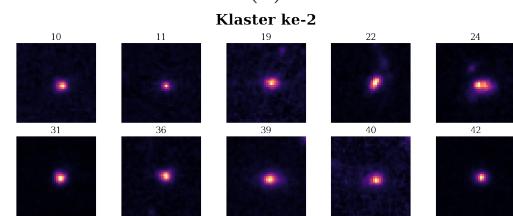
Pengelompokan untuk galaksi pada *bin* keempat ini dapat dilihat dari dendrogram pada Gambar IV.63. Pada *bin* keempat, galaksi akan dikelompokan ke dalam 3 klaster, dan sampel galaksi dari masing-masing klaster ditunjukkan pada Gambar IV.64.



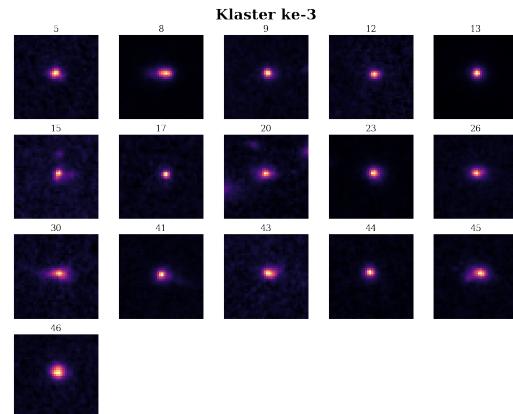
Gambar IV.63: Dendrogram galaksi pada rentang $z_{spec} > 4$.



(a)



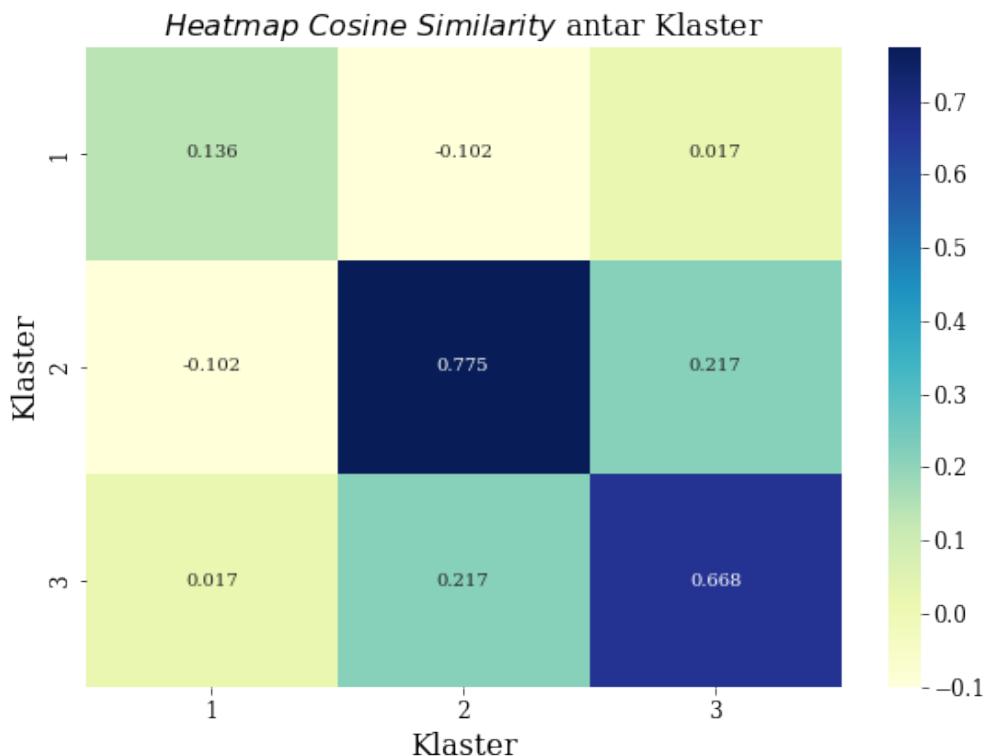
(b)



(c)

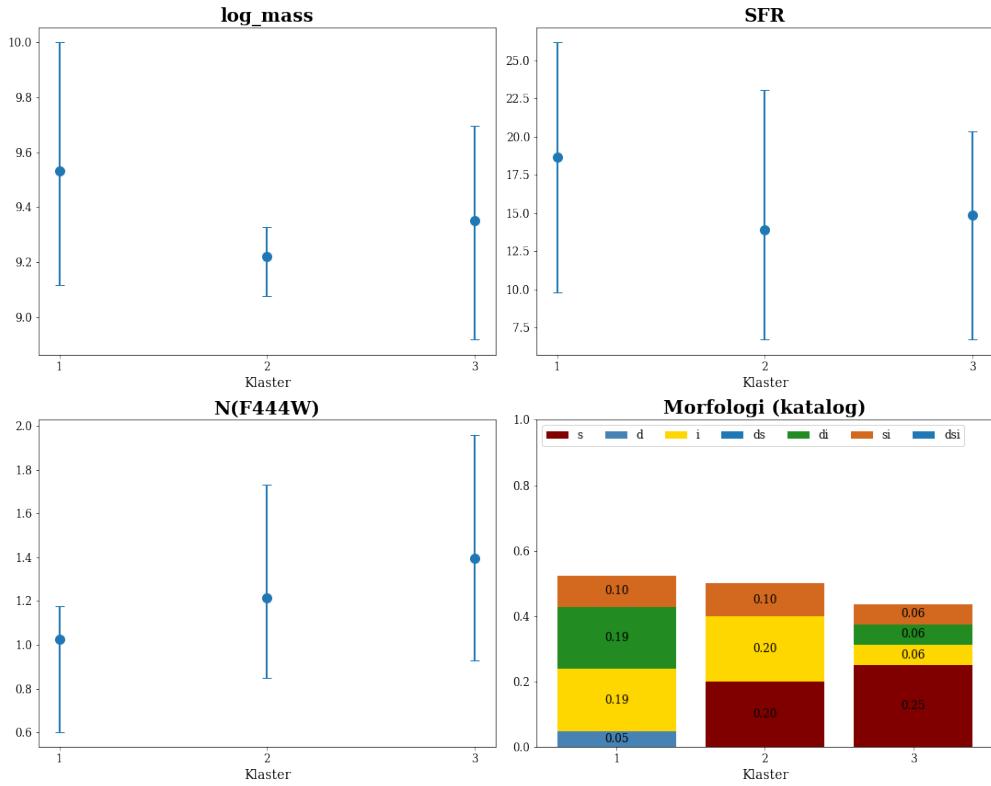
Gambar IV.64: Sampel galaksi dari masing-masing klaster pada rentang $z_{spec} > 4$.

Secara visual, klaster kedua dan klaster ketiga tampak berisi galaksi dengan bentuk *spheroid*, sementara klaster pertama didominasi oleh galaksi *irregular*, *clumpy*, atau galaksi yang berinteraksi. Meski demikian, terlihat beberapa galaksi dengan bentuk *spheroid* yang juga masuk ke dalam klaster pertama.



Gambar IV.65: *Heatmap cosine similarity* antarklaster untuk galaksi pada rentang $z_{spec} > 4$.

Jika dilihat dari *heatmap* nilai *cosine similarity* antarklaster pada Gambar IV.65, terlihat bahwa galaksi-galaksi pada klaster kedua cukup seragam karena memiliki nilai kemiripan yang cukup besar. Klaster ketiga juga memiliki nilai kemiripan yang besar, yang menunjukkan bahwa klaster ketiga juga berisi galaksi-galaksi dengan fitur yang seragam. Klaster pertama memiliki nilai *similarity cosine* yang rendah, sehingga mendukung penampakan visual galaksi-galaksi dengan bentuk yang acak dan *clumpy*.



Gambar IV.66: Distribusi beberapa parameter galaksi terhadap klaster pada rentang $z_{spec} > 4$.

Distribusi beberapa parameter galaksi untuk masing-masing klaster ditunjukkan pada Gambar IV.62. Secara umum, tidak ada perbedaan yang signifikan antara distribusi parameter galaksi untuk ketiga klaster. Klaster kedua dan ketiga tampak memiliki indeks *Sérsic* yang paling tinggi dibandingkan klaster pertama. Dilihat dari sebaran morfologi galaksi dari inspeksi visual, galaksi di klaster ketiga didominasi dengan tipe galaksi *spheroid*.

Analisis Terhadap *Redshift*

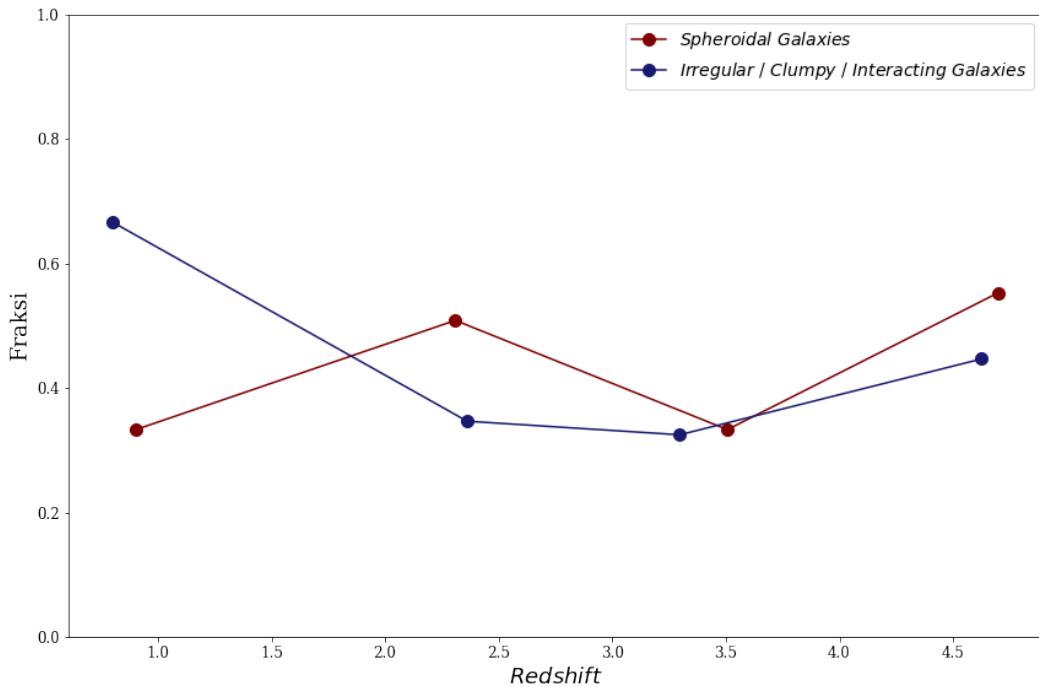
Hasil pengelompokan galaksi dalam empat rentang *redshift* menunjukkan bahwa secara umum, terdapat klaster galaksi yang didominasi dengan galaksi dengan bentuk *spheroid*, serta klaster galaksi yang tampak berisi galaksi dengan bentuk *irregular* atau galaksi yang berinteraksi. Pada bagian ini akan dilakukan analisis terhadap parameter galaksi dari kategori galaksi *spheroid* dan *irregular* di berbagai rentang *redshift*. Dalam analisis ini, digunakan asumsi bahwa galaksi *spheroid* dan *irregular* berada dalam klaster-klaster sebagaimana ditunjukkan pada Tabel IV.1. Klaster yang tidak menunjukkan karakteristik khusus dan tampak berisi gabungan galaksi dengan bentuk *spher-*

orid dan *irregular* tidak dikategorikan dalam analisis ini, misalnya klaster 1 dan klaster 4 di *bin redshift* $2 < z_{spec} \leq 3$ serta klaster 3 dan klaster 5 di *bin redshift* $3 < z_{spec} \leq 4$.

Tabel IV.1: Asumsi klaster-klaster yang berisi galaksi tipe *spheroid* dan *irregular*.

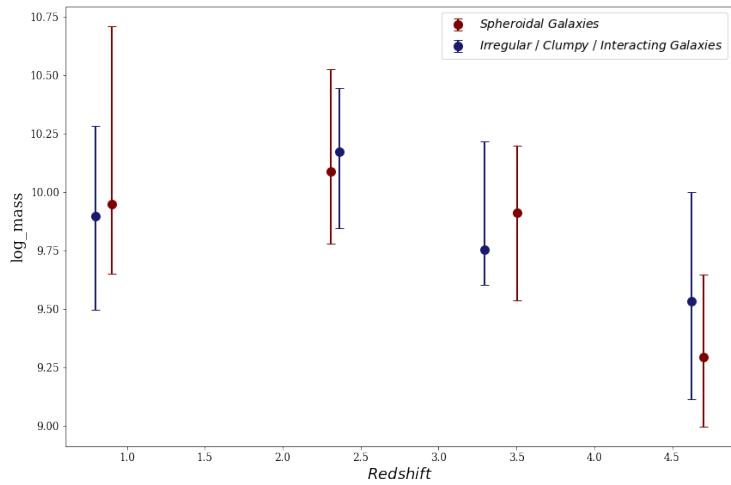
Bin redshift	Galaksi <i>Spheroid</i>	Galaksi <i>Irregular</i>
$z_{spec} \leq 2$	Klaster 1	Klaster 2, Klaster 3
$2 < z_{spec} \leq 3$	Klaster 2, Klaster 3	Klaster 5, Klaster 6, Klaster 7
$3 < z_{spec} \leq 4$	Klaster 3, Klaster 4	Klaster 1, Klaster 2
$z_{spec} > 4$	Klaster 2, Klaster 3	Klaster 1

Jika dibandingkan jumlah galaksi *spheroid* dan *irregular* berdasarkan asumsi hasil pengelompokan yang ditunjukkan pada Tabel IV.1, maka distribusi jumlah galaksi *spheroid* dan *irregular* di berbagai *redshift* dapat dilihat dalam Gambar IV.67. Berdasarkan Gambar IV.67, fraksi galaksi *spheroid* lebih rendah di $z \leq 2$ dan lebih tinggi di $z > 4$. Sementara itu, fraksi galaksi *irregular* cukup stabil di $2 < z \leq 4$, dan lebih tinggi di *redshift* yang lebih rendah dan *redshift* yang lebih tinggi. Namun, perlu kehati-hatian dalam menginterpretasikan tren yang didapatkan pada Gambar IV.67 karena pengkategorian galaksi sebagai galaksi *spheroid* dan *irregular* hanya dilakukan berdasarkan analisis visual sampel galaksi setiap klaster.

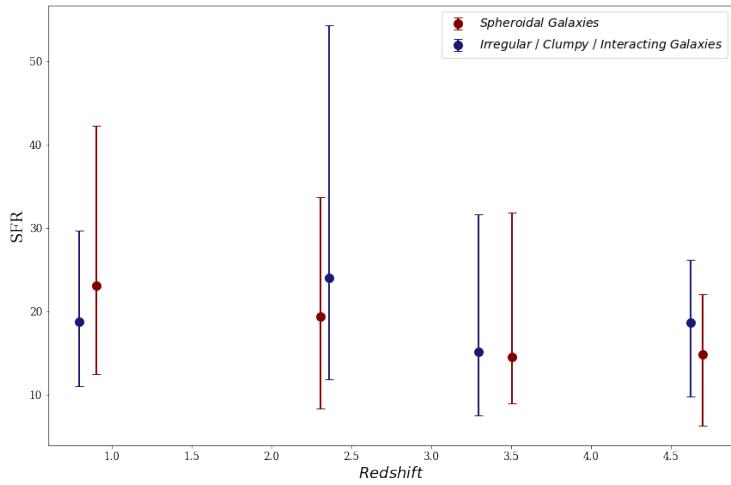


Gambar IV.67: Distribusi fraksi galaksi hasil pengelompokan yang dikategorikan secara visual pada berbagai *redshift*.

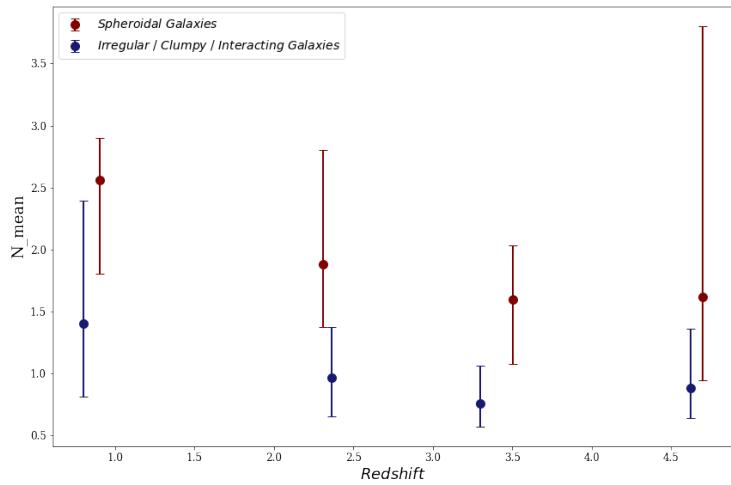
Perbandingan beberapa parameter properti galaksi seperti massa, SFR, dan indeks Sérsic untuk kategori galaksi *spheroid* dan *irregular* di berbagai *redshift* dapat dilihat pada Gambar IV.68. Distribusi massa menunjukkan bahwa untuk galaksi *spheroid* maupun *irregular* memiliki puncak di $2 < z \leq 3$. Ditinjau dari sebarannya, galaksi *spheroid* memiliki sebaran massa yang lebih besar dibandingkan galaksi *irregular*. Sementara itu, parameter SFR menunjukkan bahwa galaksi *spheroid* memiliki nilai SFR yang sedikit lebih tinggi dibandingkan galaksi *irregular* pada $z \leq 2$, meski secara umum, nilai median SFR untuk kedua kategori galaksi tersebut tidak jauh berbeda di berbagai *redshift*. Nilai indeks Sérsic untuk galaksi *spheroid* selalu lebih besar dibandingkan galaksi *irregular* di berbagai *redshift*.



(a)



(b)



(c)

Gambar IV.68: Distribusi parameter properti galaksi di berbagai *redshift*, untuk massa (panel a), SFR (panel b), dan indeks *Sérsic* (panel c).

Jika kategori morfologi hasil *clustering* sebagaimana ditunjukkan pada Tabel IV.1 dibandingkan dengan kategori morfologi dari inspeksi visual, maka nilai akurasi dan presisi dapat menjadi acuan seberapa baik pengelompokan galaksi telah dilakukan. Nilai akurasi akan memperhitungkan jumlah galaksi yang sama-sama dikategorikan sebagai satu tipe baik melalui *clustering* maupun inspeksi visual, dibandingkan dengan seluruh data. Sementara itu, nilai presisi akan memperhitungkan ketepatan prediksi *clustering* terhadap suatu kategori yang sama menurut inspeksi visual. Nilai akurasi dan presisi akan dihitung melalui persamaan IV.3 dan IV.4.

Tabel IV.2 menunjukkan nilai akurasi dan presisi hasil pengelompokan galaksi dibandingkan informasi morfologi dari inspeksi visual dari penelitian Effendi (2024). Nilai akurasi untuk rentang $2 < z \leq 4$ lebih rendah dibandingkan *redshift* lainnya karena pada rentang tersebut terdapat beberapa klaster yang tidak dimasukkan ke dalam kategori *spheroid* maupun *irregular*, sehingga perhitungan nilai akurasinya menjadi lebih rendah. Sebagian besar galaksi *spheroid* dan *irregular* hasil *clustering* dibandingkan inspeksi visual memiliki nilai presisi di atas 70%. Perlu ditekankan bahwa informasi morfologi dari inspeksi visual ini bersifat subjektif karena hanya dilakukan oleh satu responden.

Tabel IV.2: Nilai akurasi dan presisi galaksi hasil *clustering* dibandingkan morfologi dari inspeksi visual dari penelitian Effendi (2024).

<i>Bin redshift</i>	Akurasi	Presisi	
		Spheroid	Irregular
$z_{spec} \leq 2$	0.717	0.720	0.714
$2 < z_{spec} \leq 3$	0.629	0.739	0.738
$3 < z_{spec} \leq 4$	0.488	0.777	0.758
$z_{spec} > 4$	0.783	0.666	0.909
<i>All data</i>	0.615	0.737	0.754

BAB V

SIMPULAN DAN SARAN

V.1 Simpulan

Instrumen pada teleskop JWST menjadi salah satu instrumen pengamatan astronomi terbaik saat ini. JWST berhasil mengonfirmasi galaksi-galaksi pada *redshift* yang lebih jauh dibandingkan pengamatan instrumen terdahulu. Di samping berbagai keberhasilannya dalam pengamatan objek astronomis, pengamatan JWST tetap memiliki beberapa kekurangan.

Dalam sampel data penelitian ini, banyak dari data fotometri galaksi memiliki *noise* yang sangat tinggi, sehingga penampakan struktur galaksi sulit dikenali dengan jelas. Pada *redshift* yang semakin tinggi, galaksi semakin redup, sehingga *noise* dari lingkungan sekitar menjadi lebih dominan. Galaksi berbentuk *clump* pada data dengan *noise* yang tinggi membuat sulit untuk memisahkan antara *clump* atau *noise* latar belakang.

Metode *variational autoencoder* yang digunakan pada penelitian ini berhasil mempelajari fitur-fitur struktur galaksi, sehingga mesin dapat menemukan galaksi-galaksi yang 'serupa' dengan galaksi yang telah dikategorikan sebagai galaksi *spheroid*, *disk*, dan *irregular*, seperti ditunjukkan dalam Gambar IV.31, IV.30, dan IV.32. Metode ini juga berhasil meminimalisir *noise* di dalam data, sehingga citra rekonstruksi galaksi seperti dalam Gambar IV.24 tampak lebih bersih dari *noise*, dan struktur utama galaksi tetap terjaga.

Meski demikian, visualisasi parameter PCA dan UMAP menunjukkan bahwa walaupun secara visual data rekonstruksi telah meminimumkan besarnya *noise* dalam data, *noise* masih menjadi salah satu informasi yang dipelajari dan disimpan oleh mesin. Hal ini terlihat dari Gambar IV.47 dan IV.48. Selain dari representasi parameter PCA dan UMAP, pengelompokan galaksi menunjukkan kecenderungan pengelompokan berdasarkan besar kecilnya *noise* di dalam data, seperti yang tampak dalam Gambar IV.42.

Penelitian ini menunjukkan bahwa dengan arsitektur yang sama, VAE dapat merekonstruksi galaksi di *redshift* dekat maupun *redshift* tinggi. Namun, galaksi dekat perlu direpresentasikan dalam dimensi laten yang lebih tinggi dibandingkan galaksi jauh, karena bentangan sudut galaksi dekat lebih besar

dibandingkan galaksi *redshift* tinggi. Kurangnya jumlah representasi laten dapat menyebabkan fitur-fitur detil seperti lengan spiral di galaksi dekat menjadi tidak terekstrak dengan baik.

Selain itu, hasil penelitian ini menunjukkan bahwa *fitting* dengan *multi Sérsic* bisa memberikan parameter morfologi galaksi yang lebih baik dibandingkan *fitting* dengan *single Sérsic*, khususnya pada galaksi yang berada dekat dengan objek terang didekatnya.

Hasil penelitian ini menunjukkan sebaran massa terhadap radius efektif galaksi selaras dengan relasi yang dijelaskan dalam van der Wel dkk. (2014). Relasi massa-radius untuk galaksi *star forming* tampak lebih landai dibandingkan relasi massa-radius untuk galaksi *quiescent*. Hal ini terkait dengan distribusi massa di dalam galaksi *quiescent* yang terdistribusi merata di dalam galaksi, sementara pada galaksi *star forming*, massa terkonsentrasi di pusat galaksi.

Sebaran radius efektif galaksi terhadap *redshift* menunjukkan bahwa pada *redshift* yang semakin tinggi, ukuran galaksi menjadi semakin kecil. Tren ini tampak jelas untuk distribusi galaksi *star forming*. Sementara untuk galaksi *quiescent*, tren serupa dapat ditemukan dalam galaksi *quiescent* bermassa menengah. Perlu kehati-hatian dalam mendeskripsikan tren untuk galaksi *quiescent* bermassa rendah dan tinggi, salah satunya karena jumlah galaksi *quiescent* yang lebih sedikit. Jika dilihat dari sebaran rapat massa galaksi berdasarkan radius efektifnya, secara umum terlihat bahwa galaksi dengan ukuran $0.5 - 1\text{kpc}$ mendominasi di berbagai *redshift*. Semakin besar ukuran galaksi, jumlahnya semakin sedikit.

Hubungan radius efektif galaksi terhadap panjang gelombang menunjukkan bahwa radius efektif galaksi pada pengamatan di panjang gelombang pendek lebih tinggi dibandingkan radius efektif galaksi dari pengamatan pada panjang gelombang panjang. Hal ini dapat disebabkan karena dua hal menurut penelitian Baes dkk. (2024), yakni gradien populasi bintang dengan persentase 80% dan *dust attenuation* dengan persentase 20%. Hal ini menjelaskan tren radius terhadap panjang gelombang yang ditemukan pada penelitian ini. Secara umum nilai radius efektif dari pengamatan tiga filter tidak menunjukkan perbedaan yang cukup signifikan, namun pada panjang gelombang pendek, nilai radius efektif sedikit lebih tinggi.

V.2 Saran

Dalam penelitian ini, dilakukan banyak penyederhaan dalam perhitungan maupun metode yang digunakan. Dalam Subbab ini penulis memberikan beberapa saran untuk pengembangan penelitian ini kedepannya.

1. Pengelompokan galaksi yang dilakukan dalam penelitian ini menunjukkan bahwa *noise* menjadi salah satu parameter yang mengganggu dalam proses pengelompokan galaksi berdasarkan morfologinya. Kedepannya, perlu dilakukan tahap *denoising* citra galaksi sehingga data galaksi yang digunakan untuk *training* menjadi lebih homogen, dan mesin dapat mempelajari data galaksi berdasarkan morfologinya dengan lebih baik.
2. Salah satu tahap *pre-processing* yang dilakukan pada penelitian ini adalah pemangkasan ukuran citra galaksi sedemikian sehingga diharapkan dalam satu citra hanya terdapat satu galaksi. Namun tetap ditemukan beberapa citra galaksi yang terdapat lebih dari satu objek di dalamnya. Hal ini mempengaruhi kinerja mesin dalam mempelajari morfologi galaksi. Untuk mengatasi hal ini, perlu dilakukan pembersihan data dengan lebih baik misalnya dengan membangun peta segmentasi dan memotong citra sesuai dengan segmentasi tersebut. Aplikasi *Galclean* juga dapat digunakan untuk membersihkan data, namun perlu kehati-hatian dalam penyesuaian *hyperparameter* yang tersedia. Selain itu, dapat dilakukan pengelompokan galaksi awal untuk menemukan klaster berisi citra yang terdapat lebih dari satu galaksi, lalu menyisihkan klaster tersebut.
3. Data yang digunakan pada penelitian ini menggunakan data pengamatan JWST yang diarsipkan dalam katalog *EAZY* versi 6. Untuk memberikan hasil dengan kualitas yang lebih baik, penelitian selanjutnya dapat mempertimbangkan untuk menggunakan data katalog *EAZY* versi terbaru.
4. Selain pengelompokan dengan metode *k-means clustering* dan *hierarchical clustering*. Penelitian ini dapat dikembangkan dengan melakukan pengelompokan menggunakan metode lainnya, seperti *HDBSCAN*, atau bahkan menggunakan metode *supervised learning* dalam mengklasifikasikan galaksi berdasarkan katalog morfologi galaksi yang telah diperoleh dari inspeksi visual.
5. Algoritma VAE dari penelitian ini menunjukkan kemampuannya untuk menemukan galaksi-galaksi yang 'mirip' dengan galaksi yang telah diberi label morfologi sebelumnya. Algoritma ini dapat dicoba untuk diaplikasikan dalam mencari galaksi-galaksi eksotis atau galaksi yang mengalami pelensaan.
6. Algoritma VAE yang dikembangkan pada penelitian ini dapat didorong

untuk mengelompokan galaksi bukan hanya untuk kategori *disk*, *spheroid*, dan *irregular*, tetapi dapat mengelompokan galaksi berdasarkan jumlah *clump* di-dalamnya.

DAFTAR PUSTAKA

- H. Mo, F. van den Bosch dan S. White, 2010. *Galaxy Formation and Evolution*. New York: United States of America by Cambridge University Press.
- Valentino, dkk. 2023. An Atlas of Color-selected Quiescent Galaxies at $z > 3$ in Public JWST Fields. *The Astrophysical Journal*. **947**.
- Speagle, dkk. 2014. A Highly Consistent Framework for the Evolution of the Star-Forming "Main Sequence" from $z = 0\text{--}6$. *The Astrophysical Journal Supplement Series*. **214**.
- Hubble, E.P. 1926. Extragalactic Nebulae. *The Astrophysical Journal* **64**. hlm. 321-369.
- de Vaucouleurs, G. 1959. Classification and Morphology of External Galaxies. *Handbuch Der Physik / Encyclopedia of Physics*. hlm. 275-310.
- Dieleman, S. dkk. 2017. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *MNRAS*. **450(2)**. hlm. 1441-1459.
- Tohill, C. dkk. 2024. A Robust Study of High-redshift Galaxies: Unsupervised Machine Learning for Characterizing Morphology with JWST up to $z \sim 8$. *The Astrophysical Journal*. **962**.
- Sérsic, J.L. 1963. Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletín de la Asociación Argentina de Astronomía*. **6**. hlm. 41-43.
- Kartaltepe, dkk. 2023. CEERS Key Paper III: The Diversity of Galaxy Structure and Morphology at $z = 3\text{--}9$ with JWST. *The Astrophysical Journal Letters*. **946**.
- Rodriguez-Gomez, dkk. 2019. The optical morphologies of galaxies in the IllustrisTNG simulation: a comparison to Pan-STARRS observations. *MNRAS*. **483**. hlm. 4140–4159.

- Ferrari, dkk. 2015. Morfometryka – A New Way of Establishing Morphological Classification of Galaxies. *The Astrophysical Journal*. **814**.
- van der Wel, dkk. 2014. 3D-HST+CANDELS: THE EVOLUTION OF THE GALAXY SIZE–MASS DISTRIBUTION SINCE $z = 3$. *The Astrophysical Journal*. **788**.
- Drouart, G. 2013. Relation noyau actif et histoire de la formation d'étoiles dans les radio galaxies distantes.
- Alzubaidi, dkk. 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*. **8**.
- Hossain, dkk. 2023. Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison. *Iran Journal of Computer Science*. **6**. hlm. 397–417.
- Jaelani, dkk. 2024. Survey of gravitationally lensed objects in HSC imaging (SuGOHI) - X. Strong lens finding in the HSC-SSP using convolutional neural networks. *MNRAS*. **535**. hlm. 1625–1639.
- Brammer, G. 2023. grizli. *Zenodo*.
- Brammer, dkk. 2008. EAZY: A Fast, Public Photometric Redshift Code. *The Astrophysical Journal*. **686**. hlm. 1503-1513.
- Casey, dkk. 2023. COSMOS-Web: An Overview of the JWST Cosmic Origins Survey. *The Astrophysical Journal*. **954**.
- Abdurrouf, dkk. 2021. Introducing piXedfit: A Spectral Energy Distribution Fitting Code Designed for Resolved Sources. *The Astrophysical Journal Supplement Series*. **254**.
- Iyer, dkk. 2019. Nonparametric Star Formation History Reconstruction with Gaussian Processes. I. Counting Major Episodes of Star Formation. *The Astrophysical Journal*. **879**.
- Vika, dkk. 2013. MegaMorph–multi-wavelength measurement of galaxy structure: Sérsic profile fits to galaxies near and far. *MNRAS*. **435**. hlm. 623-649.
- Peng, dkk. 2002. Detailed Structural Decomposition of Galaxy Images. *The Astronomical Journal*. **124**. hlm. 266-293.

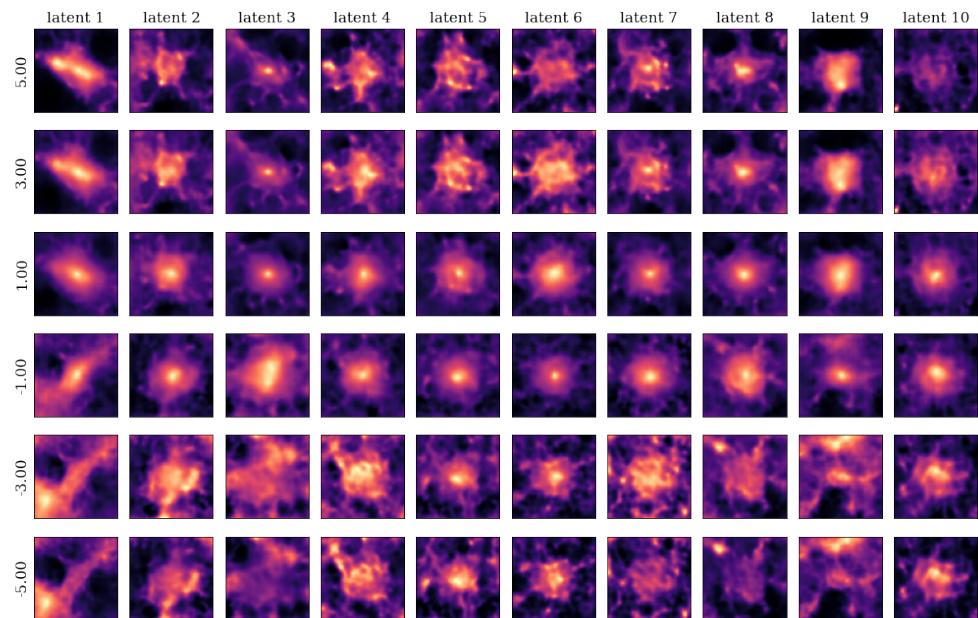
- Peng, dkk. 2010. Detailed Decomposition of Galaxy Images. II. Beyond Axi-symmetric Models. *The Astronomical Journal*. **139**. hlm. 2097-2129.
- Effendi, M.N.I 2024. Studi Struktur dan Morfologi Galaksi Pada $z > 2$ dengan Data JWST.
- Lotz, dkk. 2010. A New Nonparametric Approach to Galaxy Morphological Classification. *The Astronomical Journal*. **128**. hlm. 163-182.
- Abraham, R.G. 1998. Perspectives in Physical Morphology.
- Williams, dkk. 2009. Detection of Quiescent Galaxies in a Bicolor Sequence from $Z = 0\text{-}2$. *The Astrophysical Journal*. **691**. hlm. 1879-1895.
- Kingma, D.P. dan Welling, M. 2019. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*. **12**. hlm. 307-392.
- McInnes, dkk. 2018. UMAP:Uniform Manifold Approximation and Projection for Dimension Reduction.
- Willett, K.W. dkk. 2013. Galaxy Zoo 2: detailed morphological classifications for 304.122 galaxies from the Sloan Digital Sky Survey. *MNRAS*. **435(4)**. hlm. 2835-2860.
- Hart, R.E. dkk. 2016. Galaxy Zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias. *MNRAS*. **461(4)**. hlm. 3663–3682.
- Ferreira, L. dkk. 2018. galclean. *Zenodo*.
- Scott, D.W. 2010. Scott's rule. *Wiley Interdisciplinary Reviews: Computational Statistics*. **2(4)**. hlm. 497-502.
- Delaye, L. 2013. EVOLUTION DES PROPRIETES STRUCTURELLES DES GALAXIES 'DE TYPE PRECOCE DANS DIFFERENTS ENVIRONNEMENTS.
- Bertin, E. dan Arnouts, S. 1996. SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement*. **117**. hlm. 393-404.
- Barbary, K. 2016. SEP: Source Extractor as a library. *The Journal of Open Source Software*. **1(6)**. hlm. 58.
- Baes, M. dkk. 2024. The TNG50-SKIRT Atlas: Wavelength dependence of the effective radius. *Astronomy & Astrophysics*. **683**.

LAMPIRAN

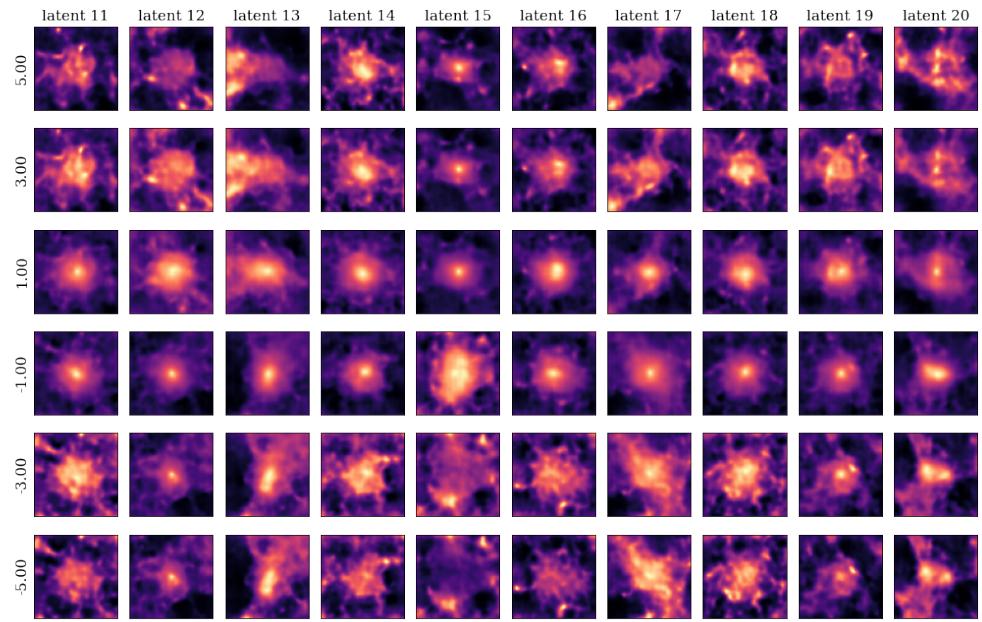
LAMPIRAN A

Visualisasi Parameter Laten dari Galaksi Dekat

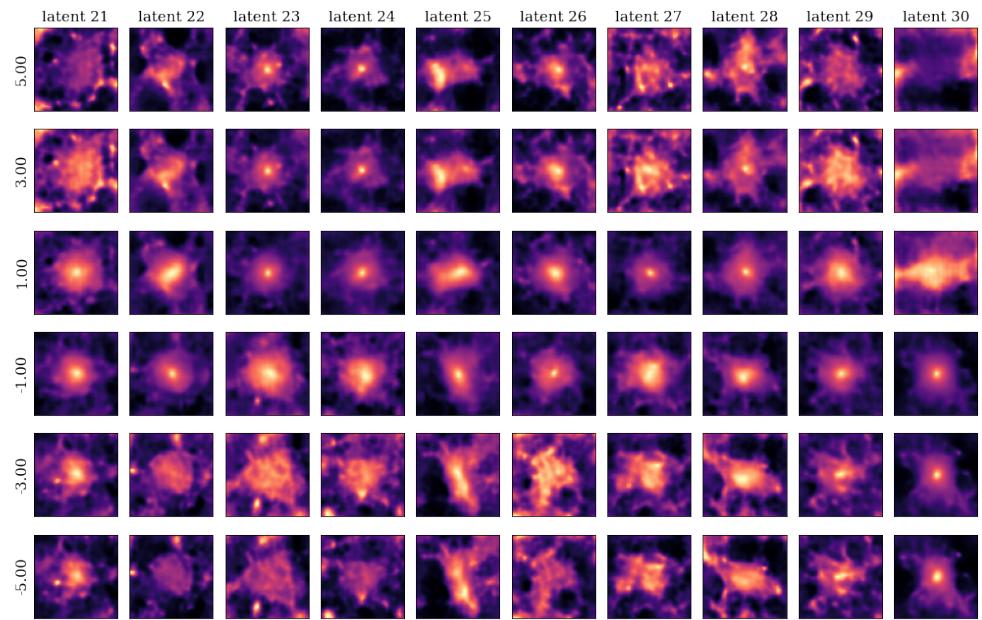
A.1 Visualisasi 100 Parameter Laten dari Galaksi Dekat



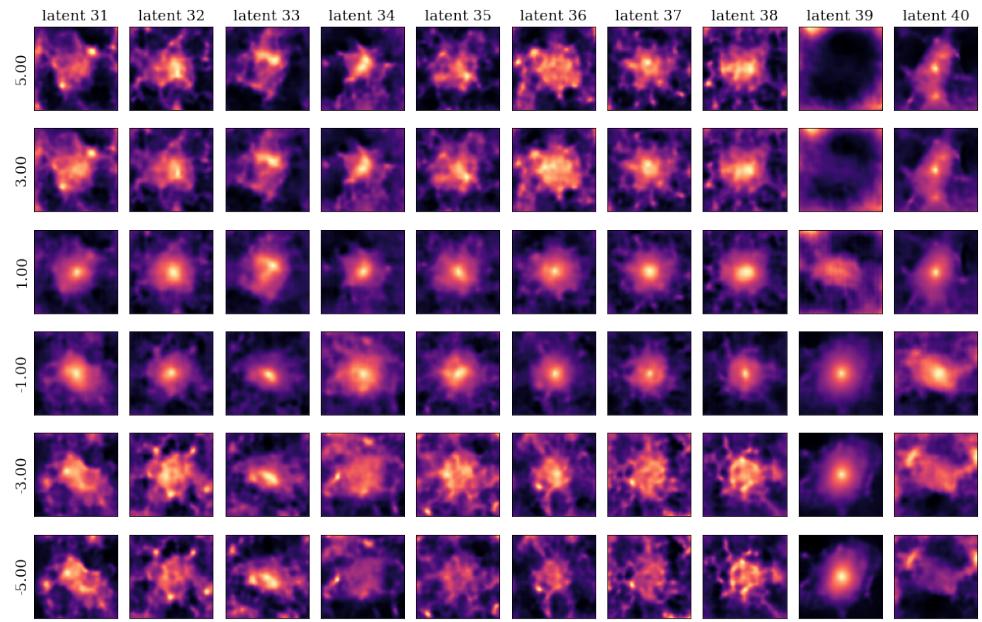
Gambar A.1: Visualisasi parameter laten 1 hingga 10.



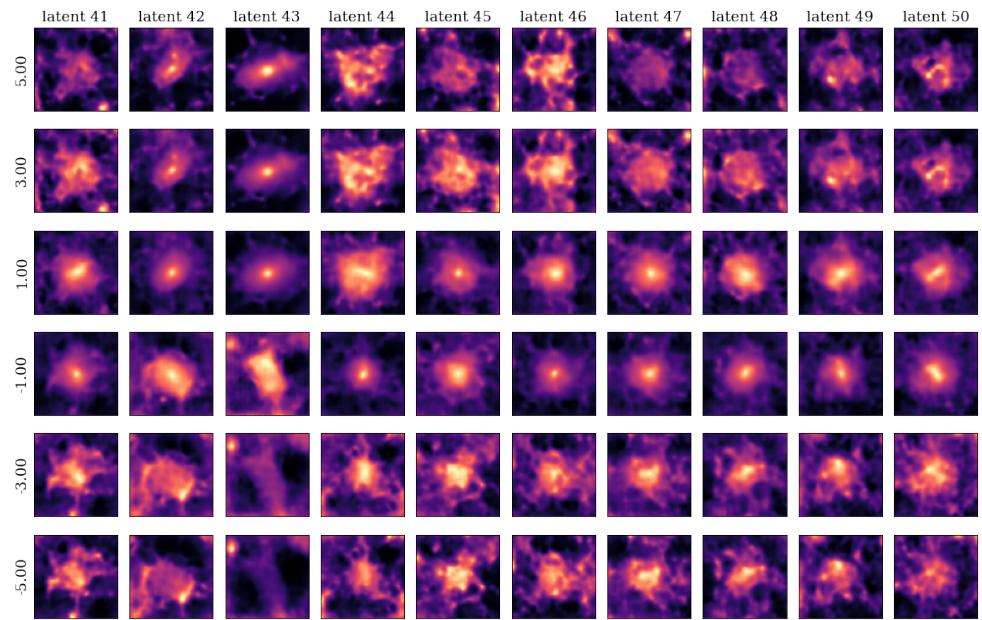
Gambar A.2: Visualisasi parameter laten 11 hingga 20.



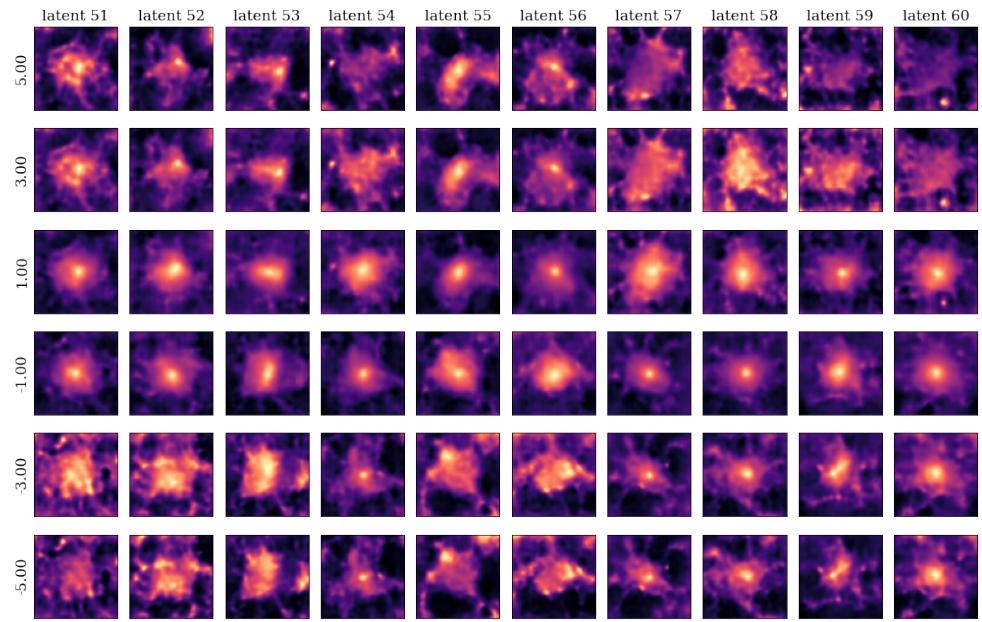
Gambar A.3: Visualisasi parameter laten 21 hingga 30.



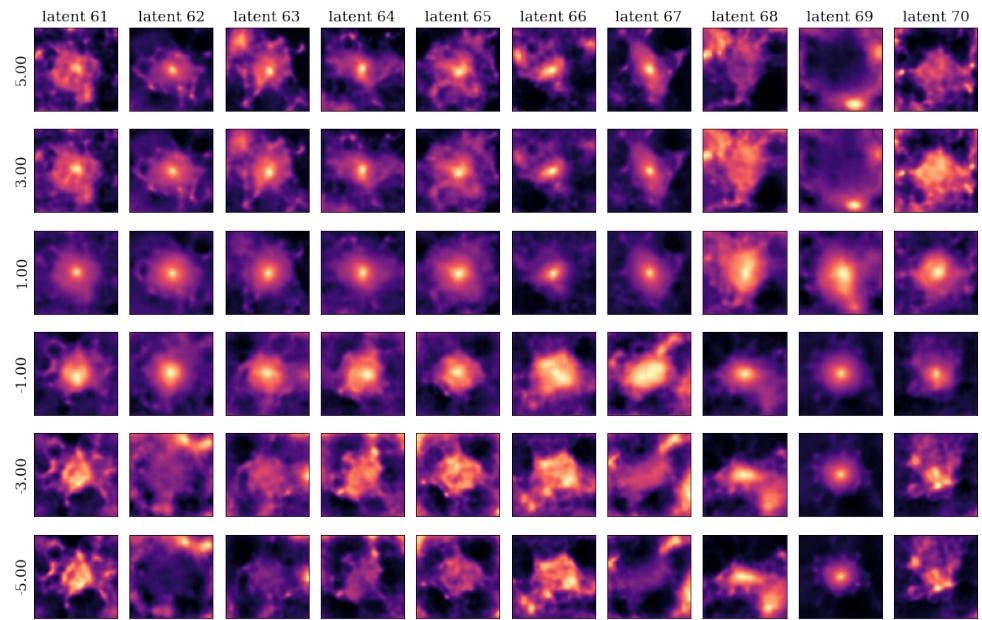
Gambar A.4: Visualisasi parameter laten 31 hingga 40.



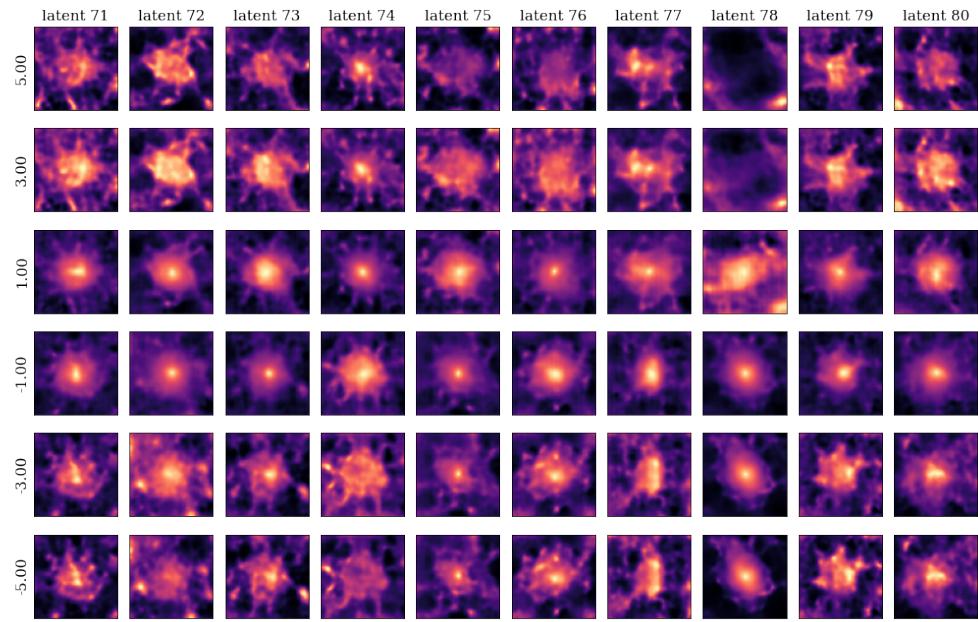
Gambar A.5: Visualisasi parameter laten 41 hingga 50.



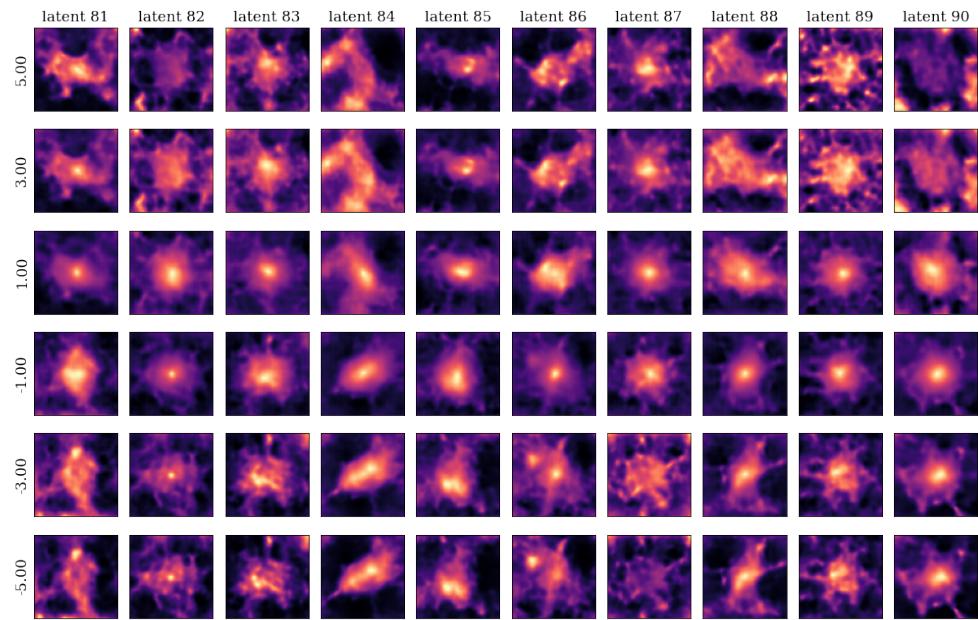
Gambar A.6: Visualisasi parameter laten 51 hingga 60.



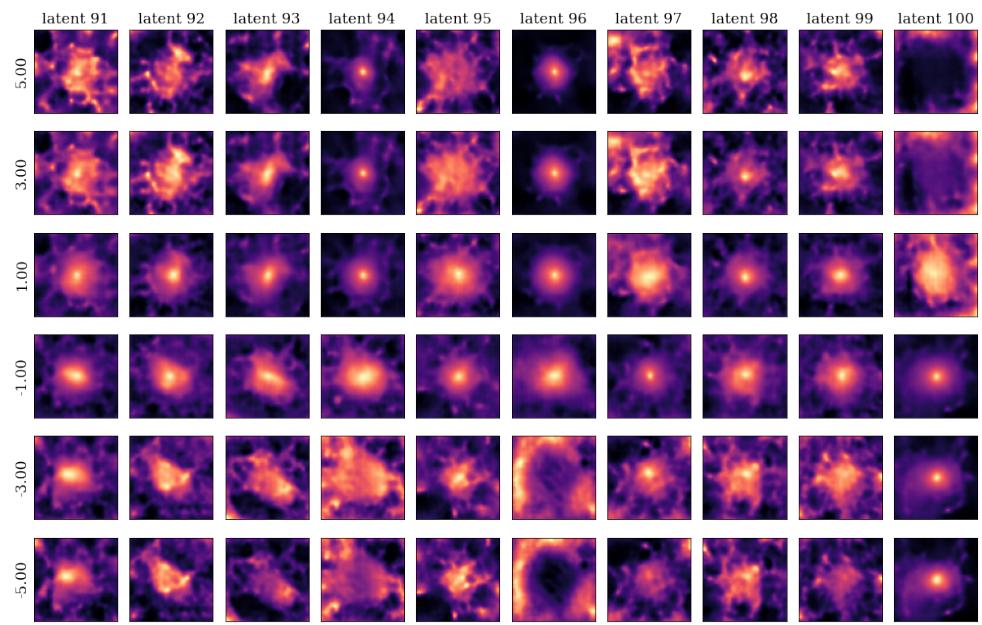
Gambar A.7: Visualisasi parameter laten 61 hingga 70.



Gambar A.8: Visualisasi parameter laten 71 hingga 80.



Gambar A.9: Visualisasi parameter laten 81 hingga 90.



Gambar A.10: Visualisasi parameter laten 91 hingga 100.